All Those Wasted Hours: On Task Abandonment in Crowdsourcing

Lei Han[+], Kevin Roitero[*], Ujwal Gadiraju[◇], Cristina Sarasua[‡], Alessandro Checco[▽], Eddy Maddalena[△], and Gianluca Demartini[+]

[+]University of Queensland, Australia. [*]University of Udine, Italy. [◇]L3S Research Center, Germany. [‡]University of Zurich, Switzerland. [▽]University of Sheffield, UK. [△]University of Southampton, UK.

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 the first investigation into the phenomenon of task *abandonment*, the act of workers previewing or beginning a task and deciding not to complete it. We follow a threefold methodology which includes 1) investigating the prevalence and causes of task abandonment by means of a survey over different crowdsourcing platforms, 2) data-driven analyses 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 IR systems).

ACM Reference Format:
Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291035

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 [7, 41] to profiling crowd workers in order to assign them tasks they can perform well on [5, 9]. 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. This may have an impact on the quality of the data collected by means of crowdsourcing. 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, we present the first work that 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). With respect to workers, in our large-scale experiment, we observed 1K distinct workers who completed HITs, and 4K distinct workers who started but then abandoned. The total effort invested by abandoning workers accounts for 616 hours of work which are equivalent to about 3.5 months FTE.
- 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.

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291035

[1]AMT: http://www.mturk.com, F8: http://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.
- 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 IR systems.

2 RELATED WORK

Crowdsourcing has recently become a Web-based model that leverages distributed human intelligence to solve highly complex data problems [6, 13]. 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 [8, 18, 43]. One of the main challenges in applying crowdsourcing to data problems is the *quality* of crowdsourced data [14, 27]. 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. [23] found that the profile of the workers can significantly affect the accuracy of the tasks. Li et al. [30] proposed a crowd targeting framework to improve accuracy at the same or even lower budgetary cost. McDonnell et al. [34] 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. [21] provided a solution to distinguish true errors from individual's systematic biases and Eickhoff [12] looked at the effect of cognitive biases in the crowd on IR evaluation. Detecting malicious workers was also discussed in [16, 22]. 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 [1, 4, 25], 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 [29].

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 [38] where authors use behavioral traces to predict the quality of crowd worker answers in a supervised manner. More recently, in [24] authors 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 [17] authors 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 [10], 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 [3], 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 [28]. 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., [20, 25, 35]. 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 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 [15] 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

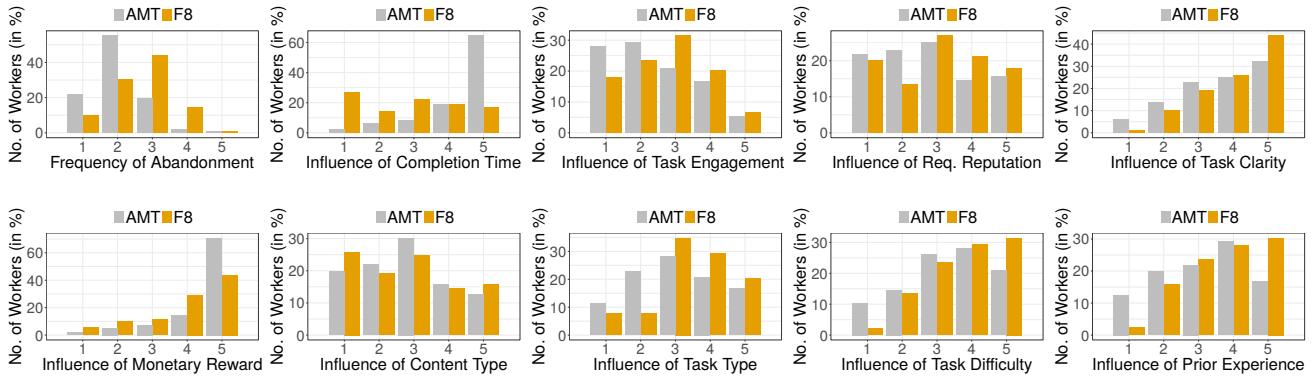


Figure 1: (*Top-left*) Frequency of task abandonment as perceived by workers on Amazon MTurk (AMT) and FigureEight (F8), and (*remaining sub-figures*) influence of various factors that affect task abandonment on the two platforms.

