Information Retrieval Journal manuscript No.
(will be inserted by the editor)

An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments

Gabriella Kazai · **Jaap Kamps** ·
Natasa Milic-Frayling

Received: date / Accepted: date

Abstract Crowdsourcing relevance judgments for the evaluation of search engines is used increasingly to overcome the issue of scalability that hinders traditional approaches relying on a fixed group of trusted expert judges. However, the benefits of crowdsourcing come with risks due to the engagement of a self-forming group of individuals–the crowd, motivated by different incentives, who complete the tasks with varying levels of attention and success. This increases the need for a careful design of crowdsourcing tasks that attracts the right crowd for the given task and promotes quality work. In this paper, we describe a series of experiments using Amazon's Mechanical Turk, conducted to explore the 'human' characteristics of the crowds involved in a relevance assessment task. In the experiments, we vary the level of pay offered, the effort required to complete a task and the qualifications required of the workers. We observe the effects of these variables on the quality of the resulting relevance labels, measured based on agreement with a gold set, and correlate them with self-reported measures of various human factors. We elicit information from the workers about their motivations, interest and familiarity with the topic, perceived task difficulty, and satisfaction with the offered pay. We investigate how these factors combine with aspects of the task design and how they affect the accuracy of the resulting relevance labels. Based on the analysis of 960 HITs and 2,880 HIT assignments resulting in 19,200 relevance labels, we arrive at insights into the complex interaction of the observed factors and provide practical guidelines to crowdsourcing practitioners. In addition, we highlight challenges in the data analysis that stem from the peculiarity of the crowdsourcing environment where the sample of individuals engaged in specific work conditions are inherently influenced by the conditions themselves.

Keywords Crowdsourcing · Relevance judgments · Study of human factors

Gabriella Kazai and Natasa Milic-Frayling
Microsoft Research Cambridge
E-mail: {v-gabkaz,natasamf}@microsoft.com

Jaap Kamps
University of Amsterdam
E-mail: kamps@uva.nl

1 Introduction

Crowdsourcing – as defined by Jeff Howe – is the act of outsourcing work to a large group of unknown people – a *crowd* [22], usually, in order to solve a problem or complete a task, and often in exchange for micro-payments, social recognition, or entertainment. With the appearance of crowdsourcing platforms, such as CrowdFlower[1] or Amazon's Mechanical Turk (AMT)[2], crowdsourcing has become broadly accessible. Employers can easily post specific jobs in the form of Human Intelligence Tasks (HITs) and engage a large population of workers within a short period of time at a relatively low cost.

Crowdsourcing has proven particularly useful for achieving scale and efficiency in tasks where human input is necessary, such as acquiring relevance labels in information retrieval (IR) to enable the comparative evaluation of search engines [4,3,19,2,33,40]. Indeed, crowdsourcing offers an attractive alternative to traditional methods of gathering relevance labels that rely on the efforts of a fixed group of experts. With the growing size of contemporary data collections and the continuously evolving information needs of large user populations, the required effort and cost of traditional relevance assessments are increasingly prohibitive.

However, while increasingly popular, the use of crowdsourcing has been criticized for its mixed quality output. Marsden, for example, argues that 90% of crowdsourcing contributions are suboptimal [41]. On the other hand, in the context of relevance assessments, several studies concluded that crowdsourcing can lead to reliable relevance labels [3,19], provided that appropriate quality assurance steps are taken [14,36,43], especially amidst growing concerns of cheating or random behavior among members of the crowd [58,38].

Clearly, engaging a workforce through crowdsourcing brings its own challenges. While in traditional employment arrangements the relationship between employers and workers is established through standardized and legally controlled procedures, in online crowdsourcing environments this is achieved through a computer mediated interaction – largely through the HITs themselves. Given that the primary interaction between the employer and the workers is through the HIT interface, it is expected that the design of the HITs will affect the results of the crowdsourced work. For example, the quality of relevance labels may be correlated with the amount of content in a HIT that needs to be reviewed and labeled, the quality assurance measures taken, such as test and trap questions, and the level of pay offered. Furthermore, aspects of the HIT design may be correlated with characteristics of the workers who choose to engage in the task. The objective of our research is to contribute to the understanding of the factors that influence the quality of crowdsourced relevance labels, both in terms of the resulting label accuracy and the workers' experience with the task. In particular, we aim to study the characteristics of the crowd that is attracted to a given HIT under a specific set of work conditions.

Our research expands on earlier studies of crowdsourcing [14,36,2,33,29,31] with deeper insights into the crowd workers' self-reported experience with the task. We achieve this by devising HIT designs that vary specific aspects of work conditions, such as the level of pay offered or the amount of effort required of a worker to complete a HIT, and also including a series of questions about the workers themselves. The workers fill in the questionnaires as they complete each HIT, indicating their motivation for participating in crowdsourcing, their interest in the task, their familiarity with the topic, their perception of the task difficulty, and their opinion about the pay for the required effort. The collected data enables us to study relationships between the information disclosed by the individual workers, the various as-

[1] http://crowdflower.com/

[2] https://www.mturk.com/

pects of the task design, and the overall crowd behavior under specific task conditions. Such insights are essential for making informed decisions about the various trade-offs in the design of HITs for relevance assessments. For example, employers need to weigh the cost and benefit of quality assurance, e.g., collecting multiple labels or using test questions to detect spammers, vs. increased costs and effort required for task completion. Moreover, it may be equally important to understand the workers' motivations and experience with the task in order to increase the quality of the crowdsourced data. Therefore, through our investigations, we aim to answer the following research questions:

R1 How do different conditions of pay, required effort, and selection of workers based on proven reliability affect the quality of the crowdsourced relevance labels?

R2 How do various human factors, such as motivation, expertise, level of interest, perceived task difficulty and satisfaction with the offered pay, that characterize a group of workers under a specific task condition relate to the resulting label quality?

To answer these questions, we run a series of crowdsourcing tasks on AMT using the data and the format of the relevance assessment task from the INEX 2008 Book Search track. Our analysis of the collected data illustrates the issues involved in processing data collected from crowdsourcing where sampling of the participants and the task conditions are tightly coupled. The results of our investigations provide insights into the human factors behind crowdsourcing and lead to both research questions to be further explored and design recommendations for the effective crowdsourcing of relevance labels.

In the rest of the paper, we present our research results as follows. In Section 2, we provide an overview of related work and, in Section 3, we detail our experimental method. In Section 4, we discuss the collected data. In Section 5, we present a detailed analysis of the effects that different task conditions have on the accuracy of the relevance labels. In Section 6, we analyze the workers' self-reported information and, in Section 7, we correlate them with the task conditions. In Section 8, we summarize our findings and discuss design recommendations for practitioners in crowdsourcing.

2 Related Work

In this section we first briefly review relevant notions of IR evaluation then introduce the crowdsourcing terminology that we adopt throughout the paper. Finally we detail related crowdsourcing studies.

2.1 System Evaluation and Relevance Assessments in IR

A standard practice of evaluating IR effectiveness is based on the Cranfield paradigm [11], broadly adopted by TREC[3] [55] and similar IR evaluation forums. It involves purposely created test collections that comprise document corpora and search topics with corresponding relevance judgments. Traditionally, test collections are built under controlled conditions: topics are created by professional searchers and the documents retrieved by various IR systems are pooled to be judged by trusted assessors. For example, TREC employs former intelligence analysts to specify search topics and assess document relevance [55]. The compiled set of documents, topics and relevance labels are then used to compute performance metrics across IR systems, e.g., precision and recall.

[3] http://trec.nist.gov/

It is well-known that relevance judgments are subjective to some extent and may vary across assessors. For instance, the reported pairwise agreement on relevant and non-relevant documents between two TREC assessors is 70-80% on average, but varies greatly per topic [55]. Several measures can be taken in order to mitigate the inherent variability of assessments across judges and topics: obtain multiple labels and calculate majority vote, use a large set of topics, and aim for comparative system rankings under the exact same conditions and on the same test collections. Indeed, studies have shown that comparative ranking of retrieval systems is relatively robust under these variabilities. For example, studies that replicated TREC assessments with non-TREC assessors found considerable disagreements on the relevance labels among assessors but a high level of agreement on the resulting system rankings [12, 54, 8].

An alternative approach to acquiring explicit relevance judgments for search results is to use implicit judgments inferred from user clicks. It has been shown that click based relevance assessments demonstrate a reasonable level of agreement with explicit judgments [48]. However, it is still unclear whether traditional test collections can be fully replaced by implicit judgments due to their incomplete and biased nature [28, 20]. In that view, collecting relevance labels through crowdsourcing is an attractive and feasible alternative that enables to scale up both the number of topics and the number of assessments. A major challenge, however, is the question of how to ensure high quality crowdsourced data. Our research aims to investigate a range of factors that may influence the quality of the relevance labels contributed by crowd workers.

2.2 Crowdsourcing Platforms and Task Design

AMT is a popular crowdsourcing platform and labor marketplace. It allows anyone (with a US address) to create and publish HITs either through the AMT Web based dashboard or programmatically via its API and gain access to hundreds of thousands of workers.

An individual or organization that creates and publishes HITs is known as a *requester*. A person who performs the work is a *worker*. A *HIT* represents a unit of work to be performed by one or more workers. A HIT is a single instance of a HIT template with associated data, e.g., a set of documents to be labeled for relevance. In its simplest form a *HIT template* is an HTML form (with no data attached) that contains instructions of what is expected of workers and HTML controls that allow the workers to interact with the form and complete the given task.

A HIT template and, consequently, each corresponding HIT instance, has an associated pay that is offered to workers and associated cost to the requester, an allotted completion time, and HIT attributes such as title, description, and keywords. Requesters define the number of workers required to complete each HIT. An individual instance of a HIT, assigned to a worker, is referred to as a HIT *assignment*. AMT ensures that different workers complete assignments associated with the same HIT. For example, a HIT designed to collect a single relevance label for a web page may be assigned to 5 different workers, resulting in 5 HIT assignments and, once the work is completed, 5 relevance labels for the same web page. Assignments completed by workers can be approved or rejected by the requester, where rejected work does not incur any cost. Rejected HITs can then be re-published to be completed by other workers. All approved assignments incur a charge of the offered pay, plus a 10% commission by Amazon.

AMT offers two primary methods of quality assurance: 1) requesters can pre-filter workers based on a number of statistics maintained by AMT, and 2) requesters can put workers

through an initial training or qualification phase. The former reflects the key component of AMT's reputation system where workers' reputation is expressed through statistics compiled over their work history. These statistics include the total number of HITs completed since joining AMT and the global approval rate expressed as a ratio of the number of approved HITs and the total number of completed HITs. In this paper we make use of AMT's pre-filtering facility and refer to workers who meet a specified filter criterion as *qualified* workers.

HITs published by requesters appear on AMT's list of available HITs, which is displayed to potential workers. Workers can browse this list or search for HITs using query words that are matched against the HIT titles and keywords. From the experimentation perspective, the sample of workers engaged in a task is not controlled but rather self-selecting or self-forming since it is the workers who decide whether to engage in a HIT or not. That decision is only indirectly influenced through the design of the HIT.

2.3 Crowdsourcing Studies

Recent years have seen a rapid adoption of crowdsourcing methods [47,13] in order to acquire annotations in various contexts, from image labels [44] and search relevance judgments [3,19,33,2,40] to semantic labels in knowledge corpora [53,30]. While crowdsourcing provides a much needed alternative to traditional annotation approaches using skilled editorial staff, obtaining annotations from anonymous online workers brings with it a whole new set of challenges. For example, crowd workers are rarely trained as relevance judges and have diverse backgrounds and motivations to engage in crowdsourcing tasks. Consequently, the resulting labels may be of varied quality. Moreover, recent research has also reported on random behavior by dishonest workers, cheating, and adversarial conduct [58,38,14,29, 33].

Various methods of quality assurance and control have been developed as a result, aiming to enforce correct behavior during task completion or enable the identification and removal of noise or spam in the crowdsourced data [39]. The most established techniques of quality assurance include methods of defensive design, e.g., by building redundancy into the task and obtaining multiple labels from different workers. The noise from multiple assessments is then reduced by applying majority rule or calculating consensus [26,3]. Another frequently adopted technique is the use of "honey pots" or gold standard data sets [53,19]. These can provide a measure of quality based on the agreement between the labels in the gold set and the labels contributed by the crowd. The gold set can also be used to train workers, as illustrated in [38]. The design may also include features that enable the discovery of dishonest workers in the form of qualifying questions [4,5], trap questions [14], or a question sequence and challenge response test [33]. Techniques that attempt to enforce correct behavior include time control mechanisms to focus worker attention [29] and designs that increase the effort of cheating [36]. Once the data has been collected, the requester may apply post-task quality control mechanisms, identifying noise and bias [26], analyzing workers' behavior [58,50] and developing worker models, trust algorithms [57,30] and spam filters [58,56].

Demographic studies of crowd workers on AMT reveal the changing face of this labor market [23,49]. The earlier study of [23] found that workers were mostly female workers from the US, with moderately high income, who solved HITs for fun or for extra income. The more recent studies of [24,49], on the other hand, showed a startlingly different picture, with Indian workers now making up 36% of the worker population, comprising mostly

young males with high education but low regular income. These workers turn to AMT as an opportunity to make a living. This is even more extreme with other crowdsourcing platforms, such as Microworkers,[4] with 63% of workers originating from countries in Asia and only 11% from the US [21]. The increased population of workers may result in a more competitive environment in which workers are incentivized to maximize their own benefits, possibly to the detriment of the work quality, especially when there is little chance of possible reprimand for misconduct [25]. Other contributing factors include different cultural norms [51]. Attempting to categorize workers based on personality traits or behavioral aspects, instead of demographics, works like [34,56,15], showed differences between, e.g., diligent or sloppy and competent or incompetent workers.

Besides the workers' abilities and attitudes, the quality of the resulting work is also affected by the properties of the HITs and their presentation to the workers. Even workers with the best of intentions will produce erroneous data if the design of the task and the user interface (UI) are of poor quality. Examples of bad designs include unclear and ambiguous task instructions, forms that restrict user input, or scales that bias the answers [45]. Despite the awareness of these issues, most of the research on quality assurance in crowdsourcing has focused on methods to discourage unethical or sloppy workers and aid in their detection, thus neglecting aspects of the task design. This is a key issue especially as controlling the behavior of the workers indirectly through the HIT design can be a daunting task for requesters, particularly because little is known about the effect that different designs have on the workers and, ultimately, on the quality of the collected data. For example, questions needing investigation include whether the quality of output increases with pay and whether the workers who are motivated by fun would be put off by higher pay.

