The Impact of Task Abandonment in Crowdsourcing

Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini

Abstract—Crowdsourcing has become a standard methodology to collect manually annotated data such as relevance judgments at scale. On crowdsourcing platforms like Amazon MTurk or FigureEight, crowd workers select tasks to work on based on different dimensions such as task reward and requester reputation. Requesters then receive the judgments of workers who self-selected into the tasks and completed them successfully. Several crowd workers, however, preview tasks, begin working on them, reaching varying stages of task completion without finally submitting their work. Such behavior results in unrewarded effort which remains invisible to requesters.

In this paper, we conduct an investigation of the phenomenon of task *abandonment*, the act of workers previewing or beginning a task and deciding not to complete it. We follow a three-fold methodology which includes 1) investigating the prevalence and causes of task abandonment by means of a survey over different crowdsourcing platforms, 2) data-driven analysis of logs collected during a large-scale relevance judgment experiment, and 3) controlled experiments measuring the effect of different dimensions on abandonment. Our results show that task abandonment is a widely spread phenomenon. Apart from accounting for a considerable amount of wasted human effort, this bears important implications on the hourly wages of workers as they are not rewarded for tasks that they do not complete. We also show how task abandonment may have strong implications on the use of collected data (for example, on the evaluation of Information Retrieval systems).

Index Terms—Task Abandonment, Crowdsourcing, Relevance Judgments.

1 Introduction

Crowdsourcing has become a wide-spread technique to collect large amounts of manually annotated data. In paid micro-task crowdsourcing platforms like Amazon MTurk (AMT) and FigureEight (F8)[1], one of the biggest challenges lies in the low quality of the collected data. To deal with this problem, previous research has looked at different approaches ranging from truth inference methods by means of complex answer aggregation models [1], [2] to profiling crowd workers in order to assign them tasks they can perform well on [3], [4]. In *pull* crowdsourcing platforms like AMT, another aspect that impacts quality is *selection bias*, which is introduced when workers decide to work on certain microtasks (also known as Human Intelligence Tasks or HITs) from the list of all available microtasks. HITs are therefore completed on a first-come-first-served basis by the required number of workers. Some workers, however, may decide to preview or even start working on a HIT and later decide to abandon it before its completion. Abandoned HITs may then be picked up by other workers willing to complete them. Note that when requesters run a batch of HITs, they receive answers from all the workers who complete the HITs but not from those who start and then return the HIT back to the platform. Such behavior of task abandonment is largely unstudied in current literature.

Addressing this gap, in this work we comprehensively studies the phenomenon of task abandonment in crowdsourcing. The aim of this paper is to understand abandonment, quantify its occurrence and analyze its impact on quality-related outcomes. To this end, we present the results of three different types of studies: i) a survey to understand the prevalence and causes of task abandonment in different paid microtask crowdsourcing platforms; ii) the analysis of task abandonment data collected 'in the wild' during a large-scale crowdsourcing relevance judgment project involving more than 7K HITs; and iii) controlled experiments to evaluate the effect of individual task properties on task abandonment. Our findings reveal that:

- The task abandonment phenomenon is very large, accounting for up to 164% abandoned tasks relative to finished tasks (i.e., for each submitted task we observed 1.64 abandoned tasks). In terms of distinct workers, in our experiment, we observed 1'157 workers completing our tasks, while the number of workers who started but then abandoned was 4'104. The total effort invested by workers in abandoned tasks accounts for 616 hours of work which are equivalent to about 3.5 months FTE. As a comparison, the total time devoted to completing and submitting tasks is 1693 hours, equivalent to 9.6 months FTE.
- As reported by workers, task abandonment is relatively more frequent for workers on F8 than on AMT. Most workers abandon tasks early, after making a quick assessment of the effort needed to complete it. Several workers on F8 however, abandon tasks after completing more than half of the expected work.

- *L. Han and G. Demartini are with the School of Information Tecnology and Electrical Engineering, University of Queensland, Australia.*
- *K. Roitero is with the Department of Mathematics, Computer Science, and Physics. University of Udine, Italy.*
- *U. Gadiraju is with the L3S Research Center, Leibniz Universität Hannover, Germany.*
- *C. Sarasua is with the Department of Informatics, University of Zurich, Switzerland.*
- *A. Checco is with the Information School, University of Sheffield, UK.*
- *E. Maddalena is with the Web and Internet Science (WAIS) group, University of Southampton. UK.*

Manuscript received; revised .

1. AMT: www.mturk.com, F8: www.figure-eight.com.

- The quality of work done by workers who abandon tasks is significantly lower than the quality of work done by those who complete tasks.
- The workers in abandoned tasks have shown decreasing engagement in terms of the quality of judgments and the time to write label justifications when they progress, compared to those in submitted tasks.
- Important factors that affect task abandonment are (listed in order of importance): the hourly wage, assessing the effort required to complete the task, and the quality checks used in the HIT design.
- There is a significant effect of task abandonment on the crowdsourced evaluation of Information Retrieval (IR) systems.

2 Related Work

Crowdsourcing has recently become a Web-based model that leverages distributed human intelligence to solve highly complex data problems [5]. As a result, a number of research projects across difference disciplines have adopted this methodology to tackle data problems that go beyond machine intelligence abilities [6], [7], [8]. One of the main challenges in applying crowdsourcing to data problems is the *quality* of crowdsourced data [9], [10]. Existing works have proposed new methods for crowdsourcing quality improvement, by focusing on both the answers provided by, and on the characteristics of crowd workers. Dow et al. [11] claimed that providing feedback to the workers can improve their performance as well as their motivation to be involved in additional tasks. Kazai et al. [12] found that the profile of the workers can significantly affect the accuracy of the tasks. Li et al. [13] proposed a crowd targeting framework to improve accuracy at the same or even lower budgetary cost. McDonnell et al. [14] showed how asking crowd workers for an explanation of the provided answer implicitly helps increasing the quality of the collected data. At the same time, the reliability of crowdsourced data has also been studied. For example, Ipeirotis et al. [15] provided a solution to distinguish true errors from individual's systematic biases and Eickhoff [16] looked at the effect of cognitive biases in the crowd on IR evaluation. Detecting malicious workers was also discussed in [17], [18]. In order to improve the reliability of the crowdsourced outcomes, Hung et al. [19] proposed using partial-agreement aggregation model to identify consensus among crowd workers in multi-label tasks. These works, however, are dealing with the quality of the data that crowd workers submit to the crowdsourcing platform. This lies in stark contrast to our focus in this paper; we shed light on the work that is carried out but *not* submitted by the workers as a result of task abandonment. We study behavioral data and responses collected from workers who abandon tasks, until they decide to abandon a given HIT.

Research on online user behavior aims at understanding the attention focus and interests of Web users. Some popular metrics of user engagement were proposed in the past few years. *Dwell time*, a simple page-level indicator that is adopted widely [20], [21], [22], can provide information about user engagement with web pages, but it is not able to capture detailed user behavior such as finding which HTML element attracts users most [23]. Chen et al. [24] proposed to consider both distance and distribution of mouse movement to predict user satisfaction in search pages.

In our work we collect and analyze behavioral data to study the task abandonment phenomenon. Low-level task interaction data has been previously used with a focus on predicting the accuracy of crowd work as an alternative to other quality assurance approaches such as gold questions. Early work on crowd worker behavioral data include [25] where Rzeszotarski et al. use behavioral traces to predict the quality of crowd worker answers in a supervised manner. More recently, in [26] Kazai et al. show how crowd behaviors can be compared to expert behaviors as a way to measure crowd work quality and to automatically detect low performing workers without the need for expensive gold questions. In [27] Goyal et al. also use behavioral data to predict worker accuracy and to better aggregate their answers on relevance judgments tasks. In this paper, we use similar log data over similar tasks but we juxtapose workers who do not complete the tasks with those who do, to understand task abandonment.

Abandonment is a frequently occurring online behavior defined as Web users who do not want to go any further with the activity and the content provided by the web pages they are visiting. As shown in [28], such phenomenon could occur either when users are satisfied with the content (good abandonment) like, for example, when relevant direct answers are provided in search engine results pages [29], or when they are dissatisfied with the information provided by pages they have visited (bad abandonment). Whenever a user's information need has already been satisfied or can no longer be fulfilled, abandonment is often observed. Abandonment in crowdsourcing has mainly been studied from a batch point of view (i.e., how many HITs of the same type, workers are completing in a sequence). For example, methods to extend crowd work sessions have been proposed and evaluated in [30]. In comparison, we look at single task abandonment rather than dropouts from batches, thereby focusing on work completed but not rewarded. There is limited research aiming at understanding the consequences of user abandoning HITs in crowdsourcing marketplaces. Some existing studies on satisfaction have tried to analyze user interaction from different dimensions to improve their search experience, e.g., [22], [24], [31], [32]. Differently to them, we focus on crowdsourcing workers who give up before completing their HITs aiming at understanding task abandonment on crowdsourcing platforms by examining their interaction and behavior while working on tasks.

3 Study I: Prevalence and Causes of Task Abandonment

To understand the prevalence of the task abandonment phenomenon among crowd workers, we first ran a survey on two popular paid micro-task crowdsourcing platforms: Amazon MTurk (AMT) and FigureEight (F8). We collected responses from 100 distinct workers on each platform and carried out a combination of quantitative and qualitative analyses to understand the perceived factors that influence task abandonment in crowdsourcing.