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

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.

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 [2, 40]. 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 [19, 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 being 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 novel 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 two main research questions here.
RQ1: How well do workers who abandon HITs (group A) perform when compared to those who complete the HITs (group S)?
RQ2: How much work do workers in group A complete before abandoning the HITs?

4.1 Crowdsourcing Data Collection

4.1.1 Task Design. We ran a large relevance assessment experiment following the design used by [32] and [36]. The HITs are presented to workers with a topic and eight documents taken from the TREC-8 ad hoc collection [42]. 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 [34]. 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 6 of the 8 documents, and (iii) two of the eight documents were gold standard editorial judgments by [39] 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 that workers judged these documents consistently ($H > N$). These three checks are 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 over 18 topics and 7'067 HITs. These judgments have been completed by 1'154 unique workers, since we allow them to participate in multiple topics (but only one HIT per topic). At the same time, we observed 4'102 unique workers who abandoned HITs during the experiment. Overall, 11'563 HITs have been abandoned and 7'067 HITs have been completed.

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

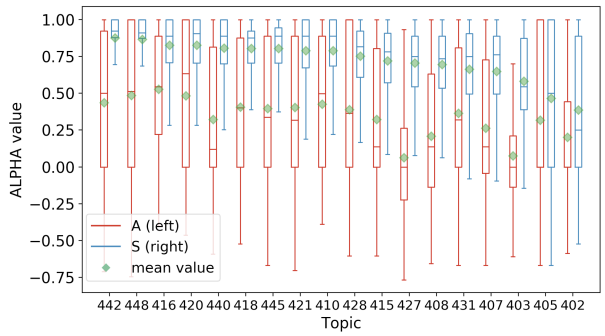


Figure 2: Judgment quality over topics comparing *S* and *A* workers. Topics are sorted by decreasing mean value for *S*.

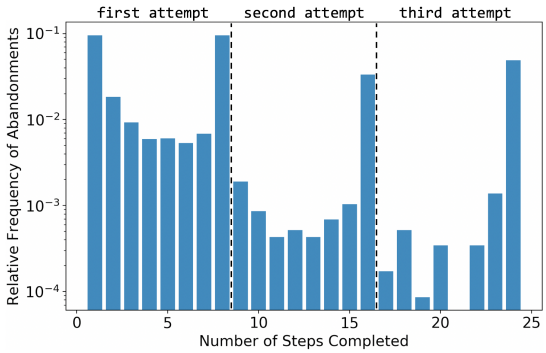


Figure 3: Relative frequency of abandonment (log scale) over the number of completed judgments.

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[2].

4.2 Methodology

Using the task design and logging infrastructure described above, we collected action logs and relevance judgments from two populations of crowd workers: those who *submitted* their judgments by completing a HIT (group *S*) and those who started the HIT but then *abandoned* it before completion (group *A*). We examined our dataset from three 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.

To measure the quality of judgments provided by workers, we compare them to ground truth editorial assessments on a 4-level scale by Sormunen [39]. Thus, we compare crowd worker judgments from both *S* and *A*, with judgments from experts, by means of agreement measures. To measure agreement between crowd workers and experts we use Krippendorff's Alpha coefficient [26] 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 *A* workers may have provided fewer labels, we only measure agreement over the subset of eight documents in each HIT for which both workers and experts judgments are available. For each HIT we compute the quality of the judgments contributed by *S* and *A* workers. We then average agreement scores across HITs for the same topic.