Studies that took the first steps towards answering such questions include, e.g., [53,19, 36,2,51,31]. For example, Grady and Lease [19] investigated how varying certain aspects of the HIT design, e.g., title, terminology, and pay, affected annotation accuracy. Shaw et al. [51] studied the effects of different social and financial incentive schemes but found that results were mainly dependent on the task difficulty. In our previous work on crowdsourcing relevance labels we investigated the effects of task design parameters, including the offered pay, effort, and document pooling and ordering on the label accuracy [31,33]. In this paper, we extended this previous work by *combining the investigation of both the task variables and the human factors*, such as workers' motivations for accepting a HIT or their satisfaction with their pay. We performed a large scale and detailed study of the crowdsourced phenomena as part of an annotation task, with specific focus on aspects of human factors that shape the users' experience. We hope that the findings will be of benefit to both current and future practitioners of crowdsourcing. In the following sections we describe in detail our experiment design and the methods we applied to gain insights into the interaction of the observed factors.

3 Experiment Design

As summarized by our research questions (Section 1), our objective is to connect the design aspects of a crowdsourcing based relevance assessment task and the achieved accuracy of relevance labels with characteristics of the workers and their feedback about their experience with the task. By deepening our understanding of the crowd workers, we hope to arrive at

[4] http://microworkers.com/

recommendations to improve the design of HITs and increase the effectiveness of future crowdsourcing engagements.

We adopt an immersive approach and conduct user research as part of an actual crowdsourcing engagement on AMT. The crowd workers are unaware of the research angle of the labeling tasks they are performing.

In the following sections, we describe the relevance assessment task and the data selected for the experiments. We sketch out the experiment grid that reflects the main HIT variables and the human factors that we wish to correlate with the accuracy of the crowdsourced labels.

3.1 Relevance Assessment Task and Experiment Data

For the crowdsourcing task we chose the assignment of relevance labels to pages from digitized books retrieved by the search systems that participated in the INEX 2008 Book Track evaluations [32]. The INEX Book corpus includes 50,239 out-of-copyright books (17 million pages), amounting to a 400GB collection of OCR text. One specific retrieval task investigated by the track is the Focused Book Search (FBS) task (a.k.a. Prove It task), where a search system is expected to point users directly to the relevant pages in the books. For each topic, the search results include a ranked list of book pages that a system considers relevant to the topic.

In order to compare the performance of the participating systems using the Cranfield approach, it is necessary to gather relevance labels for the book pages retrieved by the systems. Since the INEX community relies on the volunteers among its researchers to contribute to the relevance assessments, the scale of this challenge has become prohibitive [35]: "The estimated effort required of an assessor of the INEX 2008 Book Track to judge a single topic was to spend 95 minutes a day for 33.3 days". Through crowdsourcing this effort can be divided among thousands of people and the test collection can be completed in a much shorter time. Moving to a crowdsourcing model, thus, offers clear benefits. However, it raises questions about how to devise HITs that can effectively produce good quality relevance labels by unknown workers of varying abilities and attitudes.

In order to study the effectiveness of a crowdsourcing model to gather relevance labels, we selected 8 out of 70 topics in the INEX 2008 Book Track's test set (ID: 27, 31, 37, 39, 51, 57, 60 and 63). Figure 1 shows topic 57 as an example. We chose topics based on the number of available relevance judgments, aiming at those that had at least 40 relevant and 60 non-relevant judged pages. We chose this semi-equal ratio of relevant to non-relevant samples, following recommendations of the study by Le et al. in [38]. In the context of gathering relevance labels for product search results, they found that workers' responses were influenced by the distribution of correct answers in their training data. So, similarly to how a classifier in machine learning may develop bias towards the training data, they found that ethical workers could become predisposed to selecting labels with the highest priors and miss items that deviate from expectations. Unethical workers, on the other hand, were more likely to optimize their responses to maximize their rate of income. When label categories are highly skewed, the strategy to label all or most items with the most frequent category provided an effective way to cheat. Optimal performance was achieved using uniformly distributed category training data. Since our goal is to study the effects of other task variables, we used the 40-60 ratio to minimize the impact of label distribution in our experiments. The higher percentage of non-relevant pages is a reflection of the skewed distribution of relevance labels in the INEX test set (as is typical in most IR test collections).

```
<inex_topic track=book task=book-retrieval/book-ad-hoc topic_id=57>
 <title> Titanic </title>
 <description> I am interested in real life factual as well as artistic accounts of the
       sinking of the Titanic.
 </description>
 <narrative>
  <task> The story of the Titanic has been made popular with the success of the movie Titanic.
        I would like to find out more about this tragic event and get a better feeling about
        the effect it had on the people of the time.
  </task>
  <infneed> I am interested in historical information about the sinking of the Titanic, both
       witness accounts and historians' take on the events. I am also interested in poems and
       other artistic expressions that relate to this tragedy. I am however not interested
       in the critiques of such arts.
  </infneed>
 </narrative>
</inex_topic>
```

Fig. 1 Topic 57 from the INEX 2008 Book Track test set

For the 8 selected topics, the INEX test set contains 4,490 judged pages in 470 books of which 1,109 pages in 149 books were judged as relevant. We consider these judgments reliable and treat them as our gold standard set for comparison with the relevance labels obtained from the crowd workers. For the crowdsourcing of relevance labels we randomly picked 100 pages per topic, ensuring a 40-60% ratio of relevant and non-relevant pages in the sample. We use the resulting 800 pages in our experiments to investigate the accuracy of crowd workers under different conditions and the human factors that characterize the workers engagement and experience.

3.2 HIT Design

We designed a number of HIT templates to collect relevance labels for the book pages in our data set and, at the same time, capture information about the workers and their experience with the task. The scanned book pages were displayed within the HITs using a web service call directly to the INEX Book Search System, developed at Microsoft Research Cambridge [35]. Each HIT included three major sections, see Figure 2:

- *Instructions* – The top part of the HIT describes the task and the specific search query (topic), and explains what information should be considered relevant to the query. It also includes a request for workers to rate their knowledge on the topic using a four point scale, see Section 3.3.2.
- *Task* – The middle part of the HIT contains the book pages that need to be labeled for relevance. For each page, the user can pick among four options: 'Relevant', 'Not relevant', 'Broken link', or 'Don't know'. They can also use a free-text comment field to provide any additional explanations. In order to detect superficial work and random clicking on answers, we used a challenge-response test or 'captcha', asking workers to enter the last word in the text of the scanned page. This, combined with information from the task logs, such as the time spent on a HIT, can aid the identification of sloppy or dishonest workers and their suboptimal output [53, 36, 58, 56, 34].
- *Questionnaire* – The bottom part of the HIT is a questionnaire aimed at collecting feedback from workers regarding their background and experience with the task, see Section 3.3.2. As common practice in crowdsourcing tasks, a free-text comment box was also provided if workers wanted to give additional information or feedback.

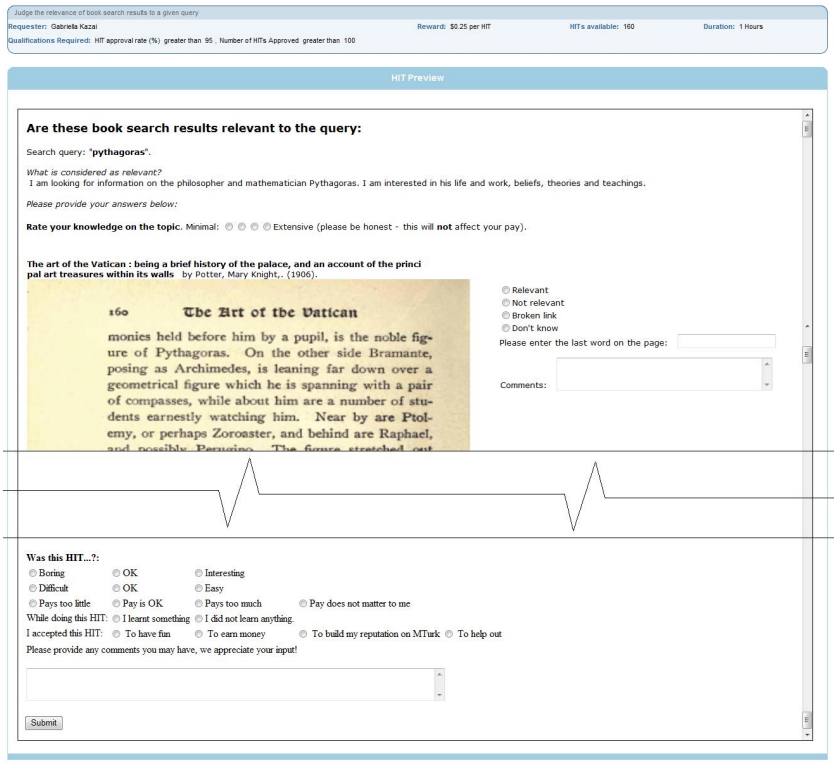


Fig. 2 Example HIT from the experiments on AMT. The ragged line represents a further 4 or 9 book pages included in a HIT, depending on the controlled 'effort' task variable

3.3 Experiment Variables

Our objective is to correlate various aspects of the task design with the quality of the crowdsourced labels and the human factors that characterize the crowd workers who chose to engage in the task. We use label accuracy, defined in terms of the agreement with our trusted gold set labels in the INEX test set, as an objective measure of the quality of the crowdsourced labels and observe how various combinations of the task properties and crowd characteristics relate to accuracy. Specifically, we explore two types of variables: the *task conditions*, reflected in the properties of the HIT design, and the *human factors* derived from the self-reported information of the crowd workers that was collected as part of the HITs.

3.3.1 Task Conditions

We focus on three key attributes of a crowdsourcing task: pay, effort and qualifying criteria. For each of the three attributes we investigate two settings, giving us a 2^3 experiment grid, corresponding to 8 distinct task conditions:

- *Pay* – We experiment with two levels of pay, i.e., the level of compensation offered to workers who successfully complete the task, paying $0.10 or $0.25 per HIT[5]. We will refer to these as c10 and c25 conditions.
- *Effort* – We vary the effort required to complete the task through the number of pages included in a HIT: requiring workers to label either 5 pages (p5) or 10 pages per HIT (p10).
- *Qualifying criteria* – We leverage AMT's worker pre-filtering feature that incorporates worker reputation measures and use two settings: open call where we require no qualifying criteria to be met by workers to gain access to the HITs (noQ) or restricting access to workers with over 95% HIT approval rate and over 100 approved HITs (yesQ).

Based on the above experiment grid, we used the data of our selected 8 topics, with 100 book pages per topic, in two series of experiments, using HIT templates that differ in the number of book pages included in the HIT (Figure 2). Series 1 involved a partition of the 800 pages into groups of 5 pages per HIT, resulting in 160 distinct HITs. Series 2 used groups of 10 pages per HIT, resulting in 80 distinct HITs. In both series we repeated the experiments under the combination of the two pay conditions and the two worker qualification conditions, paying 5 or 10 cents per HIT and drawing workers from an unrestricted pool of AMT workers or from a subset of workers who pass our pre-filter. We use the term *batch* to indicate the collection of HITs that corresponds to the full data set (800 pages) in a single experiment under the combined pay, effort and qualifying conditions. Therefore, we have 4 batches of HITs in each of the two series. In each batch, we requested labels from 3 different workers per HIT, resulting in 800x3=2,400 labels per batch, contributing to the total number of 19,200 labels collected for the 8 topics and 800 pages. Table 1 lists the resulting 8 batches of HITs, grouped by the level of pay offered, the pre-filter criteria used, and the effort required expressed in terms of the number of pages per HIT. The naming of the batches reflects the combination of three task conditions: pay—c10 or c25, qualifying criteria—noQ or yesQ, and effort—p5 or p10.

3.3.2 Human Factors

In order to expand our understanding of the characteristics of the workers who engage in crowdsourcing, we designed a short survey style questionnaire and included it in the HIT design as part of the task, see Figure 2. Specifically, we focused on the aspects of the workers' experience that are related to the task they performed:

- *Motivation* – We asked workers to indicate their main reason for accepting the HIT: 'To have fun', 'To earn money', 'To build my reputation on MTurk', 'To help out'. We will refer to these categories as Fun, Fortune, Fame and Fulfillment, respectively [41,46].
- *Familiarity* – We asked workers to rate their familiarity with the subject of the topic for which relevance labels were sought in the HIT. We used a 4 point scale (0-3) with 'Minimal' and 'Extensive' as end points. To encourage truthful answers, we emphasized that their answer would not affect their pay.
- *Task difficulty* – We asked workers' opinions on the difficulty of the task, with rating options of 'Difficult', 'OK', or 'Easy'.
- *Interest in the task* – We asked workers to indicate whether, in their opinion, the task was 'Boring', 'OK', or 'Interesting'.

[5] These pay levels are based on preliminary experiments with increasing pay up to the point that a satisfactory uptake was realized.

Table 1 Batches of HITs with different task parameter settings (pay, worker qualification criteria and required effort)

Batch	Pay	Qualif.	Effort	HITs	Assignments	Judged pages	Cost
c10-noQ-p5	$0.10	no	5 pages	160	480 (608)	2,400 (3,040)	$52.80
c10-noQ-p10	$0.10	no	10 pages	80	240 (460)	2,400 (4,600)	$26.40
c10-yesQ-p5	$0.10	yes	5 pages	160	480 (722)	2,400 (3,610)	$52.80
c10-yesQ-p10	$0.10	yes	10 pages	80	240 (358)	2,400 (3,580)	$26.40
c25-noQ-p5	$0.25	no	5 pages	160	480 (592)	2,400 (2,960)	$132.00
c25-noQ-p10	$0.25	no	10 pages	80	240 (299)	2,400 (2,990)	$66.00
c25-yesQ-p5	$0.25	yes	5 pages	160	480 (480)	2,400 (2,400)	$132.00
c25-yesQ-p10	$0.25	yes	10 pages	80	240 (304)	2,400 (3,040)	$66.00
Total				960	2,880 (3,823)	19,200 (26,220)	$554.40

– *Satisfaction with the pay* – We asked workers' opinions on the fairness of pay, indicating whether they were paid 'Too little', 'OK', 'Too much', or if 'Pay did not matter' to them.

3.4 Filtering Spam and Resulting Data Sets

As described in Section 2.3, crowdsourcing tasks can fall pray to dishonest workers who take shortcuts in completing HITs without due care in order to increase their personal gain, e.g., attain higher income from a larger number of submitted HITs. Due to the growing concerns around such cheating and random behavior [14,58,56], the filtering of unreliable output has become a norm in crowdsourcing and most crowdsourcing platforms enable requesters to reject work when it does not meet expected standards.

In our case, we applied an automatic filtering of HITs that did not meet our quality assurance criteria. We rejected assignments based on two observations: (1) the time spent on completing a HIT, and (2) the ratio of filled in captcha fields and the total number of captchas required. In particular, we first flagged workers as potentially unreliable if they completed less than 30% of the required captcha fields and spent less than 20 seconds per book page on average over all their tasks. Individual HIT assignments of flagged workers were then rejected if they failed the same tests calculated per HIT this time. Rejected HITs did not incur a charge and were republished on AMT for other workers to complete. The thresholds were determined while data was arriving based on a clear observed separation of poor and reasonable levels of performance.