3.1 Survey Design and Findings

3.1.1 Survey Design

We first asked workers to respond to some general background questions regarding demographics and their experience. Next, we collected responses about the frequency with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2948168, IEEE Transactions on Knowledge and Data Engineering

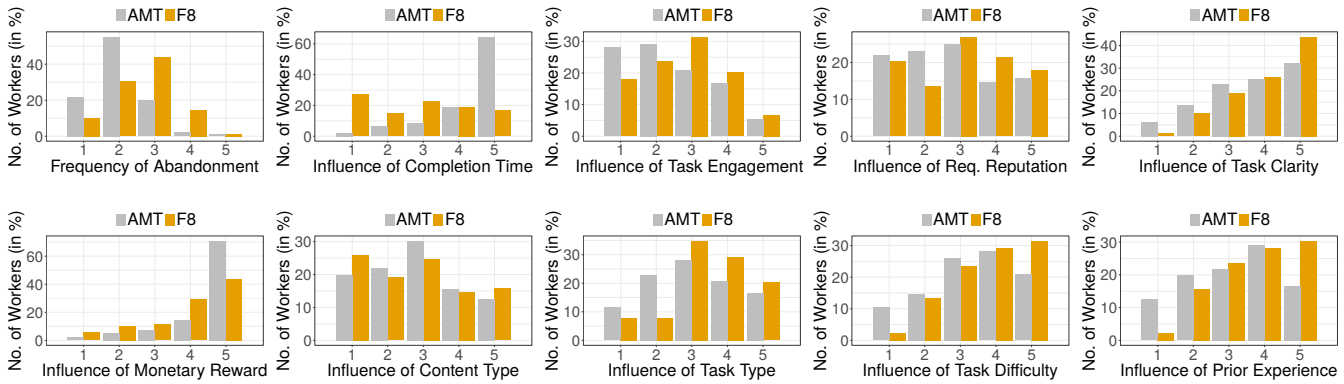


Fig. 1: (*Top-left*) Frequency of task abandonment as perceived by workers on Amazon MTurk (AMT) and FigureEight (F8), and (*remaining sub-figures*) influence of various factors that affect task abandonment on the two platforms.

which they abandoned tasks after having started them, on a 5-point Likert scale from *1: Almost Never* to *5: Almost Always*. We then asked workers the extent to which they believed that a variety of factors typically influenced their decision to abandon a task, on a 5-point Likert scale from *1: No Influence* to *5: High Influence*. These factors included task difficulty, completion time, monetary reward, requester reputation, task type, prior experience of workers, task clarity, content type (e.g., boring, explicit or disturbing), and a lack of engagement. In an open-ended text box, we also encouraged workers to reveal other factors that potentially influence task abandonment in their experience. Workers were also asked about the types of tasks [33] they abandoned most often and why they did so.

3.1.2 Frequency of Task Abandonment

As shown in Figure 1, we found that a significant fraction of workers on F8 and AMT tend to abandon tasks frequently. Nearly 60% of the F8 workers we surveyed claimed to abandon tasks with at least a level of 3 on the 5-point scale in comparison to over 22% of AMT workers. Using a two-tailed T-test, we found that F8 workers (M=2.66, SD=.89) claimed to abandon tasks significantly more frequently than AMT workers (M=2.05, SD=.77); $t(184)$=24.90, $p < .001$. Due to this reason we focus our data-driven analysis presented in Section 4 on the F8 platform.

3.1.3 Progress before Abandonment

We found that most workers on both F8 and AMT, abandon tasks either after previewing them and reading the instructions or after completing less than half of the task (see Table 1). In comparison to AMT, a greater fraction of workers on F8 abandon tasks after completing either more than half or the entire task. A relatively small fraction of workers on both platforms claimed that they typically do not abandon tasks.

3.1.4 Influence of Different Factors on Task Abandonment

We analyzed different factors that influence worker decisions to abandon tasks on F8 and AMT. Our findings are presented in Figure 1. In contrast to about 17% of F8 workers, nearly 65% of AMT workers perceived *task completion time* as being highly influential in their abandoning of tasks. Similarly, about 71% of AMT workers perceived the *monetary reward* as being highly influential in their task abandonment in comparison to 44% of F8 workers. Both F8 and AMT workers claimed a mediocre influence of *task engagement*, *requester reputation*, and *content type* on their task abandonment. F8 workers considered *task clarity*, *task difficulty*, *task type* and *prior experience* to be more influential in task abandonment than AMT workers (who also found these factors to be fairly influential). Table 2 presents a ranked list of these factors according to their level of perceived influence on task abandonment on F8 and AMT.

TABLE 1: Progress in tasks by workers before abandonment on F8 in comparison to AMT.

Progress	% Workers (F8)	% Workers (AMT)
More than half of the task	17.98	3.13
Entire task	11.24	2.08
Preview and read instructions	35.96	30.21
Less than half of the task	28.09	54.17
Typically do not abandon tasks	5.62	9.38
Other	1.12	1.04

TABLE 2: The extent to which various factors influence task abandonment on F8 and AMT (average on a 1-5 scale).

Factor	F8	AMT
Task Clarity	4.01 ± 1.07	3.64 ± 1.23
Monetary Reward	3.96 ± 1.21	4.47 ± 0.98
Task Difficulty	3.74 ± 1.11	3.34 ± 1.25
Prior Experience	3.69 ± 1.13	3.18 ± 1.27
Task Type	3.46 ± 1.13	3.08 ± 1.25
Requester Reputation	3.03 ± 1.37	2.79 ± 1.35
Task Completion Time	2.84 ± 1.44	4.38 ± 1.01
Content Type	2.75 ± 1.39	2.79 ± 1.27

3.2 Worker Remarks

We analyzed the open-ended responses from F8 and AMT workers regarding why they tend to abandon tasks by using an iterative coding process [34], [35]. In this process, we manually went through each open-ended response and categorized the theme(s) of the response. For example, a worker on AMT responded with '*The task is either too complicated or the pay figures to be too low*' (sic). This response was categorized into the themes of task difficulty and reward. We iteratively created new themes as they emerged from worker responses, and re-coded all responses to ensure accurate categorization. The main themes that were identified

as a result of our analysis are summarized below. Several workers on F8 and AMT described multiple factors playing influential roles towards task abandonment. Note that the following analysis is based on the open-ended responses alone, and does not include the responses gathered on Likert-type scales and discussed in Section 3.1.

1) *Time Constraints vs. Requirement.* Workers are constrained to complete tasks within 30 minutes by default on F8. Workers can perceive this as being restrictive, depending on the task design and the number of tasks available in the given batch. Workers abandon tasks when they believe they cannot complete tasks within this stipulated time limit. 10.64% of F8 workers cited task completion time as a factor that contributes to task abandonment in their responses. In contrast, 62.5% of the AMT workers cited completion time as a factor despite not having a default constraint on completion time. In case of AMT, time limits are enforced by the requesters. As opposed to F8 workers, AMT workers mentioned that they abandon tasks that require a lot of time for completion.
2) *Subjective Tasks.* Workers avoid subjective tasks due to the uncertainty associated with how their responses may be evaluated by the requesters. Nearly 32% of the F8 workers cited the subjective nature of tasks and the corresponding doubt over their accuracy in such tasks as being influential in task abandonment. In contrast, only 1% of AMT workers acknowledged task subjectivity as an influential factor.
3) *Poor Instructions.* Over 40% of the F8 workers and 24% of AMT workers referred to the poor quality of instructions that typically influence their decisions to abandon tasks.
4) *Maintaining Accuracy.* Workers aim to maintain a high level of accuracy in tasks in order to build a good reputation, giving themselves the best opportunity to qualify for and complete more future tasks. It is well known that several crowd workers turn to crowdsourcing microtasks as a means to earn their primary source of income [36], [37]. Nearly 28% of F8 workers and over 5% of AMT workers referred to potential threats to their overall accuracy as being influential in task abandonment.
5) *Monetary Reward.* Nearly 30% of the F8 workers and 62.5% of AMT workers cited poor pay with respect to the expected work as a factor that results in their abandoning tasks. Since workers aim to maximize their earnings, tasks that pay little for relatively more effort from the workers dissuade workers from participating in them. Nearly 14% of the AMT workers directly mentioned such disproportionate 'effort' in their responses.
6) *Fairness.* Almost 20% of the F8 workers and 21% of AMT workers described tasks that lack a sense of fairness with respect to several factors (either pay, time, or in the way they are evaluated), as influencing their decisions to abandon such tasks.
7) *Task Difficulty.* Just over 23% of F8 workers and over 10% of AMT workers indicated that task difficulty influenced their decisions to abandon tasks.
8) *Language Proficiency.* Just under 11% of F8 workers claimed that they abandon tasks when they feel that the language requirements are too high with respect to their proficiency. In stark contrast, not a single AMT worker referred to language proficiency as an influential factor.
9) *Other Factors.* A small percentage of F8 workers (under 7%) and AMT workers (nearly 8%) referred to different aspects that they believe to influence their decisions to abandon tasks; complicated workflows of tasks involving multiple phases, interestingness of tasks, and the opinion of other contributors (e.g., in workers' forums) about the given tasks.

3.3 Discussion

Our findings from this study shed a light on the different factors that influence task abandonment in crowdsourcing tasks to varying extents on F8 and AMT. Workers on both platforms abandon tasks frequently enough to affect market dynamics and make this phenomenon worthy to investigate. Workers on AMT abandon tasks primarily due to the disproportionate monetary reward with respect to the expected amount of time for task completion. In contrast, workers on F8 primarily abandon tasks due to a lack of clarity, associated reward and high perceived task difficulty. Workers on F8 perceive task abandonment to be more frequent, and they tend to abandon tasks after having progressed to greater lengths (more than half of the task, entire task). Due to this, we investigate task abandonment further in a large-scale crowdsourced relevance judgment experiment on the F8 platform.