4.3 Results

4.3.1 Quality. The average α agreement with experts for the *S* group is 0.74, while it is 0.33 for the *A* group. Figure 2 shows the differences of α values between the *S* and *A* groups over topics, where topics are sorted in descending order of average α value for workers in the *S* group. It is evident that the average judgment quality for the *A* group is lower than the quality for the *S* group across all different topics. The highest average α value across topics for the *A* group is 0.53. Using the Wilcoxon signed-rank test to compare the quality of the *A* and *S* groups 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 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.

Table 3 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 [33]).

Figure 3 presents the distribution of abandonment for the 3'860 workers who performed at least one relevance judgment, showing the ratio of whole *A* population who abandoned after a given step. We can see that the abandonment happening after Step 1 and 8 is the largest. These two steps represent workers abandoning after

[2]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.

Table 3: 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%)

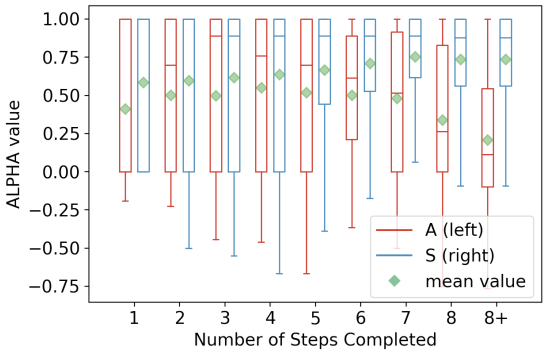


Figure 4: Judgment quality over steps for *S* and *A* workers.

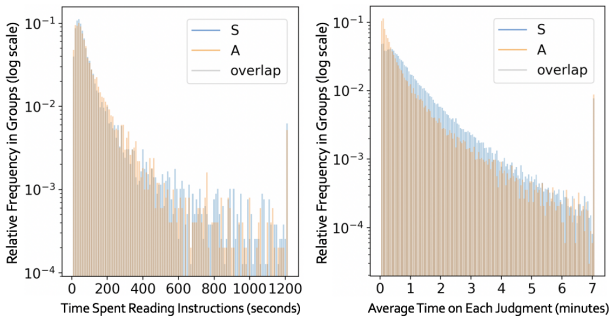


Figure 5: Time spent reading instructions (left) and judging each document (right) for *S* and *A* workers.

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 (i.e., at the end of the third and second full attempt respectively). This shows the presence of two important points of abandonment, i.e., after the *first* or *last* question in the HIT.

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 1'049 HITs (9.1% of *A* HITs) workers performed the same judgments again (Step 8+). Among *S* workers, in 1'366 HITs (19.3% of *S* 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 the quality of the judgments provided up to a given step in the HIT comparing *S* and *A* workers. For a given step, we compare the quality of judgments provided by *S* workers up to that step to the quality of judgments provided by *A* workers abandoning at that step.

Figure 4 shows the judgment quality over steps for the *S* and *A* groups. We can see that workers who submitted constantly provide higher quality labels than workers who abandoned. 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 abandon, the average quality increases from Step 1 to 4 and then drops up to Step 8 showing a decrease in engagement as they progress throughout the HIT. The quality of judgments by *A* workers differs from that by *S* workers significantly (Wilcoxon signed-rank test $p < 0.05$) at each step except for Step 2 and 4. Workers who started multiple times in the *A* group provide the lowest average quality judgments.

4.3.4 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. Figure 5 shows the distribution of *A* and *S* workers 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. The number of workers who spent more than 1'200 seconds (or 20 minutes) reading instructions is less than 1% in both the *A* and *S* groups, 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 the two groups approach the task in a

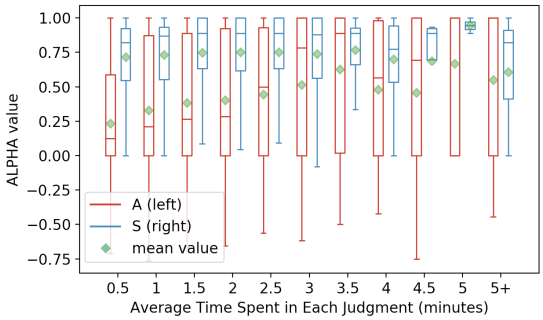


Figure 6: Judgment quality as compared to the time spent on each judgment.

similar way. On the contrary, the two groups show differences in the average time spent judging each document. 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 *A* group (40.39%) as compared to that from the *S* group (15.02%). The proportion of workers whose judging time is from 0.5 to 3.5 minutes in the *A* group is lower (55.17%) than that of the *S* group (84.28%). This observation reveals that although workers in the two groups have similar instruction reading patterns, the time devoted to judging documents is different. Next, we look more in depth at how judging time impacts judging quality.

