The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

Platform-Related Factors in Repeatability and Reproducibility of Crowdsourcing Tasks

Rehab Qarout,[1] **Alessandro Checco,**[1] **Gianluca Demartini,**[2] **Kalina Bontcheva**[1]
[1]The University of Sheffield, UK
[2]The University of Queensland, Australia

Abstract

Crowdsourcing platforms provide a convenient and scalable way to collect human-generated labels on-demand. This data can be used to train Artificial Intelligence (AI) systems or to evaluate the effectiveness of algorithms. The datasets generated by means of crowdsourcing are, however, dependent on many factors that affect their quality. These include, among others, the population sample bias introduced by aspects like task reward, requester reputation, and other filters introduced by the task design.

In this paper, we analyse platform-related factors and study how they affect dataset characteristics by running a longitudinal study where we compare the reliability of results collected with repeated experiments over time and across crowdsourcing platforms. Results show that, under certain conditions: 1) experiments replicated across different platforms result in significantly different data quality levels while 2) the quality of data from repeated experiments over time is stable within the same platform. We identify some key task design variables that cause such variations and propose an experimentally validated set of actions to counteract these effects thus achieving reliable and repeatable crowdsourced data collection experiments.

1 Introduction

The rise of several crowdsourcing platforms has enabled the collection of human labels at scale. Researchers using such platforms (as requesters) aim 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 a *"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 findings. 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), the 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, this 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 (à Campo et al. 2019; Blohm et al. 2018; Mourelatos, Frarakis, and Tzagarakis 2017; Kohler 2018) have discussed output variability across crowdsourcing platforms by studying external and internal factors affecting it. Nevertheless, reproducing the results for identical tasks over multiple platforms has not previously been explored.

Previous studies in machine learning (Rosten, Porter, and Drummond 2010) and human-computer interaction (Wilson et al. 2013; Hornbæk et al. 2014) represent 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 result in 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 paper, 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 two dif-

ferent crowdsourcing platforms, Amazon Mechanical Turk (MTurk) and Figure Eight (F8), 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.

These two platforms differ in their workers' demographics and quality control mechanisms: in MTurk requesters can reject a job and withhold its payment, while in F8 workers can be only excluded from future jobs in a batch, but they are always paid for a completed job. Furthermore, the demographic distribution of the workers on MTurk includes mainly the US and India (Difallah, Filatova, and Ipeirotis 2018) and workers are all recruited through a unique channel. The F8 worker distribution is less well known and workers are sourced to the platform from several different channels.

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 paper is organised as follows. Section 2 presents our research questions and summarises the contributions of our work. We review related work on evaluating crowdsourcing platforms and on repeating tasks in Section 3. Section 4 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 presents our experimental results and findings on obtaining repeatable results. Section 6 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 7.

2 Research Questions and Novelty

In this paper, we examine the following research questions:

- **RQ1 - 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?
- **RQ2 - Reproducibility:** Is there a significant difference in the quality of the results for the same task reproduced on a different platform?
- **RQ3 - Generalisability:** Are the results obtained consistent over different classification tasks?

To address **RQ1**, we repeated the same experiment over multiple weeks to measure the reliability and consistency of the results over time (i.e., repeatability). When addressing **RQ2**, 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 Figure Eight (F8).

To generalise our findings (**RQ3**) we used the same task design as in Experiment 1 and 2 over three different classification tasks described in Section 4. 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.

3 Related Work

3.1 Crowdsourcing Platforms Evaluation

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. Using several types of experiments, they designed them online and in a traditional lab setting. After receiving data from both experiments, their findings confirmed that the quality of the data collected under the experimental conditions in MTurk is comparable to the quality of the data collected in the traditional lab-based way. However, the consistency of the results over time is not studied.

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. The results of 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 in a few seconds (Crump, McDonnell, and Gureckis 2013). In our work, we extend this observation by measuring and comparing the completion time over different platforms.