4 Study II: Abandonment in the Wild

In this section we present findings from a large-scale relevance judgment task on F8, during which task abandonment logs were collected. We address three main research questions here.

RQ1: How well do workers perform in abandoned HITs when compared to those in completed HITs?

RQ2: How much work do workers complete in abandoned HITs before their abandonment?

RQ3: How do behavioral patterns exhibited in providing judgment justifications in abandoned HITs differ from those in submitted HITs?

4.1 Crowdsourcing Data Collection

4.1.1 Task Design

We ran a large relevance assessment experiment following the design used by [38] and [39]. The HITs are presented to workers with a topic and eight documents taken from the TREC-8 ad hoc collection [40]. Figure 2 shows the interface of the task. The topic was fixed for each HIT whereas the documents were arranged in eight sequential pages that workers can visit backwards and forward. Workers were asked to judge the relevance of each document with respect to the given topic on a four-level scale (*not-relevant*, *marginally relevant*, *relevant*, or *highly relevant*). Additionally, for each relevance assessment, a textual justification was required [14]. We implemented three quality checks: (i) an initial test question to ensure the worker understood the topic; (ii) a check that workers spent at least 20 seconds in at least six of the eight documents, and (iii) two of the eight documents were gold standard editorial judgments by [41] manually selected by experts to have one of them clearly not relevant to the topic (N) and the other one clearly relevant (H). We checked if workers judged these documents consistently ($H > N$). These three checks are

TABLE 3: Classification of workers by task completion status in topic x.

Groups		Description of workers
S		submitted a HIT for topic x and have not abandoned HITs for topic y ($\forall y \neq x$)
A		abandoned HITs for topic x and have not submitted HITs for topic y ($\forall y \neq x$)
SA	$SA^{(S)}$	submitted a HIT for topic x but abandoned HITs for topic y ($\exists y \neq x$)
	$SA^{(A)}$	abandoned HITs for topic x but have submitted HITs for topic y ($\exists y \neq x$)

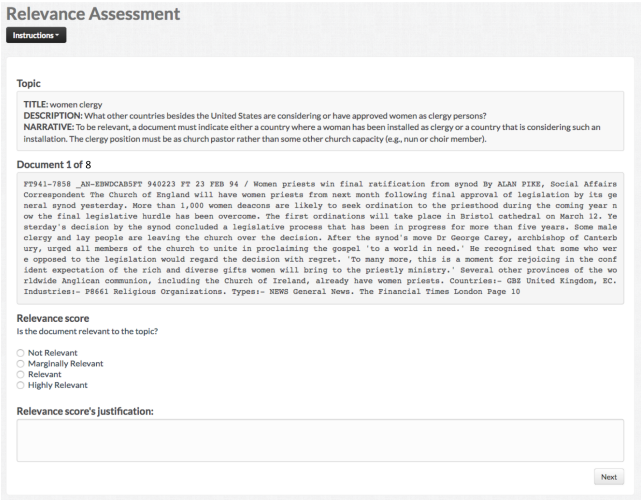


Fig. 2: Interface of the relevance assessment task.

performed at the end of the document sequence. On failing any of these checks, workers were allowed to go back and change their judgments up to three times. The time spent evaluating each document is cumulated across different attempts to reach the required 20 seconds.

Overall, we collected judgments for 4'269 documents[2] over 18 topics and 7'067 HITs. At the same time, we observed 11'563 HITs that were abandoned.

4.1.2 Logging Abandonment

Crowdsourcing platforms do not allow obtaining information about tasks which have not been correctly completed and submitted by workers. This restriction leads to a loss of the work done before task abandonment. Since this paper aims at studying task abandonment, we implemented a solution to bypass such limitation by logging each high level action performed by workers in the task. To make logging possible, we set up an external server to receive requests coming from JavaScript code embedded in the HIT. We log the following high-level actions: task begins; worker clicks the informed consent button; worker answers the initial topic understanding question and the first document is shown; worker changes page (backward or forward); worker provides a relevance judgment; one or more quality checks are failed; all quality checks are passed and the task ends successfully. Additionally, we collected the browser's HTTP user agent string[3].

2. Among these 4'269 documents, 3'881 of them are assessed by TREC [40] on a binary relevance scale, and 805 of them are re-assessed by Sormunen [41] on a 4-level ordinal scale.

3. Crowd workers were asked to read and accept an informed consent document before starting the HIT where we explain them about such behavioral action logging.

4.2 Methodology

Using the task design and logging infrastructure described above, we collected action logs and relevance judgments on a per-topic basis. For a specific topic, *submission* happens when a worker completes a HIT in this topic, while *abandonment* is defined as a worker starting tasks in this topic but abandoning all of them without a submission. Note that workers may first abandon HITs and then submit a task in the same topic, and thus such case would be classified as "submission" by our definition. Because we aim to investigate the difference between task abandonment and submission in various topics, previous abandonment in the same topic plays a training role to get used to both the topic and task. Once they submit the task they are not allowed to start a task for the same topic any more. Table 3 illustrates how we classify submission and abandonment in terms of workers for each topic. We collected data from three populations of crowd workers: (i) those who submitted HITs in all topics that they participated in (group *S*), (ii) those who started but never submitted HITs in any topics (group *A*), and (iii) those who abandoned tasks in one topic but had submissions in other topics (group *SA*).

By the classification described above, we are able to distinguish between those workers who have never failed in submitting HITs for any topics (*S* workers) from those who only abandoned in all topics that they started (*A* workers). Moreover, in order to investigate the behavioral consistency and to understand the difference of performance exhibited by *SA* workers in the tasks they abandoned compared to those they submitted, we split this group by their task completion status in each topic. If the workers have a submission in the topic, they are regarded as $SA^{(S)}$ workers in this topic. On the contrary, if there is no submission observed in the topic, they are considered as $SA^{(A)}$ workers (see Table 3). Note that $SA^{(S)}$ and $SA^{(A)}$ workers are overlapping across different topics, since one $SA^{(A)}$ worker should have submitted at least one HIT in another topic and thus be identified as $SA^{(S)}$ worker in that topic, and vice versa[4].

We examined our dataset from four perspectives: (i) the quality of the judgments performed by all workers, to answer RQ1; (ii) how many documents they judged and (iii) how much time they spent in the HITs, to answer RQ2; (iv) the content reuse (i.e., `copy` and `paste`) estimated by similarity of the text in justifications, to answer RQ3.

To measure the quality of judgments provided by workers, we compare them to ground truth editorial assessments on a 4-level scale by Sormunen [41]. Thus, we compare crowd worker judgments from all groups (i.e., *S*, *A*, $SA^{(S)}$ and $SA^{(A)}$), with judgments from human experts, by means

4. The classification of the workers is different from the definition of S and A workers in [42], where S equals to the union of S and $SA^{(S)}$ in this paper, and A equals to the union of A and $SA^{(A)}$ here.

TABLE 4: Statistics of the collected data in terms of HITs and unique workers.

Groups	S	A	$SA^{(S)}$	$SA^{(A)}$	Total
#HITs	2190	7638	4877	3925	18630 [5]
#Workers	345	3292	812	812	4449 [6]

of agreement measures. To measure agreement between crowd workers and experts we use Krippendorff's Alpha coefficient [43] owing to its ability to adapt to missing values and different number of judgments. This measurement assumes values from -1 (complete disagreement) through 0 (agreement equivalent to random evaluations) to 1 (complete agreement). Since human expert judgments do not cover all documents that we used in the experiments and the workers in group A and $SA^{(A)}$ may have provided fewer labels, we only measure agreement over the subset of eight documents in each HIT for which both crowd workers and experts judgments are available. For each HIT we compute the quality of the judgments contributed by all workers in group S, A, $SA^{(S)}$ and $SA^{(A)}$. We then average agreement scores in each group across HITs for the same topic.

To understand the behavioral patterns in terms of content reuse (i.e., `copy` and `paste`) exhibited while providing justifications, we compute the similarity of the justifications for the current judgment to two sources of text: (i) the document presented in the current task page, and (ii) the topic shown throughout the entire task. We adopt the rationale of the Ratcliff-Obershelp similarity metric [44] to compute the longest common sub-strings, and define the similarity as the number of matching characters divided by the number of total characters in current justification. With this definition, content reuse is able to be estimated by means of similarity, which goes from 0 (typing everything from scratch) to 1 (complete content reuse).

4.3 Results

During the experiments, we collected in total 7'067 submitted HITs, completed by 1'157 unique workers, since we allow them to participate in multiple topics (but only one HIT per topic). Meanwhile, we observed 4'104 unique workers who abandoned 11'563 HITs altogether (including A and $SA^{(A)}$ group). Table 4 shows the statistics in terms of tasks and unique workers with a breakdown of different groups. Again, the workers in group $SA^{(S)}$ overlap with $SA^{(A)}$ across different topics, as explained in Section 4.2.

4.3.1 Quality

The average α agreement with experts for the submitted HITs (completed by S and $SA^{(S)}$ workers) is 0.74, while it is 0.33 for the abandoned HITs (by A and $SA^{(A)}$ group). Figure 3 shows the differences of α values between the submitted and abandoned HITs over topics, where topics are sorted in descending order of average α value for

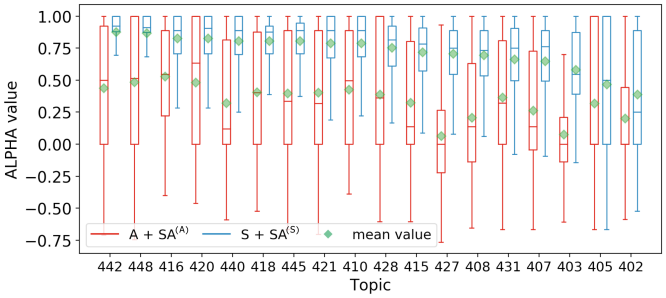


Fig. 3: Judgment quality over topics comparing submitted and abandoned HITs. Topics are sorted by decreasing mean value for submitted tasks.