The number of accepted HIT assignments and the total number of performed HITs, including the rejected ones (in brackets), are shown in Table 1. As we are interested in characterizing the workers' behaviors, in our analysis we consider all the data collected during the experiments and obtain statistics for the following three data sets:

– *All* – Comprises all the HITs and collected page labels, including the rejected HITs.
– *Cleaned* – Comprises only the approved HITs and corresponding page labels that passed our quality assurance criteria and were, thus, deemed reliable.
– *Rejected* – Comprises only the HITs that were rejected.

3.5 Analysis Methods

In the following sections we introduce the relevant terminology and describe the methods of analysis used in the rest of the paper.

3.5.1 Basic Concepts and Terminology

As discussed in the previous section, our experiments were divided into 8 batches that involve the labeling of the same set of 800 pages under different task conditions by 3 different workers per HIT. As a result, we effectively ended up with eight self-assembled crowds of workers, four from the whole population of AMT workers (no qualifying criteria) and four from the population of experienced and reputable AMT workers (requiring qualifying criteria).

Unlike in a standard laboratory setting, we did not control the distribution of workers across the HITs, for example, by assigning the same group of three workers to label pages across two different task conditions. Although such a setup would be possible to achieve, it would mean that workers no longer decided for themselves the task conditions under which they would choose to engage. The essence of our investigation is to uncover the impact that task characteristics have on the properties of the self-assembled crowd and the quality of their output. At the same time, we did not restrict workers to complete HITs only in one particular batch either. This approach has its benefits and its drawbacks. The key benefit is the realistic data. However, this comes at the price of reduced control over the samples of crowd workers involved in the different task conditions. This, in turn, restricts the types of statistical analyses that we can apply to provide conclusive evidence for the observations made about the crowd workers.

For the purposes of our study, we differentiate between workers who engaged with HITs under different task conditions, and we introduce the notion of a *worker instance*. When a worker completes a HIT under a given condition, that engagement is counted as a worker instance. If the same worker performs HITs in multiple batches, each engagement is counted separately, as distinct worker instances. This is important because the information collected via the questionnaires embedded in the HITs is specific to the task conditions and may differ across the HIT assignments for the same worker. In our analysis, we also collate information related to individual workers across all their HIT instances.

3.5.2 Measuring Label Accuracy

Since all the book pages included in the HITs have already been labeled for relevance by trusted judges from the INEX research community, we use these labels as our gold standard set to assess the accuracy (Acc) of the labels collected through crowdsourcing. For any set of judged pages, e.g., in a batch or per task condition, we can calculate the ratio of the total number of correct labels and the total number of required labels:

$$Acc = \frac{\sum_{set} page\_with\_correct\_label}{\sum_{set} page}. \quad (1)$$

We note that during the experiments, 45 of the 800 book pages were not reliably displayed by the Book Search server and, in some instances, appeared as missing when the HIT was displayed. Thus, for simplicity, we accepted the 'Broken link' option as an additional correct answer in these cases.

In order to explore the impact of task variables and human factors on the accuracy of label assignments, in addition to the above set based accuracy, we also aggregate the label precision statistics as follows:

- *HIT Accuracy (H_Acc)* – We calculate accuracy over the 5 or 10 pages included in a single HIT instance first and then average over the HITs included in a given set, e.g., batch or task condition.

– *Worker Accuracy (W_Acc)* – We calculate accuracy over the set of pages across all the HITs that a given worker completed. This accuracy reflects the worker's reliability. We then average these across the workers in a given set, e.g., workers who contributed to a batch or workers who are motivated by Fun.

Unlike Acc, worker accuracy is not effected by the skewed distribution of workers to HITs that is typical in crowdsourcing, i.e., workers who complete many HITs can dominate the average performance measured by Acc.

The relationship between the set based accuracy and the HIT accuracy for a set of HITs is more subtle: in cases when all the HITs in the set have the same number of pages, the two statistics are identical. In other words, grouping pages into sets of 5 and 10 has no impact. However, if we consider labels for a group of HITs with different effort, the two accuracies are different: the set based measure serves as a micro average, disregarding the page groupings, while H_Acc computes the macro averages, considering the groupings. In cases when we merge batches with 5 page and 10 page HITs, we will have twice the number of 5 page HITs and these will, therefore, dominate the average in H_Acc.

3.5.3 Statistical Analysis

Within each batch we have 3 distinct workers judging the same HIT. Thus, for each page we have a total of 24 labels over the 8 batches and, due to the experimental design, we have 12 labels for each of the task variable values (e.g., pay level of $0.10). We can compute the accuracy statistics over these sets against the gold standard labels. We can also calculate the level of agreement among the workers using Fleiss' kappa [18], which ranges from 0 to 1. Although there is no strict interpretation of the agreement levels, as it depends on the context in which it is applied, Landis and Koch [37] use a five level scale: slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–0.99) agreement.[6] In our case, Fleiss kappa can be applied only for each individual batch and the three task conditions. Computing agreement levels, for example, for a group of workers who share the same motivation for HITs, would mean considering unequal sets of workers and labels per page.

When calculating statistical significance values, we proceed as follows. When we compare two different HIT batches or task conditions (e.g., c10 vs. c25) we have equal size samples in terms of HIT instances. Thus, we conduct an equal variance test; specifically, the F-test comparing the variance of the two samples. If the test is positive, we have an equal sample size and equal variance and can apply the standard Student's t-test. If the test is negative, we have equal sample size with unequal variance and we apply the Welch's t-test. We report significance levels at $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***), e.g., as in Table 4.

When we consider the relationship of human factors and label accuracy (per HIT or per worker) in a sample, we are dealing with multiple responses. For that reason, we use the one-way ANOVA test. Specifically, we first test for the homogeneity of variances using Bartlett's test. If that test succeeds ($p < 0.05$), we issue the ANOVA test. If that test succeeds we test for the normality of the residuals using the Shapiro test. Essentially, when we combine HITs across batches, in order to consider human factors per worker across the task conditions of pay and effort, we do not have properly balanced samples. Therefore, we have to be cautious and perform statistical tests only under the conditions that approximate

[6] Some implementations test significance against the hypothesis that the ratings are completely random (kappa = 0), which we pass in all cases.

Table 2 Number of HITs, unique workers, average time spent per page, and accuracy (per batch, per HIT, and per worker) of the crowdsourced relevance labels across the different batches, corresponding to different task combinations of pay (c10 or c25), worker qualification criteria (noQ or yesQ) and effort (p5 or p10).

#	Batch	HITs	Wkrs	Time	Acc	H_Acc (sd)	W_Acc (sd)	Kappa
All data								
1	c10-noQ-p5	608	70	42	0.60	0.60 (0.27)	0.53 (0.26)	
2	c10-noQ-p10	460	69	26	0.35	0.35 (0.32)	0.50 (0.27)	
3	c10-yesQ-p5	722	66	42	0.61	0.61 (0.27)	0.59 (0.22)	
4	c10-yesQ-p10	358	35	42	0.59	0.59 (0.21)	0.62 (0.23)	
5	c25-noQ-p5	592	71	51	0.52	0.52 (0.28)	0.52 (0.25)	
6	c25-noQ-p10	299	58	41	0.52	0.52 (0.33)	0.53 (0.28)	
7	c25-yesQ-p5	480	43	61	0.71	0.71 (0.24)	0.66 (0.18)	
8	c25-yesQ-p10	304	54	33	0.67	0.67 (0.20)	0.65 (0.23)	
	Total	3,823	466	43	0.56	0.57 (0.29)	0.57 (0.25)	
Cleaned								
1	c10-noQ-p5	480	66	51	0.62	0.62 (0.25)	0.56 (0.24)	0.21
2	c10-noQ-p10	240	63	42	0.59	0.59 (0.22)	0.54 (0.24)	0.17
3	c10-yesQ-p5	480	62	58	0.63	0.63 (0.26)	0.61 (0.20)	0.20
4	c10-yesQ-p10	240	33	59	0.60	0.60 (0.21)	0.67 (0.17)	0.19
5	c25-noQ-p5	480	68	61	0.55	0.55 (0.27)	0.54 (0.24)	0.17
6	c25-noQ-p10	240	54	49	0.64	0.64 (0.25)	0.58 (0.24)	0.26
7	c25-yesQ-p5	480	43	61	0.71	0.71 (0.24)	0.66 (0.18)	0.39
8	c25-yesQ-p10	240	48	38	0.70	0.70 (0.18)	0.71 (0.14)	0.35
	Total	2,880	437	54	0.63	0.63 (0.25)	0.60 (0.22)	0.25

the test's assumptions of independence, homogeneity, and normality. Consequently, we can apply statistical tests only in some instances.

Finally, when we compare the responses to different questions in the questionnaire (where responses can be nominal or ordinal) and aim to test for the relationships among different human factors, we use the chi-square test and the significance level of 0.05, two-tailed.

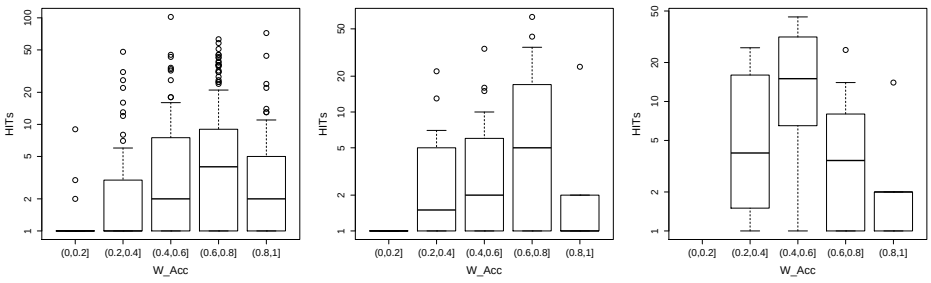
4 Overview of the Collected Data

In this section we provide an overview of the data collected from the crowdsourced tasks and statistics related to label accuracy.

4.1 Crowdsourced Labels

Table 2 presents overall statistics for the 8 batches of HITs that correspond to the different task condition combinations: effort is expressed in terms of 5 or 10 pages per HIT, pay is based on $0.10 or $0.25 reward per HIT, and the reliability of the ATM workers is defined as qualified or non-qualified workers. We report the number HITs, the number of workers, the average time spent per book page, the various label accuracy statistics calculated over all the collected data and over the cleaned data set, as well as Fleiss' kappa for the cleaned set.

Because of the uniform size of page sets per HIT within the batches, the batch accuracy (Acc) is the same as the average HIT accuracy (H_Acc). The worker accuracy (W_Acc) statistics are the averages of the accuracies achieved by the individual worker instances who



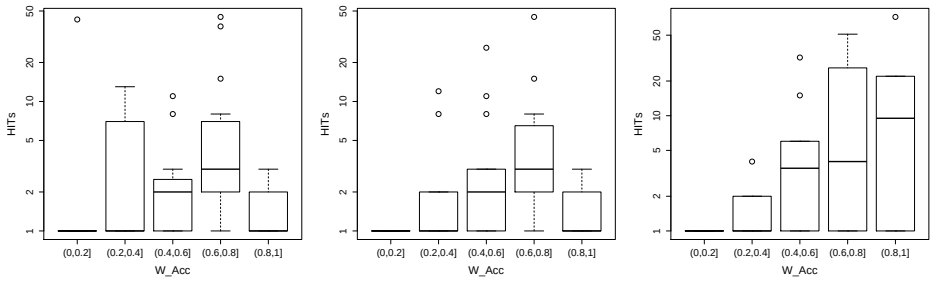
a) Cleaned b) `c10-noQ-p5` c) `c10-yesQ-p10`

Fig. 3 Distribution of HITs over worker accuracy bins for the whole cleaned data set (a), the cleaned `c10-noQ-p5` batch (b), and the cleaned `c10-yesQ-p10` batch (c).

performed HITs in a given batch. The relative differences between H_Acc and W_Acc are indicative of the distribution of the HITs to workers. For example, when H_Acc is greater than W_Acc, we expect that only a few well-performing workers contributed lots of high accuracy HITs. Conversely, when the H_Acc is lower than W_Acc, then a few badly performing workers contributed lots of low accuracy HITs. If we plot the distribution of HITs over W_Acc, the former results in a right-shifted worker-HIT distribution while the latter leads to a left-shifted worker-HIT distribution. As an example, Figures 3b) and c) show the right-shifted and left-shifted distributions as we plot the number of HITs completed by workers over increasing levels of accuracy for the cleaned `c10-noQ-p5` and `c10-yesQ-p10` batches.

As it can be seen in Table 2, we obtain label accuracy levels in the range of 52-71%, with the exception of `c10-noQ-p10` where accuracy is only 35% (all data). This is, perhaps, not so surprising since this batch was open to all workers (noQ) and presented the 'harshest conditions' where workers needed to judge 10 pages per HIT for only $0.10 payment. However, the higher value of W_Acc indicates that the low batch accuracy is a result of a handful of poorly performing workers who completed a lot of HITs with low accuracy. The increase in batch accuracy (to 59%) in the cleaned set tells us that these workers were in fact flagged by our spam filter as being unreliable and were consequently rejected. The best accuracy of 71% is obtained in the `c25-yesQ-p10` batch. Figure 4 shows the distribution of HITs over worker accuracies for these two batches.

We include the Fleiss kappa statistics computed for the workers agreement over the cleaned set. Although the values are in general low, they produce a similar relative ranking of the batches as the ranking by the batch accuracy; only batches `c10-noQ-p5` and `c10-yesQ-p5` swap places. This is significant since it indicates that, even without a gold standard set, we can assess the relative performance of the HITs by considering only agreement among the workers in assigning labels. Comparing the ranking of HIT batches based on worker accuracy and Fleiss' kappa, we see a disagreement mostly due to the pair of batches of `c10-yesQ-p5` and `c10-yesQ-p10`, which achieve relatively high worker accuracies (61% and 67%, respectively) compared with the overall average.



a) `c10-noQ-p10` (All) b) `c10-noQ-p10` (Cleaned) c) `c25-yesQ-p5` (Cleaned)

Fig. 4 Distribution of HITs over worker accuracy bins for the `c10-noQ-p10` batch over all data (a) and cleaned data (b), and the `c25-yesQ-p5` batch over cleaned data (c). The high volume of HITs completed by low accuracy workers, seen in (a) was removed by our spam filter, resulting in (b).

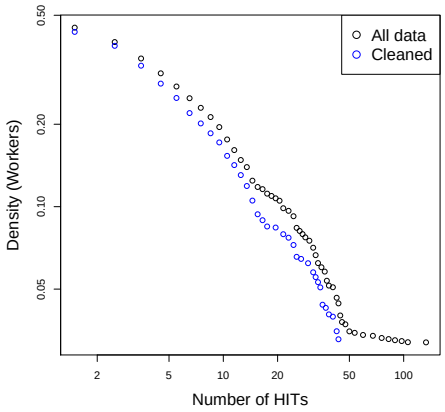


Fig. 5 Distribution of workers over the number of HITs (All and Cleaned sets), shown as the complementary cumulative distribution function (CCDF) $P(x) = Pr(X > x)$ showing the fraction of workers doing at least x HITs

4.2 Workers

In crowdsourcing, workers' participation in HITs often varies quite markedly. The majority would perform only a single HIT while a few may complete many HITs, resulting in the characteristic power law like distribution of HITs to workers. Indeed, similarly to previous findings [2,25], we observe that the distribution of HITs per worker is skewed, see Figure 5.[7]

In total, 348 distinct workers contributed to our experiments (all data) of which 52 workers had some or all of their HITs rejected. In fact, 19 workers were effectively removed because our spam filter (see Section 3.4) deemed all their contributions unreliable.