In terms of comparing crowdsourcing platforms, Peer et al. (2016) introduced a comparison study between Prolific Academic (ProA), Figure Eight, and MTurk. Our work extends this approach by studying the factors affecting result consistency and stability over multiple weeks.

In the same context, Mourelatos, Frarakis, and Tzagarakis (2017) presented a ranking model (based on Alexa ranking) for crowdsourcing platforms. They also compared platforms over time according to: *type of service provided, quality and reliability, region, online imprint*, and discussed the impact of the platform characteristics on traffic data and popularity. Our work differs in the fact that we make use of an experiment-based comparative analysis.

3.2 Reliability of Repeating Tasks over Time

Williams et al. (2017) studied the consistency of results when crowd workers repeat the same task twice. They used a method to duplicate a task in a queue of tasks presented to the same worker. This method examines the reliability of workers when completing duplicated tasks consistently. In our approach, we aim to evaluate the consistency and reliability of the platforms, so we target new workers at each task repetition.

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 the 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. Our work differs in the usage of a shorter time scale, multiple repetitions and multiple crowdsourcing platforms.

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 paper we perform a longer-term analysis by means of data collected during a longitudinal study over different crowdsourcing platforms and different kinds of classification tasks.

4 Methodology

We performed the first experiment to address **RQ1** and **RQ2**. 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 6.

Furthermore, to answer **RQ3** 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.

For both Experiment 1 and 2, we used three different classification tasks with three different kinds of labelling: documents, tweets, and images. More details about the used datasets are presented in Section 4.1.

4.1 Dataset

We used three datasets in our crowdsourcing experiments, each with a different classification task and difficulty level.

The first dataset (Dataset 1) is a collection of tweets gathered during a crisis/emergency situation (Imran, Mitra, and Castillo 2016). The goal of the crowdsourcing experiment is to categorise each tweet content into one of nine possible categories. The high number of labels make this task the most challenging, and it allows us to easily detect underperforming workers or bots.

The second dataset (Dataset 2) is a collection of product reviews related to fashion items (Chernushenko et al. 2018). 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').

The third dataset (Dataset 3) is a collection used in the Eighth Text Retrieval Conference[1] (TREC8) (Hawking et al. 2000) which contains documents, queries, and editorial relevance judgements[2] from a general web search. In this task, 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.

4.2 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. Three separate Human Intelligent Task (HITs) (one for each dataset) have been published each week on each platform. The number of documents was selected to ensure each task could be finished in approximately 5-6 minutes.

The interface has been designed to appear identical in both platforms. We used an external server to host the task interface and visualised it into each platform using iframes. The only differences 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. [3]

Crowd workers were rewarded according to US minimum wage rates ($8 per hour) after internal tests to estimate the

[1] http://trec.nist.gov,

[2] Assessors are human judges hired and trained by NIST.

[3] The GUI for the task design can be found in https://github.com/AlessandroChecco/crowd-reproducibility .

average task execution time. Since our focus was on the differences between platforms, we run a unique 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 for each dataset only once: after that, worker identifiers were not allowed to participate in future editions of the same task. However, there is the possibility that the same worker would label items from more than one dataset. 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, thus, the ordering of the items was consistent in each task and items have been sampled without replacement. 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 20th 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.

4.3 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. Since the population size is unknown and influenced by many factors, we used 30 participants per platform for the pilot experiment (Isaac and Michael 1995; Hill 1998).

Following (Thompson 2012), we estimated the sample size when comparing the means of a continuous outcome variable in two independent populations. To obtain 75% statistical power, the sample size needs to be of 150 workers on each platform 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 to use a small number of workers confidently, knowing that the obtained variability 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. This information can be used as priors for techniques of aggregation number estimation, e.g. (Chen, Lin, and Zhou 2013).

5 Experiment 1 - Achieving Repeatability

For Experiment 1, we used the same task design as presented in Section 4.2. 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.

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 1 and Table 4 (Average time per assignment).

Moreover, Figure 2 and Table 4 (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 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.