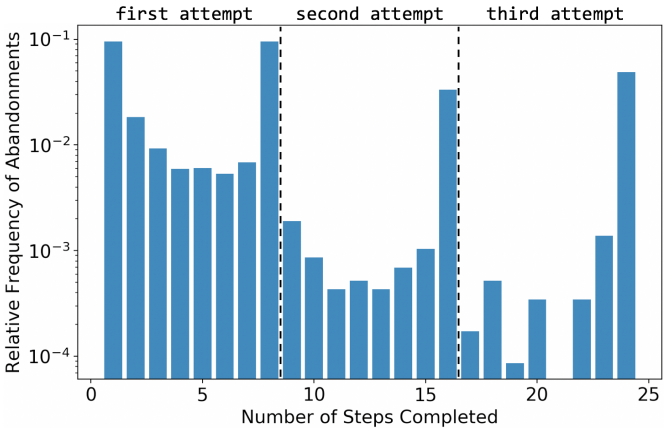


Fig. 4: Relative frequency of abandonment (log scale) over the number of completed judgments.

submitted tasks. It is evident that the average judgment quality for the abadoned HITs group is lower than the quality for the submitted HITs across all different topics. We can also observe that the standard deviation of α values for the abandoned HITs is larger, compared to submitted HITs. This shows that there exists a higher uncertainty of the judgment quality in abandoned tasks. The highest average α value across topics for the abandoned HITs is 0.53. Using the Wilcoxon signed-rank test[7] to compare the quality of the abandoned and submitted tasks we found that the difference is statistically significant ($p < 0.05$) over all topics.

4.3.2 Task Engagement and Abandonment Rate

Since we used eight documents in each HIT and allowed workers to start the same tasks up to three times if they failed the quality checks before completing their submission, the maximum number of questions that a worker might have seen is 24. Workers could abandon the task at any point when answering these 8 to 24 questions. We define each judgment as a *step* in the HIT. Before Step 1, each worker had to click a 'start button' (*Step* -1) and was consequently presented with the task instructions (*Step 0*).

Among the 11'563 abandonments observed, in two-thirds of the cases workers abandoned the task without any engagement with documents (i.e., either Step -1 or

5. In the same topic, if a worker first abandoned tasks and then successfully submitted a task, we consider such abandonments prior to the submission as 'practice' to get familiar with the topic and task. Therefore, such attempts are regarded as 'one' task in counting HITs for S and $SA^{(S)}$ group.

6. The overlapping workers in group $SA^{(S)}$ and $SA^{(A)}$ are removed in counting.

7. We use Wilcoxon signed-rank test to determine the statistical significance, since the assumption of normality is not satisfied.

TABLE 5: The absolute number and percentages of abandonments observed after each step with a topic breakdown.

Topic	Step −1	Step 0	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 8+	Total
402	6 (0.9%)	491 (70.5%)	68 (9.8%)	21 (3%)	5 (0.7%)	2 (0.3%)	3 (0.4%)	3 (0.4%)	1 (0.1%)	57 (8.2%)	39 (5.6%)	696 (100%)
403	0 (0%)	117 (39.3%)	18 (6%)	7 (2.3%)	1 (0.3%)	0 (0%)	4 (1.3%)	2 (0.7%)	1 (0.3%)	33 (11.1%)	115 (38.6%)	298 (100%)
405	1 (0.4%)	122 (54.5%)	13 (5.8%)	10 (4.5%)	2 (0.9%)	6 (2.7%)	2 (0.9%)	1 (0.4%)	11 (4.9%)	26 (11.6%)	30 (13.4%)	224 (100%)
407	3 (0.8%)	201 (54.2%)	27 (7.3%)	7 (1.9%)	3 (0.8%)	0 (0%)	2 (0.5%)	1 (0.3%)	3 (0.8%)	46 (12.4%)	78 (21%)	371 (100%)
408	0 (0%)	100 (51.5%)	15 (7.7%)	3 (1.5%)	1 (0.5%)	0 (0%)	3 (1.5%)	8 (4.1%)	2 (1%)	26 (13.4%)	36 (18.6%)	194 (100%)
410	3 (0.8%)	287 (71.8%)	29 (7.3%)	3 (0.8%)	2 (0.5%)	1 (0.3%)	7 (1.8%)	9 (2.3%)	2 (0.5%)	32 (8%)	25 (6.3%)	400 (100%)
415	2 (0.8%)	148 (59.7%)	25 (10.1%)	6 (2.4%)	3 (1.2%)	3 (1.2%)	3 (1.2%)	1 (0.4%)	7 (2.8%)	21 (8.5%)	29 (11.7%)	248 (100%)
416	4 (1.7%)	156 (67%)	12 (5.2%)	3 (1.3%)	1 (0.4%)	0 (0%)	3 (1.3%)	0 (0%)	0 (0%)	32 (13.7%)	22 (9.4%)	233 (100%)
418	2 (0.7%)	181 (66.5%)	25 (9.2%)	6 (2.2%)	2 (0.7%)	0 (0%)	0 (0%)	3 (1.1%)	5 (1.8%)	17 (6.3%)	31 (11.4%)	272 (100%)
420	2 (1%)	117 (60.9%)	18 (9.4%)	2 (1%)	2 (1%)	3 (1.6%)	0 (0%)	1 (0.5%)	4 (2.1%)	21 (10.9%)	22 (11.5%)	192 (100%)
421	4 (0.5%)	555 (74.2%)	61 (8.2%)	17 (2.3%)	7 (0.9%)	3 (0.4%)	0 (0%)	1 (0.1%)	2 (0.3%)	48 (6.4%)	50 (6.7%)	748 (100%)
427	1 (0.1%)	389 (50.7%)	45 (5.9%)	8 (1%)	15 (2%)	3 (0.4%)	5 (0.7%)	4 (0.5%)	8 (1%)	120 (15.6%)	170 (22.1%)	768 (100%)
428	15 (1.3%)	826 (69.9%)	135 (11.4%)	33 (2.8%)	8 (0.7%)	17 (1.4%)	14 (1.2%)	3 (0.3%)	10 (0.8%)	73 (6.2%)	47 (4%)	1181 (100%)
431	7 (1.7%)	278 (65.7%)	32 (7.6%)	7 (1.7%)	7 (1.7%)	2 (0.5%)	4 (0.9%)	3 (0.7%)	2 (0.5%)	44 (10.4%)	37 (8.7%)	423 (100%)
440	4 (0.7%)	364 (66.4%)	58 (10.6%)	10 (1.8%)	7 (1.3%)	5 (0.9%)	0 (0%)	2 (0.4%)	4 (0.7%)	60 (10.9%)	34 (6.2%)	548 (100%)
442	38 (1.8%)	1257 (67.9%)	161 (8.7%)	27 (1.5%)	16 (0.9%)	11 (0.6%)	7 (0.4%)	7 (0.4%)	2 (0.1%)	171 (9.2%)	157 (8.5%)	1850 (100%)
445	14 (0.8%)	1166 (69.4%)	213 (12.7%)	29 (1.7%)	16 (1%)	6 (0.4%)	7 (0.4%)	7 (0.4%)	10 (0.6%)	153 (9.1%)	58 (3.5%)	1679 (100%)
448	5 (0.4%)	841 (67.9%)	150 (12.1%)	14 (1.1%)	9 (0.7%)	7 (0.6%)	6 (0.5%)	6 (0.5%)	5 (0.4%)	126 (10.2%)	69 (5.6%)	1238 (100%)
Total	107 (0.9%)	7596 (65.7%)	1105 (9.6%)	213 (1.8%)	107 (0.9%)	69 (0.6%)	70 (0.6%)	62 (0.5%)	79 (0.7%)	1106 (9.6%)	1049 (9.1%)	11563 (100%)

0). While the overall volume of observed abandonment is massive, most of it happens very early in the HIT. This shows that many workers read the instructions or preview the task itself to make a quick assessment of the effort required to complete it in light of the allocated reward, deciding whether or not to invest their time in it. This is consistent with the open-ended responses workers provided in Study I, regarding why they abandon tasks.

Figure 4 presents the distribution of abandonment for the 3'860 abandoned tasks in which workers performed at least one relevance judgment, showing the ratio of abandonments after a given step to the whole abandoned HITs. We can see that the abandonment happening after Step 1 and 8 is the largest. These two steps represent workers abandoning after judging the first document and those abandoning at the end of the HIT because of not passing the first quality checks. The second two largest abandonment points happen after Step 24 and 16. This shows the presence of two important points of abandonment, i.e., after the *first* or *last* question in the HIT[8].

Table 5 shows the absolute number of abandonments observed after each step over different topics, together with percentages relative to the overall number of abandonments observed in the topic. We merged the steps from 9 to 24 and used *Step 8+* to indicate abandonments happening after the first full judging attempt. We can see that, on average, 67% of all abandonments happen before judging the first document (Step 0) and 76% up to the first document judgment (Step 1). An additional 10% of abandonments happen after judging all 8 documents (Step 8) because of not passing the quality checks. Another observation we can make is that abandonment behavior may vary across topics. For example, Topic 403 has more than one third of workers reaching Step 8+ showing how judging documents for this topic was particularly difficult. This is in line with other research where documents for this topic have been judged by means of crowdsourcing (e.g., Fig. 6 in [45]).