[7] We do not attempt to model the exact distribution. Instead, we look at the appropriate discrete powerlaw fit based on [10]. This powerlaw fit for all data is from 31 HITs onwards (i.e., $x_{min} = 31$, $\alpha = 3.08$, $D = 0.07229$) due to the discontinuity at the tail (spammers doing larger numbers of HITs than expected). For cleaned data, the fit is much better starting from 4 HITs onwards (i.e., $x_{min} = 4$, $\alpha = 1.93$, $D = 0.08019$).

Table 3 Number of unique workers in each of the conditions of the experimental matrix, e.g., c10 pay, and the number of workers under pairs of conditions, e.g., pay c10 and c25, and their total number of HITs

#	Task variable	Condition	Wkrs	Condition	Wkrs	Wrks in both		HITs	
						#	%	#	%
All data (348 workers and 3,823 HITs in total)									
1	Pay	c10	199	c25	197	48	13.8%	1,508	39.4%
2	Qualification	noQ	225	yesQ	155	32	9.2%	1,553	40.6%
3	Effort	p5	225	p10	189	66	19.0%	2,176	56.9%
Cleaned (329 workers and 2,880 HITs in total)									
1	Pay	c10	186	c25	184	41	12.5%	947	32.9%
2	Qualification	noQ	211	yesQ	148	30	9.1%	1,038	36.0%
3	Effort	p5	215	p10	175	61	18.5%	1,578	54.8%

Since workers could contribute to multiple batches, some may have performed HITs that fall under both conditions of a task variable, e.g., HITs that paid \$0.10 and HITs that paid \$0.25. Table 3 shows the overlap of unique workers in the pairs of the observed task conditions. Wee see that while the majority of workers (more than 80% in both All and Cleaned) only participated in a single batch, the overlap of workers in a pair of conditions ranges between 9% and 19% (both All and Cleaned). In addition, we see that those who worked on multiple batches tended to complete more HITs: 39–57% of all the HITs and 33–55% of the cleaned HITs were done by such workers. However, we note that workers who contributed to multiple batches tended to favor one condition over another. For example, workers who completed HITs in two batches (50 in All HITs and 48 in Cleaned), did on average 72% (in All and 70% in Cleaned) of their work in one batch (median 75% in All and 68% in Cleaned). Workers in more than two batches (26 in All and 23 in Cleaned) were less discriminating, concentrating only 55% (in All and 56% in Cleaned) of their work on HITs in one batch. Indeed, the correlation between the number of HITs and the number of batches that a worker would contribute to is highly significant (Pearson's correlation, $p < 0.001$ (All and Cleaned). This is a result of the skewed distribution of work in combination with the availability of the 8 batches of HITs, varying in different small ways. Hence workers interested in a large number of HITs could quickly appear in multiple batches. Interestingly, we can observe clear differences between the patterns of worker overlaps across the task variables. We see that the overlap of workers across the two pay conditions is lower than across the two effort conditions, implying that pay creates a clearer divide in the worker population. In contrast to the pay and effort conditions, participation under the two qualification conditions were more controlled in the sense that the workers who did not pass the qualification criteria did not have access to the yesQ batches of HITs. Admittedly, while this setup reflects the reality of crowdsourcing work, it raises complications in the interpretation of some of the findings, mixing the effects of the task conditions on the same type of worker with the effects of the self-selection of different types of workers to particular task conditions. On the other hand, our current study aims to focus on the workers' experiences that is specific to the given task and associated conditions. Our worker instances based analysis therefore aims to reflect the different behaviors of (the same or different) workers under different task conditions. Overall, we count a total of 466 worker instances in the full data set and 437 in the cleaned set.

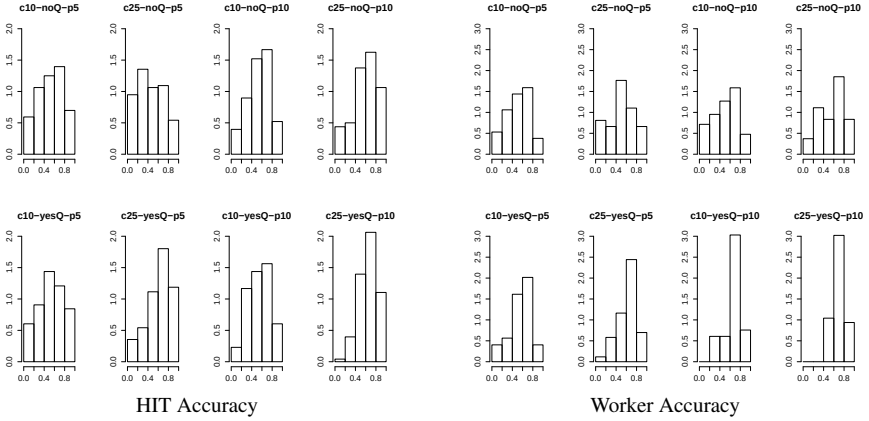
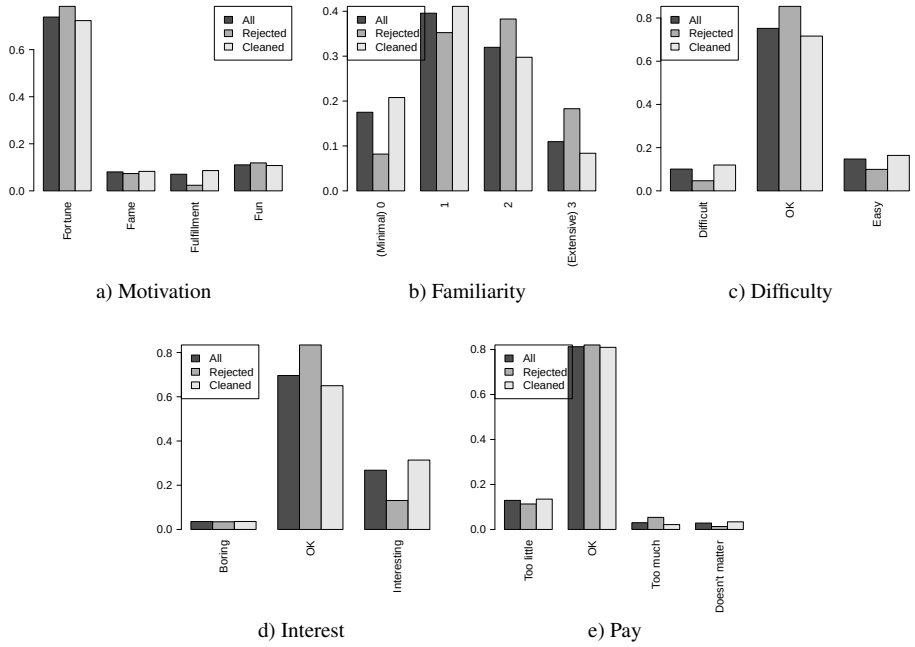


Fig. 6 Distribution of HIT Accuracy and Worker Accuracy for each of 8 batches for the Cleaned HITs

4.3 First Impressions of the Collected Data

Figure 6 shows two types of histograms for each of the 8 batches (Cleaned data): (a) the percentage of HIT assignments that achieved a particular level of accuracy and (b) the percentage of workers who achieved a given level of accuracy across all their labels within the HITs that belong to a particular batch. We observe that none of the conditions is dominated by low quality HITs or low performing workers. In fact, the majority of HITs and workers are in the middle range of label accuracy.

Crowd workers participate in HITs with varying motivation and produce results of mixed quality. In order to gain insights into the human factors that may affect label accuracy, and in particular, the workers' experience and perception of the task, we analyze the self-reported information from the questionnaires embedded in the HITs. Figure 7 shows the total number of HITs across all the batches that include a specific answer to the question of motivation, familiarity, perceived difficulty, interest in the task, and fairness of pay. We report these statistics for All, Cleaned and Rejected HITs. The histogram in Figure 7a suggests that the majority of workers, across all three data sets, are motivated by earning money. Interestingly, rejected workers seem to be rarely motivated by fulfillment, i.e., a desire 'to help out'. Figure 7b shows that the majority of workers reported to be moderately familiar with the subject matter of the topic for which they labeled book pages. Yet, rejected workers claim noticeably higher levels of familiarity. We speculate that this may be an effect of random clicks or a concerted effort to provide a socially more desirable response. In response to the question about the perceived task difficulty, the majority of workers answered that the task was neither difficult nor easy (Figure 7c). Again, the rejected workers seem particularly likely to select this middle-ground option. Similar findings are observed for the responses about their interest in the task (Figure 7d): most rejected workers chose the middle-ground response. Finally, Figure 7e shows that the majority of responses expressed satisfaction with the offered pay. Among those who stated that they were paid too much, we see a relatively high number of rejected workers. These findings are suggestive of differences in the behavior of workers under different task conditions, which we explore in more detail in the next sections.



a) Motivation b) Familiarity c) Difficulty

d) Interest e) Pay

Fig. 7 Workers' feedback on human factors over All the collected data, the Rejected data, and the Cleaned set: motivation (a), familiarity (b), task difficulty (c), interest in the task (d), and satisfaction with pay (e)

5 Effect of Task Variables on Label Accuracy

In this section we focus on the research question *R1–How do the different conditions of pay, effort, and selection of workers based on proven reliability affect the quality of the crowdsourced relevance labels?* (Section 1). To answer this question, we conduct two types of statistical analysis and hypothesis testing for the average label accuracy:

- First, by considering paired batches of HITs that vary in exactly one task variable (e.g., pay, or effort, or worker qualification), keeping the other two variables fixed, e.g., c10-noQ-p5 vs. c25-noQ-p5. This leads to 12 hypothesis tests involving corresponding pairs from the 8 batches.
- Second, by considering cross-batch collections of HITs that belong to a given task variable and corresponding parameter value, e.g., all the c10 HITs vs. all the c25 HITs, disregarding changes in the other task variables. This leads to 3 hypothesis tests (for pay, effort, and worker qualification).

We carry out hypotheses testing for the average label accuracy per HIT (H_Acc) and per worker (W_Acc). We note that, in the first type of tests, the average accuracy per HIT is equivalent to the average batch accuracy (Acc) since all the HITs in the pair of batches have the same number of pages per HIT. That does not apply to the collections of HITs considered in the second type of tests, which mix HITs with 5 and 10 pages.

Table 4 shows the results of the statistical significance tests for H_Acc and W_Acc in the observed pairs of conditions, keeping the remaining task variables fixed. For example, the tests for the pair c10-noQ-p5 and c25-noQ-p5, show that the Student t-test does not apply

Table 4 Results of statistical significance tests on pairs of task conditions, W=Welch, S=Student test, $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***), or p value is shown

				All data		Cleaned	
#	Var	Batch1	Batch2	H_Acc	W_Acc	H_Acc	W_Acc
1	Pay	c10-noQ-p5	c25-noQ-p5	W***	W, p=0.85	W***	W, p=0.66
2	Pay	c10-noQ-p10	c25-noQ-p10	W***	W, p=0.54	S, p=0.05*	W, p=0.42
3	Pay	c10-yesQ-p5	c25-yesQ-p5	S***	W, p=0.08	W***	W, p=0.16
4	Pay	c10-yesQ-p10	c25-yesQ-p10	W***	W, p=0.62	S***	W, p=0.24
5	Qualif.	c10-noQ-p5	c10-yesQ-p5	W, p=0.48	W, p=0.12	W, p=0.62	W, p=0.24
6	Qualif.	c10-noQ-p10	c10-yesQ-p10	S***	W, p=0.02*	W, p=0.82	S**
7	Qualif.	c25-noQ-p5	c25-yesQ-p5	S***	S**	S***	W**
8	Qualif.	c25-noQ-p10	c25-yesQ-p10	S***	W, p=0.02*	S, p=0.00**	S***
9	Effort	c10-noQ-p5	c10-noQ-p10	S***	W, p=0.62	S, p=0.19	W, p=0.66
10	Effort	c10-yesQ-p5	c10-yesQ-p10	S, p=0.33	W, p=0.51	S, p=0.14	W, p=0.10
11	Effort	c25-noQ-p5	c25-noQ-p10	S, p=0.88	W, p=0.73	W***	W, p=0.38
12	Effort	c25-yesQ-p5	c25-yesQ-p10	S**	W, p=0.77	S, p=0.34	S, p=0.12

Table 5 Results of statistical significance tests on pairs of task conditions, W=Welch, S=Student test, $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***), or p value is shown

#	Var	Val1	Val2	H_Acc1	H_Acc2	H_Sig.	W_Acc1	W_Acc2	W_Sig.
All data									
1	Pay	c10	c25	0.55	0.60	W***	0.55	0.58	W, 0.22
2	Qualif.	noQ	yesQ	0.50	0.64	S***	0.52	0.63	S***
3	Effort	p5	p10	0.60	0.52	S***	0.56	0.57	W, 0.89
Cleaned									
1	Pay	c10	c25	0.61	0.64	W**	0.58	0.61	W, 0.18
2	Qualif.	noQ	yesQ	0.59	0.66	S***	0.55	0.66	S***
3	Effort	p5	p10	0.63	0.63	S, 0.67	0.58	0.61	W, 0.15

(due to unequal variance) but the Welch test confirms statistical significance for the difference in H_Acc with $p < 0.001$ and no significant difference in W_Acc (All and Cleaned).

Table 5 reports outcomes of the statistical significance tests applied to the aggregated batches of HITs for a given task variable. As in Table 4, we report the test results for the average H_Acc and W_Acc, both for All and Cleaned data.

In Table 6 we show detailed statistics per aggregated HIT batches, providing further insights about the distribution of the observed statistics. In addition to label accuracy, we report the Fleiss kappa statistics, showing the level of agreement among the workers who participated in those HITs.

In the following sections, we use the information in Tables 4, 5, and 6 to discuss the impact of the task variables on the achieved accuracy of the crowdsourced relevance labels.