Table 1: Two-way ANCOVA for Dataset 1 in Experiment 1.

	sum_sq	df	F	PR(>F)
Platform	1.96	1.0	47.27	9.6×10^{-12}
Week	0.00002	1.0	0.0006	9.8×10^{-1}
Platform: Week	0.02	1.0	0.54	4.6×10^{-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 2: 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.

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 1-8). 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

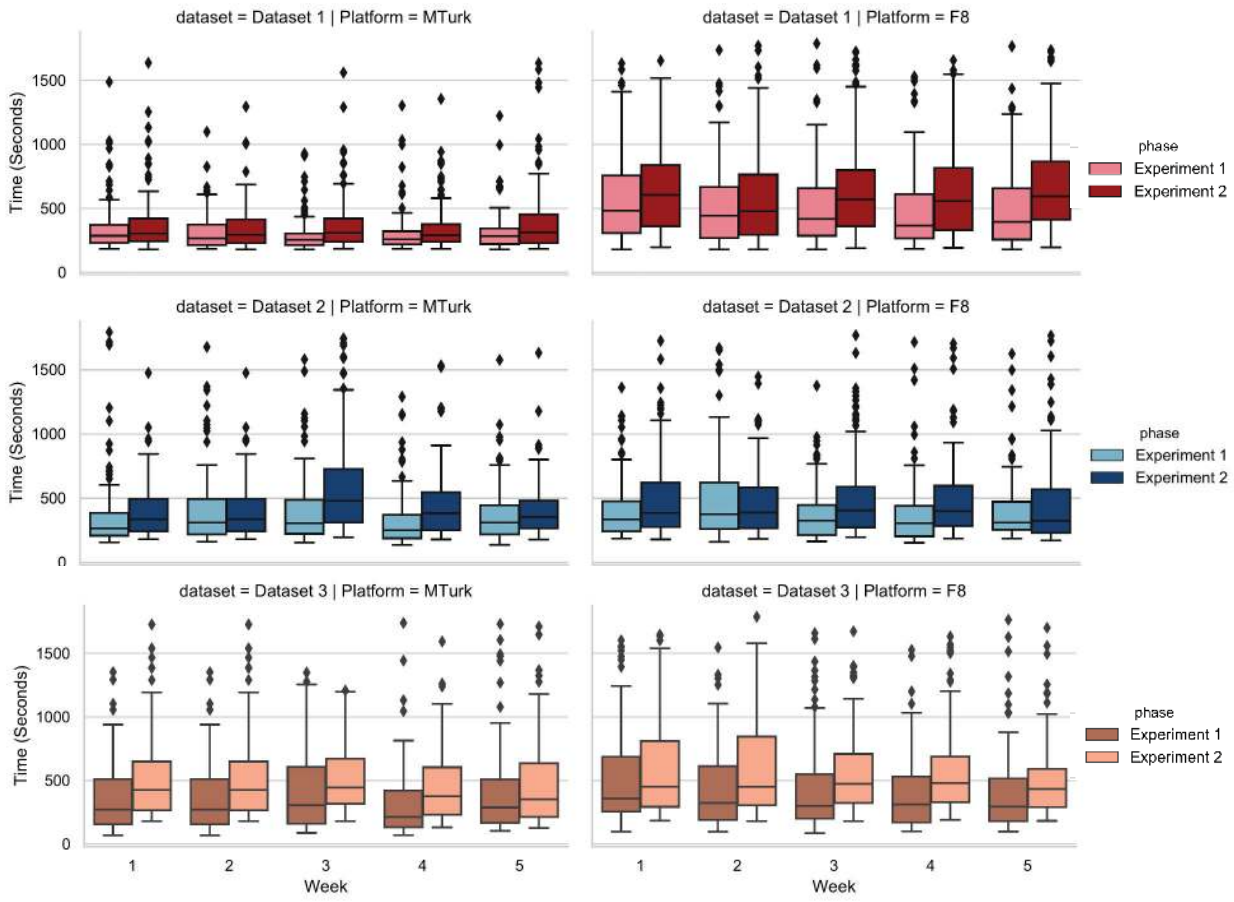

Figure 1: Average time per assignment for Experiment 1 and Experiment 2 for all 3 datasets.