Abandonments after the first judgment (Step 1) may be caused by workers' assessment of the task effort/reward ratio. If workers decide to continue the task after the first document, however, they typically aim to complete and submit the entire HIT. The number of HITs abandoned after Step 1 and 8 is 1'105 and 1'106 respectively, while in another

8. Note that in our task design we have eight documents to judge and we allow workers to have up to three attempts if they do not pass the quality checks. Therefore, abandonment that happens after Step 8, 16 and 24 represents abandoning the task after answering the *last* question in the first, second and third full attempt, respectively.

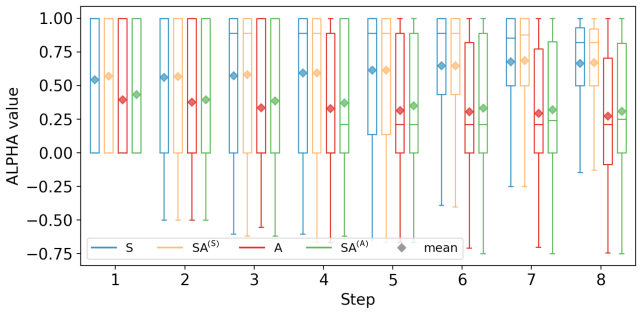


Fig. 5: Judgment quality over steps for S, $SA^{(S)}$, A and $SA^{(A)}$ workers.

1'049 HITs (9.1% of abandoned HITs) workers performed the same judgments again (Step 8+). Among the workers in group S and $SA^{(S)}$, in 1'366 HITs (19.3% of submitted HITs) workers reached step 8+ before submitting their judgments. Workers who abandoned after Step 24 have reached the maximum number of attempts allowed by our HIT design.

4.3.3 Quality Over Steps

Next, we look at how the judgment quality changes with the progress in the HIT. To this end, we compared the quality across the judgments provided by S, A, $SA^{(S)}$ and $SA^{(A)}$ workers up to each step in the HIT[9]. We focus on the quality evolution in the first attempt on the HIT if they failed the quality checks and started the task again. For those who ever went back to revise their judgments in the first attempt, we use their final version of judgments for each document to compare the quality across different groups. Due to the limited number of judgments performed in abandoned HITs, for a given step, we only consider the HITs where all the judgments up to this step are available.

Figure 5 shows the judgment quality over steps for all groups. We can see that, on average, workers who submitted (group S and $SA^{(S)}$) constantly provide higher quality labels than workers who abandoned (group A and $SA^{(A)}$). For those who submitted, the average quality steadily rises from Step 1 to 7, which indicates a positive learning effect; workers get used to the task and provide better judgments as they progress through the task. For those who abandoned, the average quality continuously drops from the beginning to the end throughout the task, showing a

9. We consider the quality of all available workers up to each given step for A and $SA^{(A)}$ group (i.e., all those whose abandonment points are at and beyond the given step).

decrease in engagement as they progress. Interestingly, the workers in the $SA^{(S)}$ group demonstrated opposite quality in performance to those in the $SA^{(A)}$ group. Actually, they are the same group of workers as a whole, and therefore, this shows that there are no absolute good or bad workers but rather that they are not good at judging documents for certain topics. There is no significant difference between the quality of the judgments given by *S* and $SA^{(S)}$ workers, nor between *A* and $SA^{(A)}$ workers. But the quality of judgments by *A* and $SA^{(A)}$ workers differs from that by *S* and $SA^{(S)}$ workers significantly (*t*-test $p < 0.05$) at each step.

4.3.4 Content Reuse in Justifications

In order to capture the behavioral patterns of content reuse when workers are providing justifications, we look at the similarity of the content (defined in Section 4.2), and use the similarity score to estimate how much text comes from `copy/paste` actions by means of the number of characters. Figure 6 shows the similarity of the text in justifications to both topic (6a) and documents (6b) that are presented in the HIT. It is evident that the similarity of the justifications given by those who submitted HITs (group *S* and $SA^{(S)}$) is maintained at a stable level, showing that their working style in terms of content reuse patterns remains unchanged throughout the task. On the contrary, *A* workers exhibit similar `copy/paste` patterns to that of the submitting workers at Step 1 but they tend to reuse less content gradually when they move on. From the fourth judgment, the reused text from topic and documents becomes significantly less than *S* and $SA^{(S)}$ workers (*t*-test $p < 0.05$) till the end of the task. This explains why *A* workers spend remarkably more time at the beginning of the task and then lose engagement step after step (shown in Section 4.3.5 below). On the other hand, low similarity of the justifications provided by *A* workers to the documents (Figure 6b) implies that their justifications contain more content that does not exist in the documents. This may bear the implication that such workers, for example, are more likely to be spammers and thus cannot pass the quality checks in the end. The justifications given by workers in $SA^{(A)}$ group are constantly less similar to the topic than those in $SA^{(S)}$ group. However, there is no significant difference in content reuse from documents performed by workers in both groups. In fact, these two groups of workers are fully overlapped across different topics and they show similar patterns of content reuse from documents. But the fact that the similarity of their justifications to the topic stays at a low level across all the four groups as they progress in the tasks is indicative of deviation from the topic content. Because of this, they are less likely to pass the quality checks by the end of the tasks and abandonment has a higher chance to happen. Overall, the level of content reuse from documents among *A* workers is the lowest compared to the workers in the other three groups.

Considering that the text of documents is much longer than the topic text as a source of content reuse, we focus on `copy/paste` actions on document text presented in the HITs. For each justification, we estimate the length of copied content by the longest common sub-strings between the justification and the document (defined in Section 4.2), and consider the remainder of the justification as newly typed content after `copy/paste`. Figure 7 shows the length of copied and newly typed content. It is clear that the distribution of the justification length is heavily skewed in all groups[10], caused by a small number of workers writing extremely long justifications.

Compared to the workers who submitted HITs (group S and $SA^{(S)}$), those who abandoned HITs (group A and $SA^{(A)}$) tend to write shorter justification text (both copied from documents and written by themselves), showing that they have a lower level of engagement in the tasks. Workers in the S and $SA^{(S)}$ groups provide longer justifications for highly-relevant documents as compared to non-relevant ones, which is explained by the higher content reuse in high-relevant documents and that the workers are inspired by such content to write more. For A workers there is no significant difference between writing justifications after a `copy/paste` action for high- and non-relevant documents. On average (based on median values), workers in all groups have typed longer text than copied content. This is also confirmed by the calculation of content similarity (less than 0.5), showing that the longest common sub-strings of justification to document is less than a half (see Figure 6b). Using Mann-Whitney *U* test[11] [46], we conclude that all the differences mentioned above are statistically significant ($p < 0.01$).

4.3.5 Time to Judge

To understand how much time workers spent on each judgment, we used the timestamp of each logged action as provided by worker browsers. We analyzed the overall time spent on each HIT: (i) time to read instructions, and (ii) time to judge documents (including providing judgments and writing justifications). Figure 8 shows the distribution of workers in submitted (group *S* and $SA^{(S)}$) and abandoned (group *A* and $SA^{(A)}$) HITs with respect to the time spent on reading instructions (left) and judging documents (right). Both distributions are long tailed with many workers spending little time on instructions and judgments. We can see that 99% of the workers spent less than 1'200 seconds (or 20 minutes) reading instructions and less than 1% of each group population took more than 7 minutes to judge a document.

The distributions of instruction reading time for both groups are very similar. This tells us that all workers approach the task in a similar way regardless of submission or abandonment. On the contrary, the workers show differences in the average time spent judging each document when they abandon tasks compared to workers submitting tasks. Those who abandoned tend to spend less time judging documents (which also influences their judgment quality as shown below). The proportion of workers spending less than half a minute to judge a document is greater for the workers in abandoned HITs (40.39%) as compared to those in the submitted HITs (15.02%). The proportion of workers whose judging time is from 0.5 to 3.5 minutes in the abandoning group is lower (55.17%) than that of the submitting group (84.28%). This observation reveals that although workers have similar instruction reading patterns, the time devoted to judging documents in abandoned HITs is different from that in submitted HITs.

10. For this reason, we use median value instead of arithmetical mean for the analysis of content reuse.

11. We use Mann-Whitney *U* test because of the existence of extreme values contributed by a few workers (e.g., a few workers providing extremely long text as their justifications), and thus the distribution is not interval-scaled.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2948168, IEEE Transactions on Knowledge and Data Engineering

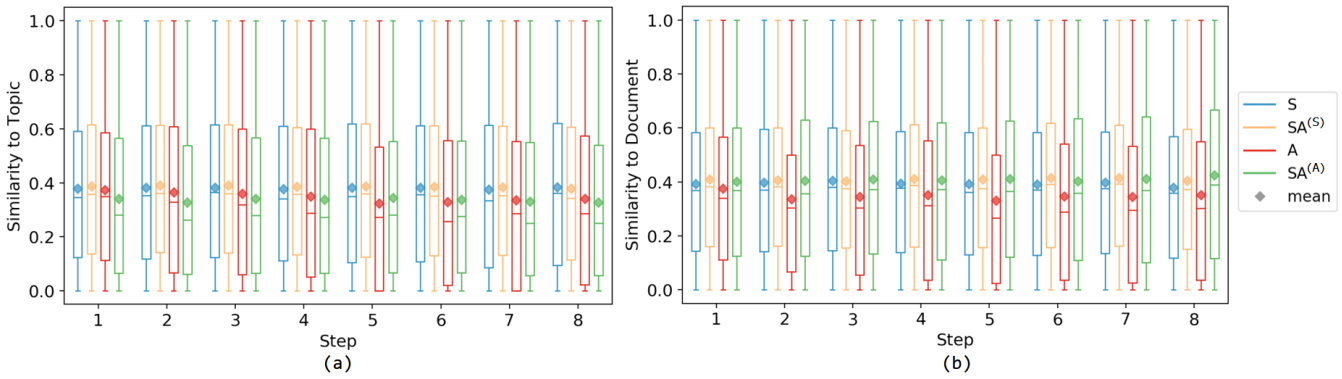


Fig. 6: Similarity of justifications to topic (a) and documents (b) over steps for S, $SA^{(S)}$, A and $SA^{(A)}$ workers.