5.1 Effect of Pay on Label Accuracy

The statistical significance tests in the first four rows of Table 4 show that the level of pay affects HIT accuracy over all the observed conditions in both All and Cleaned sets ($p \leq 0.05$). On the other hand, the difference in pay does not produce significantly different worker accuracy levels ($p > 0.05$). This means that although workers with similar levels of accuracy contribute to both pay conditions, different quality workers contribute different volumes of

Table 6 Number of HITs, number of unique workers, average time spent per page, and accuracy of the crowdsourced relevance labels across the different subsets, corresponding to different task variable values

#	Var	Value	HITs	Wkrs	Time	Acc	H_Acc (sd)	W_Acc (sd)	Kappa
All data									
1	Pay	c10	2148	240	39	0.52	0.55 (0.29)	0.55 (0.25)	
2	Pay	c25	1675	226	49	0.60	0.60 (0.28)	0.58 (0.25)	
3	Qualif.	noQ	1959	268	41	0.48	0.50 (0.31)	0.52 (0.26)	
4	Qualif.	yesQ	1864	198	45	0.64	0.64 (0.24)	0.63 (0.22)	
5	Effort	p5	2402	250	48	0.60	0.60 (0.27)	0.56 (0.24)	
6	Effort	p10	1421	216	34	0.52	0.52 (0.30)	0.57 (0.26)	
Cleaned									
1	Pay	c10	1440	224	53	0.61	0.61 (0.24)	0.58 (0.22)	0.21
2	Pay	c25	1440	213	55	0.65	0.64 (0.25)	0.61 (0.22)	0.27
3	Qualif.	noQ	1440	251	52	0.60	0.59 (0.26)	0.55 (0.24)	0.20
4	Qualif.	yesQ	1440	186	56	0.66	0.66 (0.24)	0.66 (0.18)	0.29
5	Effort	p5	1920	239	58	0.63	0.63 (0.26)	0.58 (0.22)	0.24
6	Effort	p10	960	198	47	0.63	0.63 (0.22)	0.61 (0.22)	0.25

work under the two conditions, resulting in, overall more accurate labels under the higher pay condition, where more accurate workers perform more HITs.

From Table 6 (rows 1–2) we see that, regardless of the other task variables, increasing pay from \$0.10 to \$0.25 per HIT leads to a significantly improved label quality (W***, Table 5) both over All HITs (52% vs. 60%, resp.) and over the Cleaned HITs (61% vs. 65%, resp.). Unlike previous reports, e.g., [43], this finding suggests that pay does impact the quality, and not just the quantity, of the performed work. The results for the Cleaned set points out that this difference is not only due to sloppy work but is inherent to the overall attitude of workers toward the offered reward.

Considering the quality of the collected labels in terms of the number of reliable HITs and correct labels per pay conditions, we see that more inaccurate labels are contributed when pay is lower. Indeed, during HIT filtering, 4,293 incorrect labels (29% of the collected labels) were removed from the c10 set compared to 1,990 incorrect labels (17% of the collected labels) from the c25 set. Thus, we find that pay affects label quality also indirectly, where higher pay leads to fewer unreliable HITs and incorrect labels.

Looking at the average time that workers spent on judging a page, we find that workers, over All HITs, spent 39 seconds per page in the lower pay condition vs. 49 seconds under the higher pay condition. Interestingly, after excluding rejected work, the difference is reduced to only 2 seconds, i.e., 53 and 55 seconds, respectively. This means that the difference in the average time spent across the two pay levels was mainly a result of the faster HIT completions by spam workers in the c10 set. However, we are aware that the time spent per page is influenced by a range of factors, including the level of expertise and the ratio of filled captcha fields, which makes time a more complex signal and a weaker predictor of label accuracy in our case.

Comparing quality between different pay per label sets (instead of pay per HIT), confirms the same trend: quality increases with pay. However, we can also observe evidence of a diminishing return effect [9]: Acc for pay of \$0.01 per page is 60% , increasing to 62% for pay of \$0.02 per page and then to 67% for \$0.025 per page pay, but then drops back down to 63% when pay is \$0.05 per page (Cleaned data). Worker accuracy also seems to stagnate beyond \$0.025 pay per page: 58% for \$0.01-0.02 and 61% for \$0.025-0.05 pay.

Looking at the intersection of batches for c10 and c25 for qualified and unrestricted selection of crowd workers, we find that, with increasing pay, accuracy (Acc) increases

from 45% to 52% for noQ and from 60% to 69% for yesQ on All HITs (not in table). On the Cleaned data set, Acc increases with pay from 61% to 71% for yesQ. However, accuracy actually drops as pay increases for the noQ workers: 51% for c25-noQ compared with 61% for c10-noQ. A closer look at the data revealed that this performance dip is due to a relatively high number of HITs (115) contributed by workers who filled in over 30% of captcha fields, and therefore passed our spam filter, but labeled over 80% of the pages as relevant and thus introduced incorrect labels. In comparison, under the other conditions, the number of HITs with over 80% of the answers being relevant were significantly fewer: 5 HITs in the c10-noQ subset, 27 in c10-yesQ and 40 HITs in c10-yesQ. This is further confirmed by the observed levels of worker accuracy, which consistently increased with increased pay: from 55% to 58% on All HITs data set, for both qualified and unrestricted crowd workers, and from 58% to 61% on the Clean HITs data set. Thus, we find that pay affects both groups equally, encouraging better work regardless of the AMT qualifications.

Overall, we conclude that higher pay can induce two different types of behavior in workers: it may encourage better work, especially among qualified workers, but it may also attract more unethical workers, especially among workers with no reputation to protect.

5.2 Effect of Pre-Filtering by Worker Qualifications on Label Accuracy

Rows 3–4 in Table 6 show that workers who passed our qualifying criteria (95% acceptance rate and over 100 HITs on AMT) and, thus, have a proven crowdsourcing record, achieve significantly (S***, Table 5) higher levels of worker accuracy than non-qualified workers: 52% for noQ vs. 63% for yesQ on All HITs and 55% for noQ vs. 66% for yesQ on the Cleaned HITs. Comparing W_Acc to H_Acc, we we see that, for noQ workers, H_Acc=50% $<$ W_Acc=52% (All data), suggesting that more HITs are completed by poorly performing noQ workers. Most of these low quality HITs are removed by our spam filter, resulting in H_Acc=59% $>$ W_Acc=55% for noQ in the Cleaned set. For the qualified crowd we observe the opposite: H_Acc=64% $>$ W_Acc=63% (All data), i.e., more HITs are performed by accurate workers.

The difference in batch accuracy levels (Table 6, Acc column) for All and Cleaned HITs clearly shows that non-qualified workers contribute more inaccurate labels achieving an Acc of 48% vs. 64% for yesQ. However, after filtering out unreliable HITs, we observe a diminished benefit of pre-filtering workers, i.e., Acc has a smaller gain from 60% for noQ to 66% for yesQ (Cleaned data). In fact, our relatively insensitive filter, which removed 41% of the total labels from the noQ set (including 78% incorrect labels), was less effective for the yesQ crowd, removing 33% of the total labels (but including only 47% incorrect labels). The latter highlights the removal of a rather large percentage of false negative labels in the case of yesQ HITs, suggesting the need for spam filtering methods that can be adjusted based on the type of workers involved.

The average time spent per page suggests that qualified workers took more care when performing assessments, investing more time than noQ workers, i.e., 52 seconds per page for noQ vs. 56 seconds for yesQ (Cleaned set). Interestingly, we also found that under-performing workers (in the rejected HITs) among the qualified crowd spent longer per page than under-performing noQ workers: 10.37 seconds vs. 8.58 seconds, respectively. Comparing the HITs they performed, we see that noQ workers would simply submit HITs without answering all the questions, thus contributing 2,303 missing relevance labels. On the other hand, yesQ under-performers, still with an AMT reputation to protect, would fill in most of the information in the HITs, albeit clicking randomly on the multiple choice relevance la-

bels. Only 125 labels were missing from the HITs performed by yesQ workers. Both groups of under-performers would, however, ignore the captcha fields: qualified ones skipped 90% of captcha fields while unqualified skipped 93%.

Overall, we found evidence that AMT's reputation system is effective in selecting workers who produce more accurate relevance labels. On the other hand, we demonstrate that even a simple HIT filter, based on time spent per page and the percentage of filled captcha fields, can be as effective in identifying unreliable workers as the reputation system.

5.3 Effect of Effort on Label Accuracy

Rows 5–6 in Table 6 show the effect of varying effort on label accuracy. Based on the batch accuracy statistics calculated over All data, we may conclude that significantly better results are produced for 5 page HITs, i.e., 60% Acc vs. 52% for HITs with 10 pages (S***, Table 5). However, the worker accuracy levels in both All and Cleaned sets suggest that a higher performance is actually achieved for the higher effort HITs: 61% worker accuracy for p10 vs. 58% for p5. That said, the relative improvements in accuracy scores as a result of our spam filtering highlights that a higher volume of inaccurate labels are contributed in the p10 HITs.

Interestingly, we find that workers spent, on average, more time judging a page in the low effort HITs. Over All HITs, workers spent on average 48 seconds per page for p5 HITs vs. 34 seconds per page for p10 HITs. A similar trend holds for the Cleaned HITs with 58 seconds per page in p5 HITs vs. 47 seconds in p10 HITs. However, as noted before, time is a less reliable indicator of user behavior and label quality in our case because of the very design of the HITs. Workers may spend more time on reading a book page and considering their judgment or simply filling in a captcha field.

Overall, we observe that the effort involved in a HIT has a varied effect on the quality of the workers' output: higher effort HITs result in more inaccurate labels, but with a removal of unreliable HITs p10 HITs lead to better accuracy over p5 HITs. A possible caveat here is that the observed improvement may be a result of the spam filter being more effective on the HITs with more exposure.

6 Human Factors and Label Accuracy

In this section we turn our attention to the second research question: *R2–How do various human factors relate to the label accuracy?* (Section 1). To answer this question, we analyze the self-reported information about the workers' experiences with the task, collected via the questionnaire embedded in each HIT assignment (see Section 3.3.2).

Our objective is to assess the relationship between the five human factors—workers' motivation for completing the HIT, familiarity with the topic, interest in the task, perceived difficulty of the task, and opinion of the fairness of the offered pay—and the label accuracy achieved by the workers.

Table 7 summarizes the results of our analysis. For each human factor, we consider the number of HIT assignments in which the particular response appears, the number of individual workers who chose such a response in at least one assignment, the average time these workers spent per book page, and the accuracy statistics Acc, H_Acc, and W_Acc for both All and Cleaned sets. Since each human factor may assume one of the multiple values, we use an ANOVA test to establish the significance of the factor values for explaining the

Table 7 Number of HITs, number of workers, average time spent per page, and various accuracy scores of the crowdsourced relevance labels, grouped by responses to the human factor questions

	HF	Value	HITs	Wkrs	Time	Acc	H_Acc (sd)	W_Acc (sd)
All data								
1	Motivation	Fun	391	92	43	0.44	0.49 (0.34)	0.55 (0.26)
2	Motivation	Fame	285	65	44	0.43	0.49 (0.34)	0.56 (0.25)
3	Motivation	Fortune	2,615	290	41	0.59	0.60 (0.27)	0.58 (0.25)
4	Motivation	Fulfillment	250	63	65	0.58	0.59 (0.27)	0.57 (0.26)
5	Familiarity	0	599	163	51	0.61	0.61 (0.29)	0.55 (0.29)
6	Familiarity	1	1,354	238	40	0.60	0.61 (0.26)	0.60 (0.21)
7	Familiarity	2	1,094	211	32	0.52	0.55 (0.31)	0.57 (0.24)
8	Familiarity	3	375	87	47	0.45	0.48 (0.32)	0.45 (0.26)
9	Interest	Boring	123	64	47	0.47	0.48 (0.33)	0.52 (0.30)
10	Interest	OK	2,394	294	38	0.57	0.59 (0.29)	0.59 (0.24)
11	Interest	Interesting	921	216	56	0.56	0.56 (0.28)	0.58 (0.24)
12	Ease	Difficult	324	96	51	0.56	0.59 (0.31)	0.51 (0.26)
13	Ease	OK	2,420	302	41	0.57	0.58 (0.29)	0.59 (0.25)
14	Ease	Easy	474	124	48	0.55	0.56 (0.30)	0.58 (0.25)
15	Pay	Too little	413	104	53	0.45	0.49 (0.35)	0.56 (0.25)
16	Pay	OK	2,596	307	41	0.60	0.61 (0.28)	0.60 (0.24)
17	Pay	Too much	96	14	22	0.34	0.35 (0.26)	0.39 (0.17)
18	Pay	Don't care	91	34	71	0.54	0.56 (0.27)	0.51 (0.27)
Cleaned								
1	Motivation	Fun	286	84	55	0.60	0.62 (0.27)	0.58 (0.23)
2	Motivation	Fame	220	58	55	0.60	0.62 (0.27)	0.58 (0.23)
3	Motivation	Fortune	1,922	267	53	0.64	0.64 (0.24)	0.61 (0.21)
4	Motivation	Fulfillment	229	60	70	0.62	0.62 (0.26)	0.57 (0.25)
5	Familiarity	0	526	146	56	0.67	0.66 (0.25)	0.61 (0.24)
6	Familiarity	1	1,040	223	50	0.64	0.64 (0.24)	0.62 (0.20)
7	Familiarity	2	753	194	42	0.63	0.62 (0.25)	0.59 (0.21)
8	Familiarity	3	212	70	75	0.58	0.56 (0.26)	0.55 (0.19)
9	Interest	Boring	93	56	59	0.61	0.59 (0.27)	0.55 (0.28)
10	Interest	OK	1,674	276	50	0.65	0.65 (0.25)	0.62 (0.21)
11	Interest	Interesting	808	199	63	0.61	0.61 (0.25)	0.60 (0.21)
12	Ease	Difficult	285	88	57	0.65	0.66 (0.26)	0.54 (0.25)
13	Ease	OK	1,707	283	54	0.64	0.64 (0.25)	0.62 (0.21)
14	Ease	Easy	391	114	55	0.64	0.63 (0.26)	0.61 (0.23)
15	Pay	Too little	318	94	65	0.61	0.62 (0.27)	0.60 (0.22)
16	Pay	OK	1,908	288	52	0.66	0.66 (0.24)	0.63 (0.21)
17	Pay	Too much	51	11	30	0.39	0.38 (0.26)	0.40 (0.16)
18	Pay	Don't care	80	29	80	0.60	0.61 (0.25)	0.52 (0.27)

variance in label accuracy across the HITs and workers. We present our results below in relation to five specific questions of interest.

Are workers who complete HITs for fun more accurate in assigning relevance labels than others? Focusing on the batch accuracy for the HIT assignments in which a worker expressed the particular motivation category: fun, fame, fortune, or fulfillment (Table 7), we see that, for All data, the lowest Acc is obtained when workers declare fun (44%) or fame (43%) as their motivation, i.e., when the objective is entertainment or building a good reputation on AMT. A possible explanation is that fun, while a powerful motivator, is not in fact a guarantee of quality in itself. Indeed, in the ESP game [1], the built-in quality control mechanism is inherently tied in with the game itself: two independent workers need to provide the same label for an image to earn points. In our case, however, there is a clear difference between the goal of the requester to get all pages in a HIT labeled and the entertainment value that a HIT may provide. Since it is more likely that only some of book pages in a HIT

would be of interest to most workers, workers who were mainly motivated by fun are more likely to skip or take less care with pages that are of little interest to them. This of course could be overcome by breaking down HITs to contain fewer pages, where workers would have the opportunity to only pick those HITs that would maximize their entertainment value (ignoring the effort involved to find such HITs). However, this has immediate implications for costs and for quality control. Clearly, when fewer pages are included in a HIT, the relative ratio of pages with existing gold set labels increases, thus increasing the associated costs. Alternatively, randomly placing gold set pages across all HITs can keep costs down, but decreases the chances of detecting unethical workers.