Table 3: 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.

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 3 and Table 4 (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 Section 6.

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

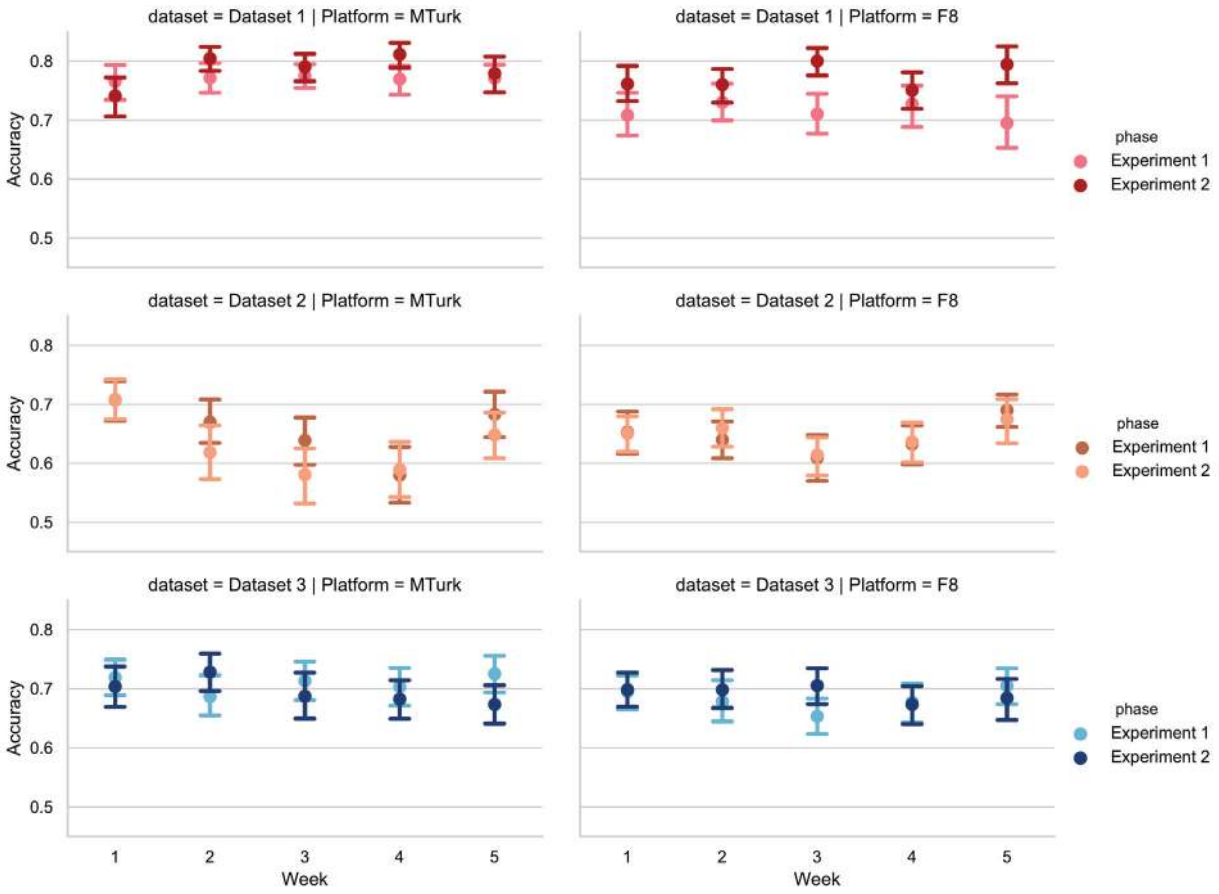

Figure 2: Accuracy distribution over time for Experiment 1 and Experiment 2 for all 3 dataset. Dataset 1 shows a statistically significant difference in accuracy between the two platforms when using the default payment scheme (Experiment 1).