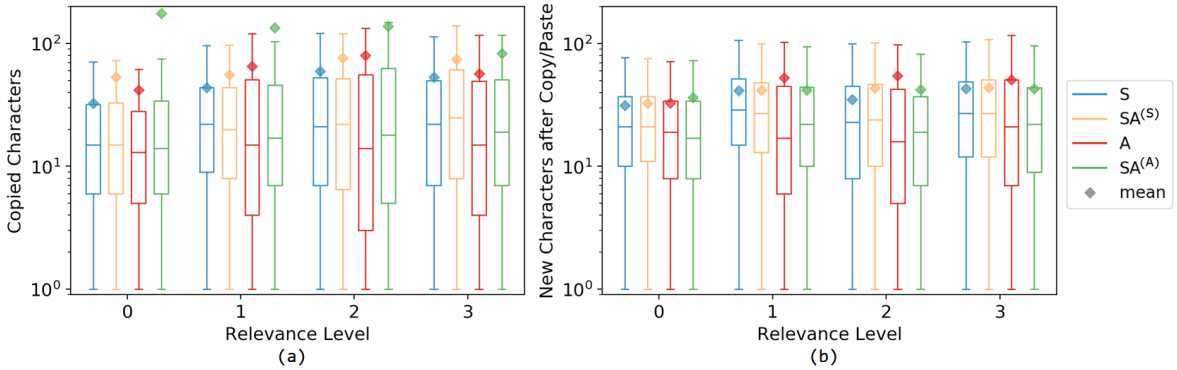


Fig. 7: Length (log scale) of text in copied content from documents (a) and newly typed justifications after `copy/paste` (b) with respect to relevance levels given by workers.

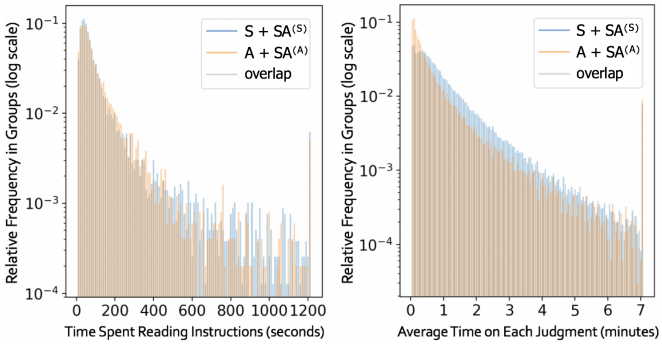


Fig. 8: Time spent reading instructions (left) and judging each document (right) in submitted and abandoned HITs.

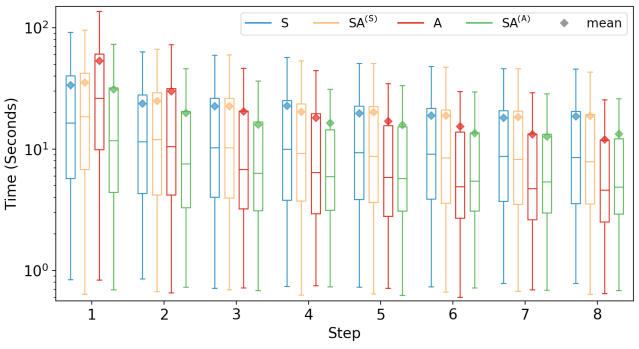


Fig. 9: Time (log scale) spent in writing justification over steps for S, $SA^{(S)}$, A and $SA^{(A)}$ workers.

We additionally look at the time that the workers spent in writing justifications for all groups (i.e., S, A, $SA^{(S)}$ and $SA^{(A)}$), aiming at understanding how the time changes when they progress throughout the HIT. Figure 9 shows the evolution of time devoted to writing justifications over steps. For a given step, we take the time for all workers who are available at this step in each group. The result reveals that all workers spend less time in providing justifications as they progress in the HIT, showing that they get used to the task when they do more and have developed some strategies to do the task faster. The time devoted by workers in groups S and $SA^{(S)}$, however, does not fluctuate too much, as compared to groups A and $SA^{(A)}$. The workers in group A spend 53 seconds on average to write justifications in Step 1, while it is 34, 35, 31 seconds for groups S, $SA^{(S)}$ and $SA^{(A)}$, respectively. The difference between group A and the other three groups is statistically significant (t-test $p < 0.05$). This manifests that A workers are struggling with constructing justifications when they start the task, although they reuse content from the topic in Step 1 (see Figure 6a). After two judgments, the workers in abandoned HITs (group A and $SA^{(A)}$) constantly spend less time in providing justifications than those in submitted HITs (group S and $SA^{(S)}$), and the difference is statistically significant (t-test $p < 0.05$) from Step 5 onward. This reveals that the workers in abandoned HITs are losing engagement in the task as they progress. Compared to workers in $SA^{(S)}$, the workers in $SA^{(A)}$ write justifications in a significantly quicker manner (t-test $p < 0.05$) at all steps. Considering the overlap of the two population groups as a whole, workers showed less engagement in the topics that they are not good

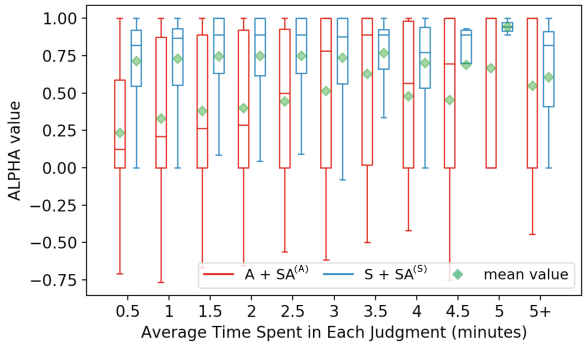


Fig. 10: Judgment quality as compared to the time spent on each judgment.

at judging documents for. Next, we look more in depth at how judging time impacts judging quality.

4.3.6 Time Impact on Quality

Figure 10 shows that judgment quality is influenced by the time spent on each document to some extent for the workers in submitted HITs but significantly more for those in abandoned HITs. For the abandoning groups, the average quality (measured by α agreement with expert judgments) improves from about 0.2 to more than 0.6 with more time spent on documents. By comparison, with an average judging time of less than 3.5 minutes the average quality of judgments by workers in the submitting group lie between 0.72 and 0.77. The quality decreases, however, when more than 3.5 minutes are spent judging documents for both groups. This is in line with previous research that shows how long judging time may result in lower quality judgments [47].

Table 6 shows how quality scores vary with the average judging time for the two groups. For those who spent less than half a minute on a document, only 4.56% of the workers in submitting groups and 30.83% of those in abandoning groups provided low quality ($\alpha \leq 0.66$) responses. This shows that despite being quicker, workers who submit tasks manage to produce higher quality judgments when compared to those who abandon tasks.

High quality ($\alpha > 0.66$) contributors with an average judging time between 0.5 and 1.5 minutes account for 38.82% of the population of submitting group and 13.56% of that of abandoning group. In summary, workers in the submitting groups spend on average more time on each document and provide better quality judgments as compared to workers in the abandoning groups, which strengthens the conclusion that if workers spend less time and provide low quality judgments, abandonment is more likely to happen also because of the quality checks present in the task.

Overall, nearly two-thirds of the workers in abandoned HITs provide low quality ($\alpha \leq 0.66$) judgments. By comparison, more than 70% of those in submitted HITs provide judgments with high agreement scores ($\alpha > 0.66$).

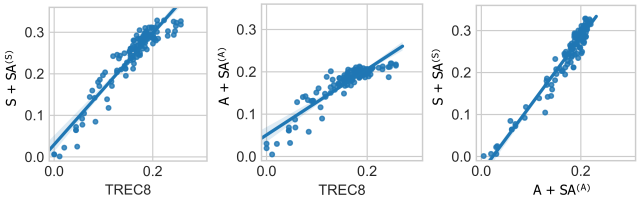
4.4 The Effect of Task Abandonment on Crowdsourced IR Systems Evaluation

Finally, we aim to understand the effect of task abandonment on the crowdsourced evaluation of IR systems. To this end, we used the judgments generated by workers who submitted their HITs to compute a relevance score for each document, and then the NDCG@10 score for each

TABLE 6: Ratio of workers with a given level of quality and average judging time in the submitted HITs (top) and abandoned HITs (bottom).