On the other hand, we see that the corresponding W_Acc levels are similar across all motivation categories (55-58%), indicating that the low batch accuracy for Fun and Fame is, in fact, a result of a small number of poorly performing workers. Indeed, statistics for the Cleaned HITs show only a slight difference both in Acc and W_Acc, with Fortune and Fulfillment resulting in highest Acc levels of 64% and 62% and Fortune in highest worker accuracy of 61%. A one way ANOVA test ($p < 0.001$) confirms that the relationship between motivation and HIT accuracy is significant over All HITs but not on the Cleaned data. In both sets, best performance is achieved by workers motivated by monetary rewards.

Considering the distribution of the motivation responses across HITs assignment and workers, it is clear that individuals are by far the most motivated by Fortune, which is hardly surprising considering that AMT is a labor market based on monetary rewards. It is also interesting to see a relatively high average time spent per page by workers motivated by fulfillment: 65 seconds vs. 41-43 seconds in other categories for All HITs and 70 seconds vs. 53-55 seconds for the Cleaned HITs. From Figure 7a we saw that the Fulfillment motivation was rarely associated with rejected HIT assignments, suggesting that these workers did in fact invest the time to provide quality results.

Are workers who claim to be familiar with a given topic more accurate in judging relevance? Accuracy statistics reported in rows 5–8 of Table 7 show a trend that is opposite to what we may normally expect. The lowest accuracy levels are achieved by the workers who claim to be 'experts' (familiarity level of 3 on the scale of 0–3), achieving only 45% batch accuracy for All HITs and 58% over the Cleaned HITs. In contrast, those with the lowest reported familiarity level achieved the highest Acc: 61% for All HITs and 67% on the Cleaned set. The one way ANOVA test, $p < 0.001$ shows statistically significant relationship of Familiarity and Acc for All HITs. On the Cleaned HITs, the data fails the homogeneity condition and, hence, ANOVA cannot be applied. The worker accuracy statistics show a similar trend but peak at the familiarity level 1 with 60% W_Acc for All HITs and 63% for the Cleaned data.

There are many reasons that may have caused this result, contrasting previous research on the impact of assessors' knowledge and the self-efficacy of AMT workers ([6, 16]. First, Familiarity, in our case, was self reported in circumstances that may have encouraged workers to represent their competence in a better light (despite assurances that familiarity with the topic would not affect pay). Second, the responses could reflect both the confidence and the attitude of the crowd workers involved in the task. In order to investigate this issue in more detail, we analyzed separately the responses from the workers who passed the qualifying requirements (yesQ) and those from the unrestricted AMT population (noQ). We found that noQ workers who highly rated their Familiarity with the topic performed the worst, achieving only 33% Acc on All HITs. At the same time, the yesQ workers, who declared themselves as highly Familiar with the topic, achieved 60% Acc on All HITs. In general, we see a negative correlation between self-reported Familiarity and accuracy for noQ workers. Indeed, Acc drops from 56% for the Familiarity level 0 (minimal knowledge of the topic)

to 55% (level 1), 42% (level 2) and to 33% for level 3 (extensive knowledge). For yesQ workers, Acc is more uniform across all Familiarity levels, ranging from 63% to 67% on All data. This may indicate that more veteran workers on AMT are better at gauging their own level of expertise while new workers may be more prone to effects of satisficing [29].

Are workers who find the task interesting more accurate in assigning relevance labels? Rows 9–11 in Table 7 show that the lowest Acc is achieved by those workers who find the task boring: 47% Acc on All HITs compared to 56–57% accuracy when some level of interest is declared. This is also confirmed by the obtained worker accuracy levels: 52% (All data) and 55% (Cleaned) W_Acc for workers bored with the task vs. 58-59% (All) and 60-62% (Cleaned) for workers with some interest in the task. Interestingly, workers with moderate interest (OK) in the task spent the least amount of time per page: only 38 seconds vs. 47-56 seconds (All data) and 50 seconds vs. 59-63 seconds (Cleaned data). A closer look at both the self-reported Familiarity and Interest levels reveals that the workers who reported higher levels of Familiarity generally reported the task less interesting, which may explain their faster pace of work.

Analyzing the HIT assignments, we find that low performing 'bored' workers (in the rejected HITs) contributed incomplete HITs with lots of missing information. On the other hand, 'bored' workers in the Cleaned set, despite their lack of interest, took time and care to complete the HITs. Interestingly, under-performing (spam-filtered) workers who were 'Interested' in the task labeled more pages per HIT and spent more time per page. Thus, despite their interest and good intentions, these workers were not able to assess the relevance of the pages accurately. Finally, the under-performing workers who selected 'OK' for their interest level spent the shortest amount of time per page and mostly assigned labels randomly. The data collected about the workers' Interest fails the homogeneity condition for both All HITs and Clean HITs; hence we could not apply the ANOVA test.

Are workers who find the task easier more accurate? From the batch accuracy statistics for the Ease factor in Table 7 (rows 12–14), we see no relation between the reported levels of task difficulty and Acc: 55-57% over All HITs and 64-65% in the Cleaned set. However, the W_Acc statistics reveal a clear drop in performance when workers find the task challenging: 51% W_Acc for 'Difficult' vs. 58-59% otherwise for All HITs and 54% vs. 61-62% for Cleaned HITs. We also note a slight increase in the average time spent per page when the task is found difficult: 51 seconds v. 41-48 seconds for All HITs. This confirms the findings of Shaw et al. [51] who studied the effects of different social and financial incentive schemes but found that results were mainly dependent on the difficulty of the task. Since the data for the Ease factor fails the homogeneity condition for both All HITs and the Clean HITs, we could not apply the ANOVA test.

Are workers who are happy with their pay more reliable in their work? Rows 15–18 in Table 7 report the accuracy statistics for the groups of workers who expressed their satisfaction with Pay as one of four levels: 'Too little', 'OK', 'Too much', and 'Don't care'. We see that those worker who are content ('OK') with the pay produce the highest accuracy: 60%Acc for All HITs and 66% Acc for the Cleaned set. This is consistent with the W_Acc statistics that are also the highest when the workers reported that the pay was 'OK': 60% for All HITs and 63% for Cleaned HITs. The lowest Acc and W_Acc are obtained when workers reported 'too much pay', e.g., Acc of 34% (All) and 39% (Cleaned). We expect that this partially results from random clicking by unethical workers. The relation between workers' satisfaction with Pay and HIT accuracy is significant over All HITs (one way ANOVA test, $p < 0.001$) and over Cleaned HITs ($p < 0.001$). This is in concert with our previous observation that the majority of workers are motivated by the financial gain and shows that the

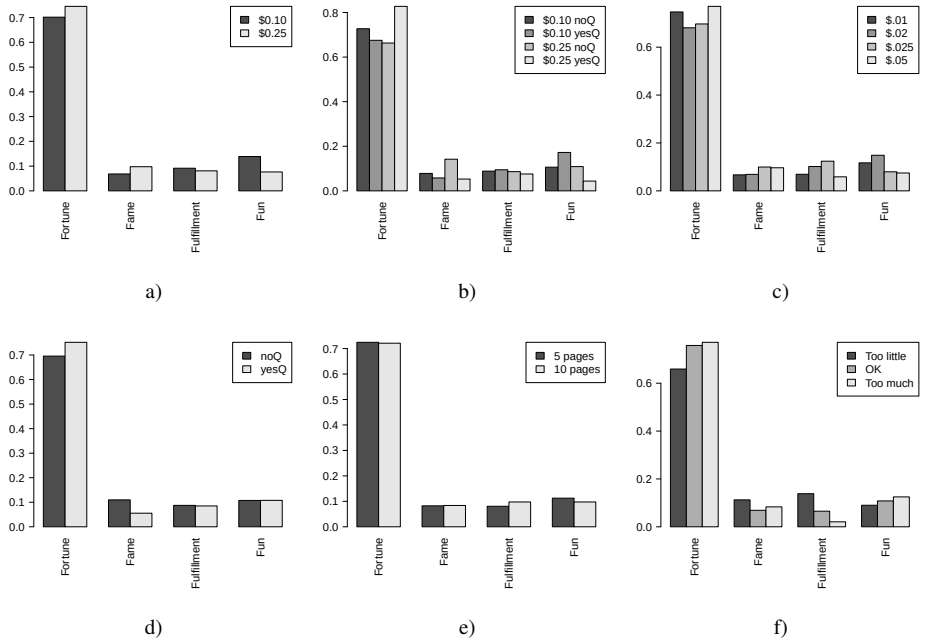


Fig. 8 Workers' feedback on the motivation for the tasks paid at \$0.10 and \$0.25level subsets (a), further broken down by the workers qualification (noQ and yesQ) (b), and by the effort of p5 and p10, resulting in different pays per page—\$0.01, \$0.02, \$0.025, \$0.05 (c). Also showing motivation for noQ and yesQ workers (d), effort as a number of pages per HIT (e), and over self-reported responses to satisfaction with pay (f)

.

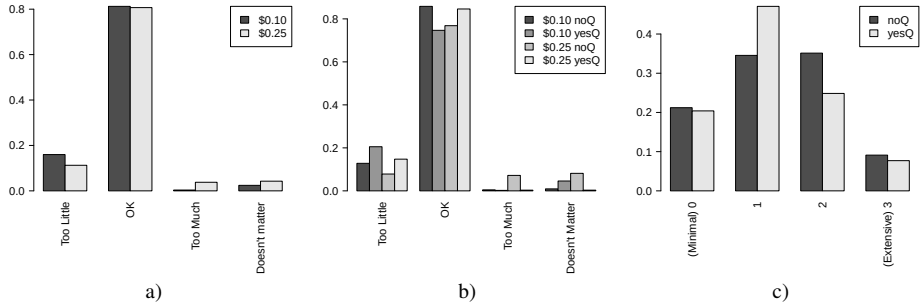
level of workers' satisfaction with the offered pay can explain the variability in the accuracy they achieve.

7 Characteristics of Crowd Workers in Different Task Conditions

In this section we investigate the impact that the task variables, i.e., pay, effort, and the requirement for a proven work record on AMT, may have on the properties of the crowds that chose to work on the task. We present our analysis of the self-reported human factors, e.g., motivation or perceived task difficulty, separately for each task variable in the following sections.

7.1 Pay and Human Factors

Do different pay levels attract people with different motivations? Figure 8a shows that the distribution of motivation categories for workers in the c10 and c25 HITs are fairly similar. One noticeable difference is that the c10 condition has a slightly higher percentage of workers who are motivated by Fun compared to the c25 condition. At the same time, we also observe a higher percentage of 'Fortune' responses in the c25 HITs than in the c10 HITs. This may be expected as the lower pay is less likely to attract workers who are primarily



a) b) c)

Fig. 9 Workers' feedback on the amount of pay per HIT in the $0.10 and $0.25 pay-level subsets (a) and further broken down by worker qualifications (b). Workers' Familiarity level (0='Minimal', 3='Extensive') for noQ and yesQ workers (c)

.

interested in earning money. It may also indicate a cognitive dissonance effect [17] where workers who feel uncomfortable with their level of pay, will subconsciously seek other motivations such as fun or fulfillment.

When separating workers by qualifications in Figure 8b, we see an interesting difference in the two pay conditions: in c10 more yesQ workers reported Fun as motivation than noQ workers. In the case of the c25 pay level, more yesQ than noQ workers declared Fortune as their motivation. Amongst workers motivated by Fortune, we see that more noQ workers complete lower paid c10 HITs while more yesQ workers perform c25 HITs.

Due to the different combinations of pay (c10 and c25) and effort (5 and 10 pages), we, in effect, have four levels of pay per page: $0.01, $0.02, $0.025, and $0.05. The breakdown shown in Figure 8c, indicates a slight increase of the Fortune motivation with increased pay, and relatively more Fun for the low paying HITs. This confirms a shift in motivation to Fortune from other motivating factors as pay increases. Indeed, a chi-square test over responses per worker also confirms that pay and motivation are significantly related ($p < 0.0001$) for both All HITs and Cleaned data.

Looking at the self-reported responses on satisfaction with the offered pay per motivation categories, in Figure 8f, we find that workers who feel underpaid are much more frequently motivated by fulfillment relative to those who feel overpaid, which is consistent with a cognitive dissonance effect.

Are better paid workers more satisfied with their pay? Unsurprisingly, we find that workers were more satisfied with higher pay, see Figure 9a. At the same time, the majority of the responses indicated that workers were content ('pay is ok') with the offered pay in both pay categories: 63% in the c10 condition and 70% in the c25 HITs. This suggests that workers are indifferent to pay in the studied spectrum: from $0.05 down to $0.01 per page judgment. Only 2% of the responses indicated that workers involved in low pay HITs did not care about pay, compared with 4% of the responses in the higher paid HITs.

Interestingly, among All HITs, we see that workers who are happier with the level of pay actually spent less time on the task: 40 seconds (pay is 'too little'), 32 seconds (pay is ok), 24 (pay is 'too much'), while workers who did not care about pay spent 47 seconds per page. The relatively short average times in this set are, however, largely due to the unreliable HITs. In the Cleaned data, the time spent per page varies between 47 and 59 seconds (no correlation with satisfaction with pay), where the longest time is invested by workers who

'don't care' about pay: 71 seconds per page, on average. Thus the crowd of workers who do not care about pay are made up of two distinct groups: under-performing workers who can be easily filtered out and diligent workers who take the task seriously.

Do better paid workers find the tasks more interesting? We find a very similar pattern and no significant difference in workers' interests in the task for the two pay levels in both All and Cleaned sets, e.g., 3/65/22 and 3/59/27 percentage of the responses indicated low/ok/high interest for the c10 and c25 conditions, respectively (All). This may however be more a result of the way workers on AMT can locate HITs of potential interest to them. The main mechanism is by browsing lists of available HITs, which workers can filter using keywords or criteria on minimum pay. The keywords entered by the worker are matched against the keywords assigned to the HITs by the requester. A worker then needs to inspect each potential HIT of interest before accepting it. Since this may be a rather laborious task, which also takes up the workers' time, workers may feel pressured into accepting HITs that may or may not be of interest to them, regardless of the level of pay.

To investigate this further, we also ask whether the workers who are more satisfied with the offered pay are also more interested in the task? Among the subset of HITs with 'too much pay' or 'pay does not matter' responses, workers indicated high interest (65% of responses) or 'ok' interest (31%) in the task. This ratio is very different for workers who selected 'pay is ok' as their feedback: only 20% of the responses expressed high interest and 71% indicated moderate (ok) interest. On the other hand, when workers were unsatisfied with the pay ('pay is too little'), 31% of the responses registered high interest, and 59% moderate interest. This again highlights the duality of pay and interest, where high interest compensates for low (or even no) pay, e.g., as demonstrated by the ESP game [1]. A chi-square test over responses per worker also confirmed that satisfaction with pay and reported interest in the task are significantly related ($p < 0.05$) for both All and Cleaned data. So, although we did not find a direct relationship between workers' interests in the task and the level of pay offered, we can see that workers who are less happy about their pay are more interested in the task and vice versa.