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.

6 Experiment 2 - Achieving Reproducibility

To equalise the conditions between platforms, 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 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 1 and Table 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 Ipeirotis 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 demo-

Table 4: 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	$0,72 \pm 0.17$	$0,70 \pm 0.17$
	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

Table 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

graphically 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 3 and Table 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.

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 2 and Table 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.

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 6 shows that none of the factors have a significant effect

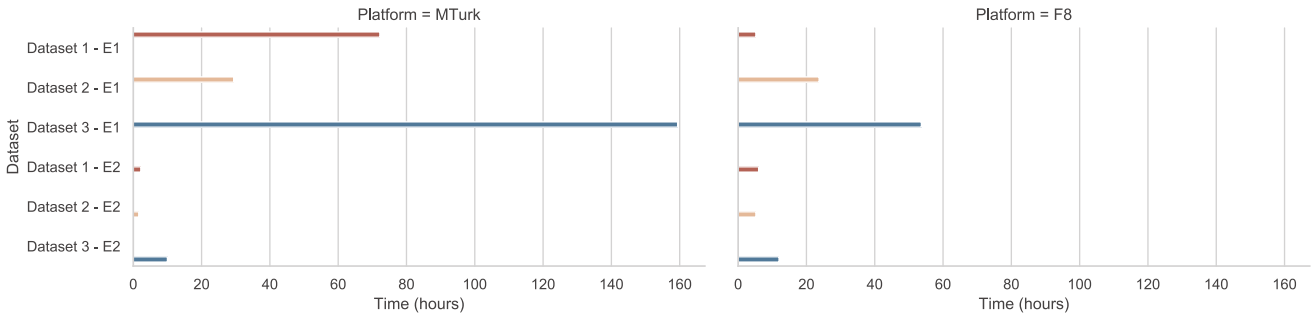

Figure 3: Average completion time for all batches in Experiment 1 and 2.

Table 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 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.

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 6, 7 and 8).

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).

Table 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.

7 Conclusions

In this paper, we have looked at how crowdsourcing experiments can be repeated and reproduced. Our findings show that: (1) using standard crowdsourcing platform settings, the completion time for the same data collection experiment may vary by orders of magnitude across different platforms, also obtaining different levels of accuracy in some cases; (2) **(RQ1)** in our results we obtained *repeatable* experiments in each of the studied crowdsourcing platforms, but we have observed a problem of *reproducibility* over different platforms when the task is difficult (Dataset 1); and that (3) **(RQ2)** by setting the same payment expectations and rejection rate across different platforms, we achieved both *repeatability* and *reproducibility* of our crowdsourcing experiments; (4) we observed similar results over time and over different platforms across different datasets. Aligning payment schemes across platforms, increased repeatability and reproducibility over different classification tasks **(RQ3)**.

While the absence of quality control does reduce labelling quality, we have observed that the threat of unpaid task rejections does not increase crowd labelling quality, but rather it reduces the attractiveness of the task and thus increases its overall completion time. On the other hand, our results confirm that introducing a bonus for high-quality labels has a positive effect on labelling quality.

This work has the following limitations: (i) By controlling for task appearance, we did not consider the effect of platform design choices. (ii) We only looked at classification tasks. (iii) We did not observe a statistically significant difference in accuracy between the two platforms for Datasets 2 and 3. This can be explained by the fact that Dataset 1 consisted of more difficult tasks, where the elements to be classified could belong to 1 of 9 possible classes. Thus, Dataset 1

has an extreme correction by chance factor, 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.

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

References

à Campo, S.; Khan, V.-J.; Papangelis, K.; and Markopoulos, P. 2019. Community heuristics for user interface evaluation of crowdsourcing platforms. *Future Generation Computer Systems* 95:775–789.