4.3.5 Time Impact on Quality. Figure 6 shows that judgment quality is influenced by the time spent on each document to some extent for the *S* group but significantly more for the *A* group. For the *A* group, 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 *S* 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 [31].

Table 4 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 *S* workers and 30.83% of *A* workers 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.

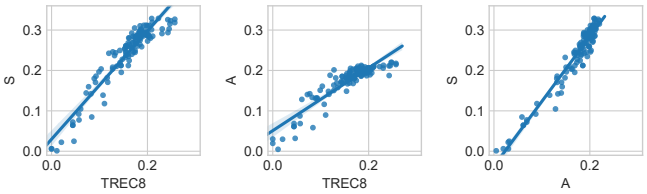
High quality ($\alpha > 0.66$) contributors with an average judging time between 0.5 and 1.5 minutes account for 38.82% of the *S* population and 13.56% of the *A* group. In summary, workers in the *S* group spend on average more time on each document and provide better quality judgments as compared to workers in the *A* group, 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 abandoning *A* workers provide low quality ($\alpha \leq 0.66$) judgments. By comparison, more than 70% of submitting *S* workers provide judgments with high agreement scores ($\alpha > 0.66$).

Table 4: Ratio of workers with a given level of quality and average judging time in the *S* (top) and *A* (bottom) groups.

	α interval of *S* group				
Time (min)	[-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 %

	α interval of *A* group				
Time (min)	[-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)**

Figure 7: NDCG@10 values computed with judgments from the different groups.

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 an information retrieval (IR) system. To this end, we used the judgments generated by workers who submitted their HITs (i.e., the standard crowdsourced IR evaluation approach) and the judgments contributed by *A* workers before they abandoned the task to generate two different relevance assessments obtained by majority vote aggregation[3]. In this way we obtain three sets of topic/document judgments (*S*, *A*, and binary editorial judgments by TREC). Figure 7 shows the three different pairwise comparisons between IR system rankings generated by the three relevance judgment sets. We can see that: i) judgments provided by *S* workers generate an IR system ranking more similar to that obtained via editorial judgments than *A* worker judgments (Kendall τ of 0.75 vs 0.68 as shown in Figure 7a and 7b)[4], especially on the most effective systems; ii) the IR system rankings produced with *S* and *A* judgments are similar ($\tau = 0.73$), but they tend to disagree on top and mid ranked systems (Figure 7c).

[3]We break ties, i.e., relevance levels with the same number of selections, at random.
[4]We focus on τ because we are interested in the final ranking of IR systems rather than on the exact evaluation measure value.

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 in the group of 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.

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 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*: Identical to 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 were required 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 the abandonment behavior: (a) *number of sessions*[5] 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[6] (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]The number of times a worker started the HIT again (e. g. refresh) in the browser.
[6]Since reward and task length are interval variables.

Table 5: 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

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%).

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 on the number of sessions 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[7]; iv) Workers who abandon also provide faster judgments as compared to those who complete. However,

[7]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.

we have also observed fast and high quality judgments by workers who complete; v) The IR evaluation results generated with judgments by workers who complete is more similar to that obtained with expert judgments as compared to judgments by workers who abandon; vi) 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.

We have also observed that behavioral logging might be used as an instrument for requesters to collect data 'for free' without rewarding workers and pushing them to abandon tasks. As our findings suggest, however, such an approach would result in low-quality data and thus cannot be used against workers and to unbalance the crowdsourcing ecosystem. Our future work will focus on better understanding the causes of abandonment to design prediction models for task abandonment. This will aim 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), and the Erasmus+ project DISKOW (Project No. 60171990).