Time (min)	α interval of submitting groups [-1, 0.66]	(0.66, 0.77]	(0.77, 0.88]	(0.88, 1]	Total
< 0.5	4.56%	1.92%	1.78%	6.76%	15.02%
< 1.5	15.07%	5.71%	6.29%	26.82%	53.89%
< 2.5	5.81%	2.75%	2.21%	11.7 %	22.47%
< 3.5	2.12%	0.81%	0.86%	4.13%	7.92%
< 4.5	0.23%	0.07%	0.07%	0.27%	0.64%
≥ 4.5	0.01%	0 %	0.01%	0.04%	0.06%
Total	27.8 %	11.26%	11.22%	49.72%	100 %

Time (min)	α interval of abandoning groups [-1, 0.66]	(0.66, 0.77]	(0.77, 0.88]	(0.88, 1]	Total
< 0.5	30.83%	1.78%	1.38%	6.4%	40.39%
< 1.5	25.28%	1.51%	1.75%	10.3 %	38.84%
< 2.5	6.26%	0.2%	0.34%	3.97%	10.77%
< 3.5	2.26%	0.17%	0.44%	2.69%	5.56%
< 4.5	0.98%	0.03%	0.03%	0.91%	1.95%
≥ 4.5	1.01%	0.1%	0.03%	1.35%	2.49%
Total	66.62%	3.79%	3.97%	25.62%	100 %



(a) (b) (c)

Fig. 11: NDCG@10 values computed with judgments from the different groups.

retrieval system. To this aim, we use the judgments by A and $SA^{(A)}$ workers, and we aggregate such judgments by using majority vote aggregation[12]. In this way we obtain three sets of topic/document judgments (results by S and $SA^{(S)}$ workers, judgments by A and $SA^{(A)}$ workers, and binary editorial judgments by TREC). Figure 11 shows the NDCG@10 scores for the retrieval systems and the induced system ranking generated by the three relevance judgment sets. We can see that: i) judgments provided by S and $SA^{(S)}$ workers generate an IR system ranking more similar to that obtained via editorial judgments than A and $SA^{(A)}$ worker judgments (Kendall τ of 0.75 vs 0.68 as shown in Figure 11a and 11b)[13], especially on the most effective systems; ii) the IR system rankings produced by judgments in submitted and abandoned HITs are similar ($\tau = 0.73$), but they tend to disagree on top and mid ranked systems (Figure 11c).

5 Study III: The Effect of Reward, Task Length, and Quality Checks

With the aim to inform future crowdsourcing experiment design and identify how we could intervene on people who abandon, we study how individual task properties influence task abandonment. Based on the results from Section 3 and 4, we analyze three factors that bear implications on task abandonment: (i) reward, (ii) task length (i.e., number of documents to be labeled in one HIT), and (iii) the presence of quality checks. Thus, we run a set of controlled experiments where we vary one condition at a time.

12. We break ties, i.e., relevance levels with the same number of selections, at random.

13. We focus on τ because we are interested in the final ranking of IR systems rather than on the exact evaluation measure value.

5.1 Experimental Design

We designed a 4-level scale relevance judgment batch of HITs and deployed it varying one of the independent variable at a time (i.e., reward, task length and quality check). We selected documents from the TREC-8 ad-hoc track [40] so as to have half of them relevant to the given topic, and the other half not relevant according to TREC assessors. To reduce the impact of other factors on the results, we selected documents of approximately the same length from the same TREC topic (i.e., 418) and from the same corpus (i.e., LA-Times). We ran a between-subjects experiment with the following conditions (i.e., a worker could only participate in one of the conditions):

- *Baseline*: The length of the HIT is fixed to 6 documents for which we reward workers $0.30. We do not use any quality check.
- *Reward*: Same as the baseline HIT, but the reward is $0.10.
- *Task Length*: The length of the HIT is 3 documents for which we reward workers $0.15 (i.e., we keep the reward fixed to $0.05 per judgment).
- *Quality Checks*: In addition to the baseline HIT, we include two quality checks; we ask a topic understanding question first and we use two manually-selected *gold* documents, one that is highly relevant (H) and another obviously not relevant (N) for which we require consistent judgments (i.e., the judgment of H should be higher than the judgment of N).

For each condition we published 100 HITs on the F8 platform employing level-2 workers. Workers were allowed to navigate back and forth across documents into the HIT, but needed to express a relevance judgment for each document.

Focusing on the abandoning group, we analyze the effect of these factors on three dependent variables related to abandonment: (a) *number of sessions*[14] that the worker completed (b) *number of steps* logged that show how far the worker went through the task (c) *average time spent per session*.

To study the individual and in-between effects of these factors, we conducted three separate two-way (reward and task length) analysis of covariance[15] (ANCOVA) on the number of sessions, the number of steps and the time per session respectively. To avoid multicollinearity, we set the intercept to zero: a natural choice since zero task length implies null dependent variables by construction. To study the effect of quality checks, we separately conducted three one-way ANOVAs on the same dependent variables. We then applied Bonferroni corrections on the group of tests.

5.2 Results

Firstly, we observe that the abandonment is inversely proportional to both reward (from 47.37% to 51.70% from *Baseline* to *Reward*) and task length (47.37% to 52.15% from *Baseline* to *Task Length*). In the case of the quality checks, when we activated them more people abandoned (from 47.37% to 91.54%)[16].

14. The number of times a worker started the HIT again (e. g. refresh).

15. Since reward and task length are interval variables.

16. We allow workers to attempt the same HIT up to three times if they fail in the quality checks. Therefore, abandonment that happens after three failures of quality checks should not be considered as intentional abandonment as this is rather representing the requester rejecting low quality work.

TABLE 7: Two-way ANCOVA with reward and task length factors, and one-way ANOVA with quality control factor.

	F	Adj.p-value	ω^2
Two-way ANCOVA			
Number of Sessions			
Reward	76.07	$p < .001$	0.11
Task Length	113.01	$p < .001$	0.18
Reward:Task Length	43.35	$p < .001$	0.07
Number of Steps			
Reward	48.05	$p < .001$	0.10
Task Length	22.96	$p < .001$	0.05
Reward:Task Length	4.04	$p = .27$	0.01
AVG Time per Session			
Reward	0.08	$p = 1$	-0.01
Task Length	1.38	$p = 1$	0.01
Reward:Task Length	1.01	$p = 1$	0.01
One-way ANOVA			
Number of Sessions			
Quality Control	0.76	$p = 1$	-0.01
Number of Steps			
Quality Control	47.31	$p < .001$	0.09
AVG Time per Sessions			
Quality Control	65.47	$p < .001$	0.12

The effect of reward and task length is statistically significant ($p < 0.05$, $\alpha = 0.008\overline{3}$ after Bonferroni correction) with medium-large effect size ($\omega^2 > 0.05$), on the number of sessions and the number of steps (also jointly for the number of steps). The effect of quality checks is statistically significant with large effect size ($\omega^2 > 0.06$), on the number of steps and on the average time spent per session.

6 Discussion and Conclusions

In this paper we have investigated the understudied phenomenon of task abandonment in crowdsourcing, i.e., crowd workers who start a HIT but do not complete it, thereby failing to submit their responses. Their responses are therefore not captured by the platform or the requesters, and as a result workers do not receive any monetary compensation. We have conducted three distinct studies by means of: i) Crowd worker surveys to understand workers' perception of abandonment; ii) A large-scale crowdsourced relevance judgment experiment to understand the different dimensions of abandonment; and iii) Controlled experiments on the factors influencing abandonment.

Our main findings show that: i) Workers tend to abandon tasks early if the reward is not considered worth the required effort; ii) Overall, task abandonment is a widespread phenomenon but most of it occurs early in the task; iii) The quality of relevance judgments provided by workers who abandon is worse than that by workers who complete the task[17]; iv) Workers in abandoned tasks show decreasing engagement as they progress in the tasks, with respect to the quality of their judgments and the time they spend on writing justifications; v) Workers who abandon also provide faster judgments as compared to those who complete. However, we have also observed fast and high quality judgments by workers who complete; vi) The IR evaluation results generated with judgments by workers who complete is more similar to that obtained with expert judgments as compared

17. Note that low quality submitted work may not always result in a rejection, as requesters may not be able to check quality without ground truth data.

to judgments by workers who abandon; vii) Quality checks in the HITs have the highest effect on task abandonment.

These results have strong implications on the use of crowdsourcing for IR evaluation. First, quality checks in crowdsourcing have proven to be an essential instrument to implicitly select a sample of the crowd that can provide high quality judgments. On the other hand, this comes with the undesired effect of unrewarded effort by crowd workers who self-select into the abandoning group of workers. In our large-scale relevance assessment experiment, some workers have shown the abilities to rapidly provide high quality judgments. But we did include a time-based check ($\geq$ 20 seconds) as one condition to pass our quality checks, which may block the submissions from those fast and high quality workers. Moreover, as another condition to pass our quality checks, understanding of the initial question can be developed when workers progress in tasks. Our results show that submitting (S and $SA^{(S)}$) workers exhibit a positive learning effect when they judge more documents, and thus the increase of the judgment quality observed over steps implies that the worker has the potential to complete the task with a quality comparable to that of submitting workers. Therefore, these shed some lights on the design of quality check mechanisms, which may replace the classic quality checks presented at the end of the HITs by continuous assessment of the worker quality at each step, for example. The quality control mechanisms we used in our task design, however, have an impact on what we observed and thus our results cannot be generalized to other task types or different experimental setups.

The decision made by workers who are self-selected to start HITs is influenced by many factors, such as task understanding, their self-perceived difficulty in completing the tasks, the evaluation on time/reward trade-off, and so forth. As they progress in the task, their self-estimation of the performance has an impact on whether they continue working on the task or abandon it. This bears the implications on the design of effective workflows in the tasks to encourage workers to provide high quality responses rather than pushing them to lose engagement step by step. In our experiment, workers in abandoned tasks have shown continuous decrease in task engagement. If this phenomenon can be observed at an early stage, for example, intervention action such as increasing the task reward may be introduced to motivate workers to continue doing the task. Therefore, the potential wasted hours caused by task abandonment could be compensated, which has a positive effect on increasing the hourly rate for workers.

In the imminent future, we will extend our controlled experiments (presented in Section 5) by adding more factors to understand the impact of task properties on abandonment. We will also build predictive models for task abandonment, aiming at reducing the dominant abandonment phenomenon we have observed in this paper and its negative effects on crowd work.