Assis Neto, F. R., and Santos, C. A. 2018. Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management. *Information Processing and Management* 54(4):490–506.

Bentley, F. R.; Daskalova, N.; and White, B. 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*, 1092–1099.

Blanco, R.; Halpin, H.; Herzig, D. M.; Mika, P.; Pound, J.; and Thompson, H. S. 2011. Repeatable and Reliable Search System Evaluation using Crowdsourcing. *Journal of Web Semantics* 21:923–932.

Blohm, I.; Zogaj, S.; Bretschneider, U.; and Leimeister, J. M. 2018. How to manage crowdsourcing platforms effectively? *California Management Review* 60(2):122–149.

Checco, A.; Roitero, A.; Maddalena, E.; Mizzaro, S.; and Demartini, G. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-17)*, 11–20.

Chen, X.; Lin, Q.; and Zhou, D. 2013. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *International conference on machine learning*, 64–72.

Cheng, J.; Teevan, J.; Iqbal, S. T.; and Bernstein, M. S. 2015. Break it down: A comparison of macro-and microtasks. In *CHI '15*, 4061–4064. ACM.

Chernushenko, I.; Gers, F. A.; Löser, A.; and Checco, A. 2018. Crowd-labeling fashion reviews with quality control. *CoRR* abs/1805.09648.

Crump, M. J. C.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8(3).

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*, volume 9, 135–143.

Harter, S. P. 1996. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* 47(1):37–49.

Hawking, D.; Voorhees, E.; Craswell, N.; and Bailey, P. 2000. Overview of the TREC-8 web track. In. In *Proceedings of Eighth Text Retrieval Conference (TREC8). National Institute of Standards and Technology Special Publication 500-246*, 131–148.

Hill, R. 1998. What sample size is "enough" in internet survey research? *An Electronic Journal for the 21st Century* 6(3-4):1–10.

Hornbæk, K.; Sander, S. S.; Bargas-Avila, J. A.; and Grue Simonsen, J. 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, 3523–3532. New York, NY, USA: ACM.

Imran, M.; Mitra, P.; and Castillo, C. 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 2016)*, 1638–1643.

Isaac, S., and Michael, W. B. 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*. Edits publishers.

Kittur, A.; Nickerson, J.; and Bernstein, M. 2013. The Future of Crowd Work. In *Proc. CSCW '13*, 1–17.

Kohler, T. 2018. How to Scale Crowdsourcing Platforms. *California Management Review* 60(2):98–121.

Mourelatos, E.; Frarakis, N.; and Tzagarakis, M. 2017. A Study on the Evolution of Crowdsourcing Websites. *ISSNOnline) European Journal of Social Sciences Education and Research* 11(1):2411–9563.

Paritosh, P. 2012. Human Computation Must Be Reproducible. In *WWW2012*.

Parkhurst, D. F. 2001. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience* 51(12):1051–1057.

Peer, E.; Samat, S.; Brandimarte, L.; and Acquisti, A. 2016. Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70(January):153–163.

Rosten, E.; Porter, R.; and Drummond, T. 2010. Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32:105–119.

Sun, P., and Stolee, K. T. 2016. Exploring crowd consistency in a mechanical turk survey. In *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering - CSI-SE '16*, 8–14. New York, New York, USA: ACM Press.

Thompson, S. K. 2012. Sample Size. In *Sampling*. Wiley-Blackwell. chapter 4, 53–56.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2015. Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval* 18(5):445–472.

Waterhouse, T. P. 2013. Pay by the bit. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 623–638.

Williams, A.; Willis, C. G.; Davis, C. C.; Goh, J.; Ellison, A. M.; and Law, E. 2017. Deja Vu: Characterizing worker quality using task consistency. In *ACM CHI Conference on Human Factors in Computing Systems*, In review.

Wilson, M. L.; Coyle, D.; Resnick, P.; and Chi, E. H. 2013. RepliCHI-The Workshop. Technical report.