Are more knowledgeable workers drawn to the higher pay condition? We find no evidence to suggest that higher paid HITs draw in more self-confessed experts. This is somewhat expected as workers of different expertise get allocated to HITs as and when they come across them on the AMT dashboard. It may, however, be more appropriate to assume that more veteran AMT workers (e.g., qualified workers) have a better chance of grabbing higher paying HITs as they are more familiar with the ways in which HITs can be located on AMT.

Do better paid workers perceive the task easier? We find few differences between workers' perception of the task difficulty in the two pay conditions in the All data set. However, in the Cleaned set, 76% of the responses registered the tasks as OK or Easy in the higher paid condition, compared with only 70% for the c10 HITs. Since neither the knowledge levels of workers, nor their interest levels showed difference over the two pay groups, this suggests that when workers are paid better, they may anticipate that more is required of them and thus rate the task easier in relation to this expectation.

7.2 Pre-selection of AMT Workers and Human Factors

Do qualified and non-qualified workers differ in their motivations? Figure 8d shows that a noticeably higher number of non-qualified workers are motivated by fame, i.e., trying to obtain a reputation on the AMT. This is again a demonstration of the reputation system at

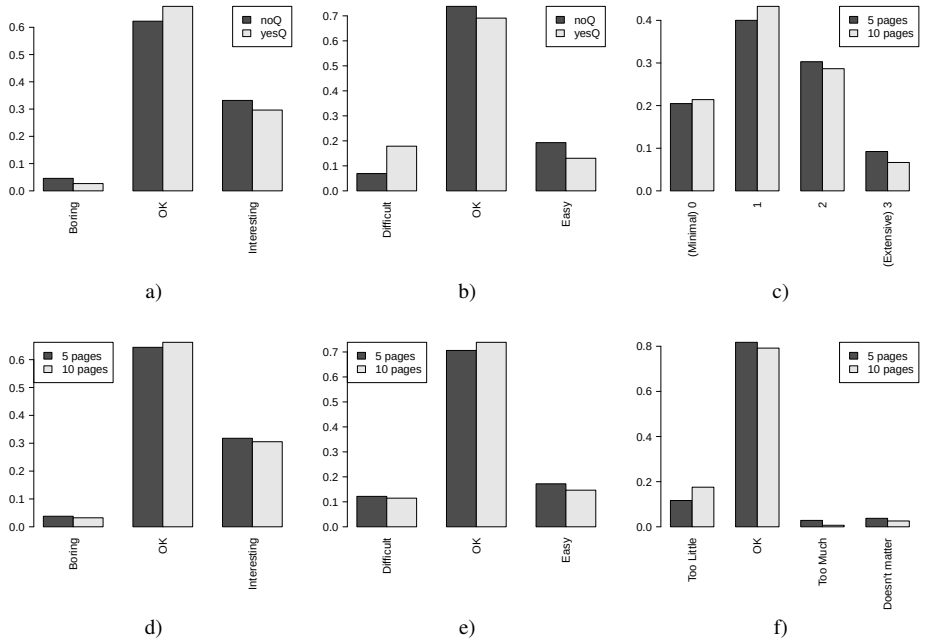


Fig. 10 Workers' feedback on human factor question over qualifications and effort: interest over qualifications (a), task difficulty over qualifications (b), familiarity over effort (c), interest over effort (d), task difficulty over effort (e), and pay over effort

work. Qualified workers, who already established their reputation, are more motivated by earning money, thus capitalizing on their earlier efforts.

Are non-qualified workers happier with their level of pay? We find that non-qualified workers are happier with the offered pay than qualified workers, perhaps somewhat surprisingly even more so in the lower pay conditions. Qualified workers were more disgruntled by lower pay than non-qualified workers: this is indicated by the drop in 'pay is ok' responses in the yesQ batches (48%) compared with 78% in the noQ batches, see Figure 9b. Thus it seems that qualified workers have higher expectations on pay, while non-qualified workers estimate the value of their work at a lower rate. This can be related back to the finding that qualified workers in the higher pay condition are more motivated by fortune (Figure 8b), thus setting their expectations accordingly.

Are qualified workers more interested in the task? Since qualified workers get a larger pick of HITs (that are open to them), it may be reasonable to assume that they are more likely to cherry-pick those HITs that are of more interest to them. Figure 10a shows the distribution of Interest responses for yesQ and noQ HIT assignments. Contrary to our expectation, we note a slightly higher level of reported interest from non-qualified workers.

Do the qualified workers find the task easier? Qualified workers have in general more experience with crowdsourcing work, so we may expect a lower perceived difficulty of the task by yesQ workers compared with noQ workers. However, Figure 10b shows that, in fact, qualified workers find the task more often difficult and less often respond with 'OK' or 'easy'.

Are qualified workers more knowledgeable? Looking at the reported levels of familiarity with the topic by qualified and non-qualified workers, we see that noQ workers tend to be more confident and report higher familiarity than qualified workers, see Figure 9c.

7.3 Effort and Human Factors

Does effort influence the type of workers who chose to work on the task? In the design of HITs, it is important to create tasks that are well received by the workers. Thus we ask ourselves about the influence that the required effort may have on the workers in selecting the task. As shown in Figure 8e, the distribution of responses to motivation are similar for two effort levels of 5 and 10 pages per HIT. Figure 10c shows the distribution of responses to familiarity with the subject matter of the task, showing slightly higher levels of Familiarity in the lower effort tasks. Considering the responses to the Interest question for each of the effort conditions, (Figure 10d), we note similar levels of interest for both p5 and p10 conditions.

Do workers feel that higher-effort tasks are more difficult? Figure 10e gives the distribution of responses to task difficulty for both 5 page and 10 page HITs. Again, we see basically the same levels of perceived difficulty for both effort conditions. Correlating with label accuracy and broken down by worker qualification criteria, we find that non-qualified workers who found the task difficult performed the worst (35% on All data, 50% on Cleaned), compared with 48-50% (All) and 60-61% (Cleaned) accuracy when noQ workers reported the task as not difficult. Qualified workers achieved a more consistent accuracy regardless of whether they found the task easy or difficult (68-72% for Cleaned data).

How do satisfaction with pay and required effort relate to each other? One can expect a relation between the amount of effort and workers' reported satisfaction with pay. Figure 10f shows the responses to satisfaction with the pay level over the p5 and p10 effort conditions. Although the 10 page HITs require essentially twice the effort, the majority of workers are happy with the level of pay offered. There is a mild change in responses with 'too little' being more frequent for the 10 page HITs and 'too much' being more frequent for the 5 page HITs. The relative insensitivity of the responses to effort is similar to that of the different pay level ($0.10 versus $0.25) as was shown in Figure 9a.

8 Discussions and Conclusions

In this paper we presented a large scale study of crowdsourcing tasks involving the gathering of relevance labels for book search evaluation. Our objective was to explore the influence that common task design decisions have on the quality of the workers' output and, at the same time, gain insights about the characteristics of the crowds who chose to work under specific task conditions and the human factors that shape their experience with the task. Here we summarize our main findings, reflecting on the methodological decisions, and provide recommendations to HIT designers.

8.1 Summary of Main Findings

8.1.1 Influence of the task conditions on the label accuracy

Unlike [43], we found that the level of pay does impact on label accuracy: higher pay encourages better work, especially among qualified workers; lower pay increases the number

of unreliable HITs and inaccurate labels. At the same time, higher pay may also attract more unethical workers, especially among those with no reputation to protect. This suggests that in applications where label quality is paramount, it may be worth paying workers more, but this should be balanced with richer quality control mechanisms built into the design (e.g., pre-filtering workers, captcha, or training [38]) and more sophisticated spam filters. Considering pay per label, rather than pay per HIT, we found evidence of diminishing return where accuracy levels flatten or even drop with increased pay. This is partly due to the fact that a higher pay attracts more sophisticated workers who are, nevertheless, dishonest in their attempt to achieve personal gain without performing quality work.

Increasing the effort required of workers in a HIT has two effects. On the one hand, higher effort HITs can result in higher volumes of inaccurate labels–however, these can be easily removed with the help of post-task quality control methods. On the other hand, higher effort tasks may be better at attracting more well-performing workers. In addition, increasing the number of pages to be judged in a HIT can also reduce the overall judging time and can thus increase productivity. At the same time, lower effort HITs can be more popular among workers, leading to faster overall task completion, e.g., our c25-p5 batches for both yesQ and noQ crowds completed in one to two days, while the c25-p10 batches took between 9-14 days (no such trend was observed in the c10 condition).

Restricting access to HITs to the reputable workers on AMT produces better quality work, providing evidence of the effectiveness of AMT's reputation system. On the other hand, we demonstrated that even a simple spam filter, e.g., based on time spent per page and the percentage of filled captcha fields, can be as effective in identifying unreliable workers as the reputation system. Furthermore, higher accuracy cannot be guaranteed just by pre-filtering workers as more sophisticated unethical workers can easily build a false reputation. At the same time, pre-filtering can exclude crowds of highly reliable workers who wish to build up their reputation on AMT. Our findings also highlighted the need to adopt different HIT designs and quality control methods (e.g., spam filters) to eliminate suboptimal contributions from dishonest workers within the two groups.

8.1.2 Influence of human factors on the label accuracy

Unsurprisingly, our analysis of the workers motivation revealed that earning money is by far the main reason for engaging in crowdsourcing work on AMT. However, we also found that those who are motivated by Fortune and Fulfillment are, in fact, the most accurate contributors of relevance labels. Workers who perform tasks for Fun or Fame turned out to be the least accurate for our task. This is in contrast to the success of human computing applications such as the ESP game [1] that are solely based on the combination of fun and fame. We expect that the difference is primarily due to the the dominant monetary aspects of the AMT platform and the fact that our tasks were optimized for utility rather than fun.

Self-declared information about Familiarity with the topics turned out to be unreliable. Low performing workers reported high levels of familiarity, perhaps, in their desire to create an impression of high competence, while individuals with low reported levels of familiarity achieved higher label accuracy. Our data, thus, seems contradictory to previous work [6], but a deeper analysis reveals that the phenomena is tied to the effect of self-declaration rather than true expertise of the users as well as the effect of spam workers' responses.

Considering the workers' interest in the task, we saw two distinct patterns. For All HITs, the individuals who find the task boring under-performed considerably. However, after the removal of unreliable HITs, workers who found the work boring actually achieved high

levels of accuracy. We found the perception of task difficulty indicative of workers' performance in terms of showing a clear drop in worker accuracy levels when workers reported the task challenging. Satisfaction with the pay was very strongly correlated with label accuracy. The highest accuracy was achieved by the workers who were content with the pay level and the lowest by those who reported 'too much pay'.

8.1.3 Characteristics of the crowds

Perhaps most insightful are the observations about the crowds that self-formed to complete the tasks under different effort and pay conditions. We saw that lower pay HITs had more workers reporting Fun as motivation. We found no significant differences between workers' levels of self-reported expertise or interests in the task across the two pay conditions. However, it turned out that HITs with higher pay were also regarded as more difficult. Finally, we saw that the higher pay led to higher satisfaction with the pay and vice versa. However, the majority of workers were contended with the offered pay for both 5 page and 10 page HITs. Given the substantial differences in pay per label, this relative insensitivity to the amount of work is striking.

Comparing the experiences of qualified and non-qualified workers, we saw that noQ workers claimed higher levels of familiarity with the topic, were slightly more interested in the task. and reported the task easier than qualified workers. Non-qualified workers were also happier with the offered pay. An interesting observation is that while qualified workers are supposed to be superior, the non-qualified workers' responses seem more attractive from a requester's point of view. We may question to what extent these answers are given truthfully, and to what extent are socially desirable answers provided. The output quality indeed suggests that qualified workers result in more reliable labels–although the difference is in fact minimal after spam filtering.

8.1.4 Conclusions of the findings

Generally, we are interested in the use of crowdsourcing for creating labeled data, in particular relevance judgments that are known to be subjective. The result of such a labeling task is complex combination of at least three factors: First, the self-formation of crowds, giving rise to a biased but unknown sample of workers. Second, the skewed distribution of HITs within the particular group of workers attracted to a batch of work, further skewing the output. Third, the HIT design promoting quality of work for a given worker completing a HIT. The main general conclusion of this paper is that task conditions do indeed attract a different crowd, and that these differences are affecting the quality of the work.

8.2 Design Recommendations

In this section, based on the empirical data from our experiments, we offer advice on practical issues that designers of crowdsourcing tasks may face.

How should I price my HIT? Our findings suggest that the best way to determine the appropriate level of pay is to estimate the price per unit of effort. This may be measured in terms of the expected time workers would need to spend on the HIT, taking into account the cognitive effort based on the difficulty of the task. As we saw, there is likely to be an equilibrium point on the pay per unit of effort scale: too little pay can be just as detrimental

to output quality as too much pay, although, evidently for different reasons. Paying too little can lead to sloppy work, mainly as a result of little commitment, while paying too much can attract sophisticated spammers who may be more difficult to detect and deter (see also e.g., [43,38,19,26]). The evolving general practice, at least on AMT, is to pay workers an hourly rate of $5–6, as suggested in the requesters' instructions on AMT. The actual rates will of course depend on the workers' pace. At the same time, many tasks pay less and, as a result, hourly rates of $1.67–1.97 were reported in [49].

As demonstrated by others, pay is related to the total time a crowdsourcing experiment takes to complete, e.g., [43]. While our findings confirm the same trend, we saw that the rate of completion is actually more influenced by the effort associated with a HIT: higher paid batches completed in 67% of the time than lower paid batches, but lower effort batches finished in half the time of the higher effort batches. This suggests again that it is the pay per unit of effort that is the key factor to be considered by requesters.

How should I package my task? It is essential to break-down tasks into easily digestible units for the workers. Smaller tasks are a better fit with AMT's crowdsourcing model of 'many hands make light work'. However, smaller tasks can attract more workers motivated by fun, who can be unreliable if the design of the HIT does not tie quality control directly into the game aspects [27].

The downside of shredding a task into many small units is that the pay per HIT is obviously reduced, which is an important factor in how workers on AMT find and select HITs. Obtaining fewer labels per HIT also reduces the ability to detect low quality work due to workers' limited exposure, and makes quality control mechanisms, like inserting pages with known labels from a gold set, more expensive. So a careful balance is needed. A positive outcome is that we found no particular correlation between effort and output quality–although packaging larger amounts of work in a HIT may make it less attractive, leading to longer batch completion times.

When posting HITs on AMT, an important aspect is to consider the representation and 'findability' of the HIT. This clearly concerns pay, but also aspects like HIT title and keywords. To increase the findability of our HITs, we included a wider range of keywords, describing the nature and subject matter of the task, e.g., "search, evaluate, relevance judgment, labels, books, history, Pythagoras, Dalai Lama, exorcism, Buddhism, ale, beer, Titanic, chess, fire of London". However, similarly to [25], we also observed that HIT completion is at its highest rate soon after a HIT batch is published, with attention trailing off as time passes. As a solution, Ipeirotis suggests to cancel and re-publish batches periodically.