Acknowledgements. This work is supported in part by the EU's H2020 research and innovation programme (Grant Agreement No. 732328), the Erasmus+ project DISKOW (Project No. 60171990), and by the ARC Discovery Project (Grant No. DP190102141).

REFERENCES

[1] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of WWW*. ACM, 2014, pp. 155–164.

[2] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *WWW*, 2012, pp. 469–478.

[3] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the right crowd: expert finding in social networks," in *Proceedings of EDBT*. ACM, 2013, pp. 637–648.

[4] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Pick-a-crowd: tell me what you like, and I'll tell you what to do," in *Proceedings of WWW*. ACM, 2013, pp. 367–374.

[5] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information science*, vol. 38, no. 2, pp. 189–200, 2012.

[6] R. L. Gruner and D. Power, "What's in a crowd? exploring crowdsourced versus traditional customer participation in the innovation process," *Journal of Marketing Management*, vol. 33, no. 13-14, pp. 1060–1092, 2017.

[7] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. Jose, and L. Azzopardi, "Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems," *Information retrieval*, vol. 16, no. 2, pp. 267–305, 2013.

[8] G. Demartini, D. E. Difallah, U. Gadiraju, and M. Catasta, "An introduction to hybrid human-machine information systems," *Foundations and Trends® in Web Science*, vol. 7, no. 1, pp. 1–87, 2017.

[9] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *Proceedings of CSCW*. ACM, 2013, pp. 1301–1318.

[10] U. Gadiraju, A. Checco, N. Gupta, and G. Demartini, "Modus operandi of crowd workers: The invisible role of microtask work environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 49, 2017.

[11] S. Dow, A. Kulkarni, B. Bunge, T. Nguyen, S. Klemmer, and B. Hartmann, "Shepherding the crowd: managing and providing feedback to crowd workers," in *CHI EA on Human Factors in Computing Systems*. ACM, 2011, pp. 1669–1674.

[12] G. Kazai, J. Kamps, and N. Milic-Frayling, "The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy," in *Proceedings of CIKM*, 2012, pp. 2583–2586.

[13] H. Li, B. Zhao, and A. Fuxman, "The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing," in *Proceedings of WWW*. ACM, 2014, pp. 165–176.

[14] T. McDonnell, M. Lease, M. Kutlu, and T. Elsayed, "Why is that relevant? collecting annotator rationales for relevance judgments," in *HComp*, 2016.

[15] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2010, pp. 64–67.

[16] C. Eickhoff, "Cognitive biases in crowdsourcing," in *Proceedings of WSDM*. New York, NY, USA: ACM, 2018, pp. 162–170.

[17] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini, "Understanding malicious behavior in crowdsourcing platforms: The case of online surveys," in *Proceedings of CHI*. ACM, 2015, pp. 1631–1640.

[18] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Identifying unreliable and adversarial workers in crowdsourced labeling tasks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3233–3299, 2017.

[19] N. Q. V. Hung, H. H. Viet, N. T. Tam, M. Weidlich, H. Yin, and X. Zhou, "Computing crowd consensus with partial agreement," *TKDE*, vol. 30, no. 1, pp. 1–14, 2017.

[20] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proceedings of SIGIR*. ACM, 2006, pp. 19–26.

[21] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds: identifying relevant websites from user activity," in *Proceedings of WWW*. ACM, 2008, pp. 51–60.

[22] Y. Kim, A. Hassan, R. White, and I. Zitouni, "Modeling dwell time to predict click-level satisfaction," in *Proceedings of WSDM*. ACM, 2014, pp. 193–202.

[23] D. Lagun and M. Lalmas, "Understanding user attention and engagement in online news reading," in *Proceedings of WSDM*. ACM, 2016, pp. 113–122.

[24] Y. Chen, Y. Liu, M. Zhang, and S. Ma, "User satisfaction prediction with mouse movement information in heterogeneous search environment," *TKDE*, vol. 29, no. 11, pp. 2470–2483, 2017.

[25] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: Using implicit behavioral measures to predict task performance," in *Proceedings of UIST*. ACM, 2011, pp. 13–22.

[26] G. Kazai and I. Zitouni, "Quality management in crowdsourcing using gold judges behavior," in *WSDM*, 2016, pp. 267–276.

[27] T. Goyal, T. McDonnell, M. Kutlu, T. Elsayed, and M. Lease, "Your behavior signals your reliability: Modeling crowd behavioral traces to ensure quality relevance annotations," in *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[28] A. Diriye, R. White, G. Buscher, and S. Dumais, "Leaving so soon?: understanding and predicting web search abandonment rationales," in *Proceedings of CIKM*. ACM, 2012, pp. 1025–1034.

[29] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz, "Direct answers for search queries in the long tail," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2012, pp. 237–246.

[30] A. Kobren, C. H. Tan, P. Ipeirotis, and E. Gabrilovich, "Getting more for less: Optimized crowdsourcing with dynamic tasks and goals," in *Proceedings of WWW*, 2015, pp. 592–602.

[31] A. Hassan, X. Shi, N. Craswell, and B. Ramsey, "Beyond clicks: query reformulation as a predictor of search satisfaction," in *Proceedings of CIKM*. ACM, 2013, pp. 2019–2028.

[32] R. Mehrotra, A. Awadallah, M. Shokouhi, E. Yilmaz, I. Zitouni, A. El Kholy, and M. Khabsa, "Deep sequential models for task satisfaction prediction," in *Proceedings of CIKM*. ACM, 2017, pp. 737–746.

[33] U. Gadiraju, R. Kawase, and S. Dietze, "A taxonomy of microtasks on the web," in *Proceedings of HT*. ACM, 2014, pp. 218–223.

[34] A. L. Strauss, *Qualitative analysis for social scientists*. Cambridge University Press, 1987.

[35] B. L. Berg, H. Lune, and H. Lune, *Qualitative research methods for the social sciences*. Pearson Boston, MA, 2004, vol. 5.

[36] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of CHI*. ACM, 2018.

[37] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2010, pp. 2863–2872.

[38] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin, "On crowdsourcing relevance magnitudes for information retrieval evaluation," *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 19:1–19:32, Jan. 2017.

[39] K. Roitero, E. Maddalena, G. Demartini, and S. Mizzaro, "On fine-grained relevance scales," in *SIGIR*, 2018, pp. 675–684.

[40] E. Voorhees and D. Harman, "Overview of The Eighth Text REtrieval Conference (TREC 8), 1–24," *NIST Special Publication*, 1999.

[41] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?" in *Proceedings of SIGIR*, 2002, pp. 324–330.

[42] L. Han, K. Roitero, U. Gadiraju, C. Sarasua, A. Checco, E. Maddalena, and G. Demartini, "All those wasted hours: On task abandonment in crowdsourcing," in *Proceedings of WSDM*. ACM, 2019, pp. 321–329.

[43] K. Kirppendorff, "Content analysis: An introduction to its methodology," *Beverley Hills: Sage*, 1989.

[44] J. W. Ratcliff and D. E. Metzener, "Pattern-matching-the gestalt approach," *Dr Dobbs Journal*, vol. 13, no. 7, p. 46, 1988.

[45] E. Maddalena, K. Roitero, G. Demartini, and S. Mizzaro, "Considering assessor agreement in ir evaluation," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 2017, pp. 75–82.

[46] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[47] E. Maddalena, M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini, "Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge," in *HComp*, 2016.

Lei Han is a PhD Student at the University of Queensland (UQ), Brisbane, Australia. His main research interests include Human Computation, Crowdsourcing, Data mining and analysis, and Information Retrieval. Currently, he is working on understanding the behaviors exhibited by crowd workers, and their impact on the quality of crowdsourced outcomes. Before joining UQ, he had been working in the software sector for eight years.

Kevin Roitero is a PhD Student at the University of Udine, North-East of Italy. His research interests include Information Retrieval Evaluation, Crowdsourcing, Data mining and analysis, Machine Learning, and Statistical Modelling. He visited and collaborated with multiple universities, where he worked on the development of crowdsourcing tasks with the aim of understanding and predicting user features such as user engagement, user agreement, and bias.

Dr. Ujwal Gadiraju is a Postdoctoral Researcher at the L3S Research Center, Leibniz University of Hannover, Germany. His main research interests include Human Computation, Crowdsourcing, Social Computing, and Information Retrieval. He has published over 60 peer-reviewed scientific articles, including papers at top-tier conferences and high-impact journals. Among other honors, Ujwal received the Douglas Engelbart Best Paper Award at the ACM Conference on Hypertext and Social Media (HT) in 2017.

Cristina Sarasua is a research scientist and lecturer in Social Computing at the University of Zurich (Switzerland). Her current research focuses on the intersection of human computation and crowdsourcing methods and various Web data-related fields including Semantic Web data integration, knowledge engineering and data mining.

Dr. Alessandro Checco is a lecturer in Business Analytics at the Information School, University of Sheffield. His main research interests are Crowdsourcing, Human Computation, Distributed Private Recommender Systems, Information Retrieval, Data Privacy, and Algorithmic Bias.

Dr. Eddy Maddalena is a Research fellow at the University of Southampton, in the United Kingdom. Eddy's research contributions range from Information Retrieval to Human Computation and Crowdsourcing.

Dr. Gianluca Demartini is a Senior Lecturer in Data Science at the University of Queensland, Australia. His research interests include Information Retrieval, Semantic Web, and Human Computation. He received Best Paper Awards at the AAAI Conference on Human Computation and Crowdsourcing (HCOMP) in 2018 and at the European Conference on Information Retrieval (ECIR) in 2016. He has published more than 100 peer-reviewed scientific articles.