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR*. ACM, 19–26.

[2] Bruce Lawrence Berg, Howard Lune, and Howard Lune. 2004. *Qualitative research methods for the social sciences*. Vol. 5. Pearson Boston, MA.

[3] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 237–246.

[4] Mikhail Bilenko and Ryen W White. 2008. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 51–60.

[5] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. 2013. Choosing the right crowd: expert finding in social networks. In *Proceedings of EDBT*. ACM, 637–648.

[6] Daren C Brabham. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* 14, 1 (2008), 75–90.

[7] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of WWW*. ACM, 469–478.

[8] Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, Michele Catasta, et al. 2017. An introduction to hybrid human-machine information systems. *Foundations and Trends® in Web Science* 7, 1 (2017), 1–87.

[9] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 367–374.

[10] Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. 2012. Leaving so soon?: understanding and predicting web search abandonment rationales. In *Proceedings of CIKM*. ACM, 1025–1034.

[11] Steven Dow, Anand Kulkarni, Brie Bunge, Truc Nguyen, Scott Klemmer, and Björn Hartmann. 2011. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI EA on Human Factors in Computing Systems*. ACM, 1669–1674.

[12] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654

[13] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science* 38, 2 (2012), 189–200.

[14] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 49.

[15] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of HT*. ACM, 218–223.

[16] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of CHI*. ACM, 1631–1640.

[17] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[18] Richard L Gruner and Damien Power. 2017. What's in a crowd? Exploring crowdsourced versus traditional customer participation in the innovation process. *Journal of Marketing Management* 33, 13-14 (2017), 1060–1092.

[19] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of CHI*. ACM.

[20] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of CIKM*. ACM, 2019–2028.

[21] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 64–67.

[22] S Jagabathula, L Subramanian, and A Venkataraman. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *The Journal of Machine Learning Research* 18, 1 (2017), 3233–3299.

[23] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of CIKM*. ACM, 2583–2586.

[24] Gabriella Kazai and Imed Zitouni. 2016. Quality Management in Crowdsourcing Using Gold Judges Behavior. In *Proceedings of WSDM*. ACM, 267–276.

[25] Y Kim, A Hassan, RW White, and I Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of WSDM*. ACM, 193–202.

[26] K Kirppendorff. 1989. Content analysis: An introduction to its methodology. *Beverley Hills: Sage* (1989).

[27] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on CSCW*. ACM, 1301–1318.

[28] Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th WWW confererence*. 592–602.

[29] Dmitry Lagun and Mounia Lalmas. 2016. Understanding user attention and engagement in online news reading. In *Proceedings of WSDM*. ACM, 113–122.

[30] Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*. ACM, 165–176.

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

[32] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Syst.* 35, 3, Article 19 (Jan. 2017), 32 pages.

[33] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering assessor agreement in ir evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 75–82.

[34] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[35] R Mehrotra, AH Awadallah, M Shokouhi, E Yilmaz, I Zitouni, A El Kholy, and M Khabsa. 2017. Deep Sequential Models for Task Satisfaction Prediction. In *Proceedings of CIKM*. ACM, 737–746.

[36] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *SIGIR*. 675–684.

[37] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[38] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of UIST*. ACM, 13–22. https://doi.org/10.1145/2047196.2047199

[39] Eero Sormunen. 2002. Liberal relevance criteria of TREC-: Counting on negligible documents?. In *Proceedings of SIGIR*. ACM, 324–330.

[40] Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge University Press.

[41] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of WWW*. ACM, 155–164.

[42] EM Voorhees and DK Harman. 1999. Overview of The Eighth Text REtrieval Conference (TREC 8), 1–24. *NIST Special Publication* (1999).

[43] G Zuccon, T Leelanupab, S Whiting, E Yilmaz, JM Jose, and L Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16, 2 (2013), 267–305.