How can I protect against spam workers? We summarized some of the most popular quality assurance and control mechanisms in Section 2.3. Using gold set data to monitor workers' performances [53,19] or obtaining multiple labels from different workers where noise is reduced through majority rule [26,3] are well-known techniques. Since, in this paper, we conducted a crowdsourcing meta-experiment with the goal of analysis, we had gold set labels available for all the judged pages. Thus, we opted to reject work on other grounds here, in particular, based on the time per page and the fraction of filled in Captcha fields[8] The effectiveness of such a simple filter highlights the usefulness of behavioral observations in quality control, e.g., [58,50].

[8] Rejecting workers based on their gold set agreement would have led to an artificial bias in the accuracy of the 'cleaned set', since the gold set is no longer an independent test of the resulting quality of work.

The main recommendation of Kittur et al. [36] was that a task should be given in such a way that cheating took approximately the same effort as faithfully completing the task. Thus crowdsourcing relevance labels using simple check-boxes is likely to be a suboptimal design, especially, as shown by [15], the use of multiple choice answers can attract spammers who can simply click randomly. Open-ended questions can be used to counter this effect, for example soliciting worker feedback [3]. Our design, showing scans of book pages, also had a simple and natural challenge-response task by asking workers to enter the last word of the scan, necessitating workers to check the pages as part of the task, functioning as a Captcha. Such a Captcha is a simple and inexpensive tool for serious workers, who can enter the information without having to do extra work since they have read the pages in order to make the required relevance judgment. It both deters sloppy workers, making the HIT unattractive for those who try to complete it as fast as possible, and gives us a good signal to detect sloppy work. We showed that rejecting work based on the fraction of Captcha fields actually completed, gave a substantial increase in accuracy. Clearly, such captchas are still rather rudimentary and could be passed without the desired effect. A better solution may be to ask workers to enter the first word of a relevant sentence on the page [33]. Recently, however, workers have been advised not to do HITs with Captchas since these may be scam jobs, for example trying to obtain credit card numbers, with no intention to pay for the work [52]. So it is essential to naturally embed a challenge-response element as part of the main task of the HIT. Other defensive-design methods, e.g., rewards linked to performance, trap questions, can be effective both in deterring and detecting spammers.

From the outcome of our experiments, we note that self-reporting as a tool can be used as an aid to identify and filter spam workers, particularly when such questionnaires are structured so that only certain combinations of answers make sense [33]. Related ideas were proposed in [29], who demonstrated that slowing down the presentation of survey questions increased comprehension and thus the quality of the collected data.

8.3 Reflections on the Research Method

There are two fundamentally different approaches to analytic crowdsourcing studies as presented in this paper, so it is worth discussing the methodological considerations in greater detail. On the one hand, one can view a crowdsourcing platform as a means to solicit respondents or test persons for surveys or experimental work. In this approach one would try to approximate a traditional controlled experiment as closely as possible, in particular, by controlling the workers inside the experiment, and by hoping that the sample of workers attracted to a task is representative enough for all practical purposes. One could design a HIT experiment setup with constrained access to HITs to ensure that the workers are assigned randomly or by stratum to particular conditions. In that manner we would control the number of HITs per worker, and ensure that workers under each conditions are drawn from the same population (see [7,42] for related ideas). On the other hand, one can view the crowdsourcing phenomenon as the object of study itself, and aim to observe workers "in the wild" as they engage in crowdsourcing tasks in the typical way that they are published on crowdsourcing platforms like AMT. Here an essential characteristic is that crowd workers voluntarily engage in HITs, leading to a skewed distribution of HITs over workers, where the specific tasks and task conditions may attract different populations of workers. In this case we should not control the assignment of workers to tasks, since the self-selection is an essential part of the market mechanism of crowdsourcing platforms.

There are pros and cons to either methods. A clear advantage of the first approach is methodological rigor. As with traditional studies, the control is exercised to give experimental power, and thus, has many advantages in terms of the analysis of the balanced data obtained. However, the generalizability of the results is potentially problematic, since in realistic crowdsourcing settings, we will have a very skewed work distribution, with learning effects as well as fatigue of workers doing many HITs, and very different types of workers such as spammers that tend to be interested in tasks with many HITs available (hence not taking part in a controlled setup). More generally, since it is unknown what biased samples of workers are attracted to the task, one should be cautious in assuming that findings in controlled conditions transfer to other cases. An advantage of using a setup resembling typical crowdsourcing tasks is its realism, since all the essential features of the work distribution are present within the experiment. However, this also poses challenges to the analysis of the data, specifically in the application of statistical measures that assume a balanced setup. Essentially we have to address three effects, i) the particular population of workers attracted to a particular task and task condition, ii) the skewed distribution of work between those engaging in the task under these task conditions, and iii) the effect of the task conditions on the task performance. Thus, we are faced with contradictory requirements imposed by strict statistical analysis on one side and the objective to inform about the behavior of the workers in their natural environment. Given the research questions of this paper, with the crowdsourcing setup as the object of study, we adopted the second approach. That is, we opted for the standard crowdsourcing setup but run experiments on a full matrix of a comprehensive set of HIT parameters (in particular pay, effort, and qualifications) that are crucial for anyone posting work on crowdsourcing platforms. In our analysis, we looked at measures both at the level of the performed work (HIT accuracy) that is representative for the quality of the crowdsourced work, and at the level of the workers involved (Worker accuracy) that is insensitive to the skewed distribution of work. In addition, we opted to contrast worker instances (workers under given task conditions).

Perhaps one of the main general insights of this paper is that the dynamics of the uncontrolled assignment have a far greater impact than we imagined beforehand. In particular, the crowds attracted to particular task conditions, even with only relatively subtle differences in pay, effort or qualifications, can lead to worker groups with radically different characteristics. This is a distinct property of the crowdsourcing setup, and a major factor affecting quality of work in typical crowdsourcing scenarios. Understanding this aspect of the crowdsourcing dynamics is one of the main open questions and, even if challenging in terms of research method and statistical analysis, deserves further study.

8.4 Concluding Remarks

In our future work, we plan to build a rich framework for modeling the comprehensive ecosystem of crowdsourcing platforms, comprising the populations of workers and requesters, and the platform owner. We aim to expand our focus through further considerations of the task design and the factors that influence the interactions and the outcome of the crowdsourcing engagements. The building blocks for such a model can be derived from the insights presented here. While in this paper we focused on the quality of the crowdsourced work in terms of label accuracy, the quality of relevance labels can be further attested through their use in system evaluation and ranking [33, 34]. Furthermore, we aim to explore a wider range of parameters to characterize the crowd and the crowd engagements, including factors such as the clarity of the task, aesthetics, pre-task qualification tests, seeding, workers' emotion,

and similar. Our ultimate goal is to provide crowdsourcing practitioners with a framework to guide the design of their crowdsourcing tasks in order to maximize label quality.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '04, pp. 319–326. ACM, New York, NY, USA (2004)
2. Alonso, O., Baeza-Yates, R.A.: Design and implementation of relevance assessments using crowdsourcing. In: Advances in Information Retrieval - 33rd European Conference on IR Research (ECIR 2011), *LNCS*, vol. 6611, pp. 153–164. Springer (2011)
3. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In: Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, pp. 557–566 (2009)
4. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. SIGIR Forum **42**(2), 9–15 (2008)
5. Alonso, O., Schenkel, R., Theobald, M.: Crowdsourcing assessments for xml ranked retrieval. In: Advances in Information Retrieval, 32nd European Conference on IR Research (ECIR 2010), *LNCS*, vol. 5993, pp. 602–606. Springer (2010)
6. Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., Yilmaz, E.: Relevance assessment: are judges exchangeable and does it matter. In: SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference, pp. 667–674. ACM, New York, NY, USA (2008)
7. Behrend, T.S., Sharek, D.J., Meade, A.W., Wiebe, E.N.: The viability of crowdsourcing for survey research. Behavior Research Methods (2011)
8. Carterette, B., Soboroff, I.: The effect of assessor error on ir system evaluation. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pp. 539–546. ACM, New York, NY, USA (2010)
9. Case, K.E., Fair, R.C., Oster, S.C.: Principles of Economics, tenth edn. Prentice-Hall (2011)
10. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review **51**(4), 661–703 (2009)
11. Cleverdon, C.W.: The Cranfield tests on index language devices. Aslib **19**, 173–192 (1967)
12. Cormack, G.V., Palmer, C.R., Clarke, C.L.A.: Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pp. 282–289. ACM, New York, NY, USA (1998)
13. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. Commun. ACM **54**, 86–96 (2011)
14. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the 28th international conference on Human factors in computing systems (CHI '10), pp. 2399–2402. ACM (2010)
15. Eickhoff, C., de Vries, A.P.: How crowdsourcable is your task? In: Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), pp. 11–14. ACM (2011)
16. Feild, H., Jones, R., Miller, R.C., Nayak, R., Churchill, E.F., Velipasaoglu, E.: Logging the Search Self-Efficacy of Amazon Mechanical Turkers. In: M. Lease, V. Carvalho, E. Yilmaz (eds.) Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010), pp. 27–30. Geneva, Switzerland (2010)
17. Festinger, L., Carlsmith, J.M.: Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology **58**(2), 203–210 (1959). `http://psychclassics.yorku.ca/Festinger/`
18. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**, 378–382 (1971)
19. Grady, C., Lease, M.: Crodsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 172–179 (2010)
20. Grimes, C., Tang, D., Russell, D.M.: Query Logs Alone are not Enough. In: E. Amitay, C.G. Murray, J. Teevan (eds.) Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007) (2007)
21. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com. In: Workshop on Future Internet and Next Generation Networks (FINGNet). Seoul, Korea (2011)

22. Howe, J.: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Publishing Group, New York, NY, USA (2008)
23. Ipeirotis, P.: Mechanical turk: The demographics. Blog post (2008). Http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html
24. Ipeirotis, P.: The new demographics of mechanical turk. Blog post (2010). Http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html
25. Ipeirotis, P.G.: Analyzing the amazon mechanical turk marketplace. XRDS **17**, 16–21 (2010)
26. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10, pp. 64–67. ACM, New York, NY, USA (2010)
27. Jain, S., Parkes, D.C.: The role of game theory in human computation systems. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09, pp. 58–61. ACM, New York, NY, USA (2009)
28. Kamps, J., Koolen, M., Trotman, A.: Comparative analysis of clicks and judgments for IR evaluation. In: Proceedings of the Workshop on Web Search Click Data (WSCD 2009), pp. 80–87. ACM Press, New York NY, USA (2009)
29. Kapelner, A., Chandler, D.: Preventing satisficing in online surveys: A 'kapcha' to ensure higher quality data. In: The World's First Conference on the Future of Distributed Work (CrowdConf2010) (2010)
30. Kasneci, G., Van Gael, J., Herbrich, R., Graepel, T.: Bayesian knowledge corroboration with logical rules and user feedback. In: Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part II, ECML PKDD'10, pp. 1–18. Springer-Verlag, Berlin, Heidelberg (2010)
31. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: Advances in Information Retrieval - 33rd European Conference on IR Research (ECIR 2011), *LNCS*, vol. 6611, pp. 165–176. Springer (2011)
32. Kazai, G., Doucet, A., Landoni, M.: Overview of the inex 2008 book track. In: INEX, pp. 106–123 (2008)
33. Kazai, G., Kamps, J., Koolen, M., Milic-Frayling, N.: Crowdsourcing for book search evaluation: Impact of quality on comparative system ranking. In: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2011)
34. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1941–1944. ACM (2011)
35. Kazai, G., Milic-Frayling, N., Costello, J.: Towards methods for the collective gathering and quality control of relevance assessments. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, pp. 452–459. ACM, New York, NY, USA (2009)
36. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08), pp. 453–456. ACM (2008)
37. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
38. Le, J., Edmonds, A., Hester, V., Biewald, L.: Ensuring quality in crowdsourced search relevance evaluation. In: V. Carvalho, M. Lease, E. Yilmaz (eds.) SIGIR Workshop on Crowdsourcing for Search Evaluation, pp. 17–20. ACM, New York, NY, USA (2010)
39. Lease, M.: On quality control and machine learning in crowdsourcing. In: Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI, pp. 97–102 (2011)
40. Lease, M., Kazai, G.: Overview of the trec 2011 crowdsourcing track. In: Proceedings of the Text Retrieval Conference (TREC) (2011)
41. Marsden, P.: Crowdsourcing. Contagious Magazine **18**, 24–28 (2009)
42. Mason, W., Suri, S.: Conducting behavioral research on amazons mechanical turk. Behavior Research Methods (2011)
43. Mason, W., Watts, D.J.: Financial incentives and the "performance of crowds". In: HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 77–85. ACM, New York, NY, USA (2009)
44. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: MIR '10: Proceedings of the international conference on Multimedia information retrieval, pp. 557–566. ACM, New York, NY, USA (2010)
45. Oppenheim, A.N.: Questionnaire Design and Attitude Measurement. Heinemann, London (1966)
46. Quinn, A.J., Bederson, B.B.: A taxonomy of distributed human computation. Tech. Rep. HCIL-2009-23, University of Maryland (2009)

47. Quinn, A.J., Bederson, B.B.: Human computation: A survey and taxonomy of a growing field. In: Proceedings of CHI 2011 (2011)
48. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval quality? In: J.G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D.A. Evans, A. Kolcz, K.S. Choi, A. Chowdhury (eds.) CIKM, pp. 43–52. ACM (2008)
49. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, pp. 2863–2872. ACM (2010)
50. Rzeszotarski, J.M., Kittur, A.: Instrumenting the crowd: using implicit behavioral measures to predict task performance. In: Proceedings of the 24th annual ACM symposium on User interface software and technology, UIST '11, pp. 13–22. ACM, New York, NY, USA (2011). doi: 10.1145/2047196.2047199. URL http://doi.acm.org/10.1145/2047196.2047199
51. Shaw, A., Horton, J., Chen, D.: Designing incentives for inexpert human raters. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW '11 (2011)
52. Silberman, M.S., Ross, J., Irani, L., Tomlinson, B.: Sellers' problems in human computation markets. In: Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10), pp. 18–21. ACM (2010)
53. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), pp. 254–263. ACL (2008)
54. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management **36**(5), 697 – 716 (2000)
55. Voorhees, E.M., Harman, D.K. (eds.): TREC: Experimentation and Evaluation in Information Retrieval. MIT Press (2005)
56. Vuurens, J., Vries, A.P.D., Eickhoff, C.: How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy. In: M. Lease, V. Hester, A. Sorokin, E. Yilmaz (eds.) Proceedings of the ACM SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR 2011), pp. 48–55. Beijing, China (2011)
57. Welinder, P., Branson, S., Belongie, S., Perona, P.: The multidimensional wisdom of crowds. In: J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (eds.) Advances in Neural Information Processing Systems (NIPS '10), pp. 2424–2432 (2010)
58. Zhu, D., Carterette, B.: An analysis of assessor behavior in crowdsourced preference judgments. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (2010)