In Search of Quality in Crowdsourcing for Search Engine Evaluation

Gabriella Kazai

Microsoft Research Cambridge
v-gabkaz@microsoft.com

Abstract. Crowdsourcing is increasingly looked upon as a feasible alternative to traditional methods of gathering relevance labels for the evaluation of search engines, offering a solution to the scalability problem that hinders traditional approaches. However, crowdsourcing raises a range of questions regarding the quality of the resulting data. What indeed can be said about the quality of the data that is contributed by anonymous workers who are only paid cents for their efforts? Can higher pay guarantee better quality? Do better qualified workers produce higher quality labels? In this paper, we investigate these and similar questions via a series of controlled crowdsourcing experiments where we vary pay, required effort and worker qualifications and observe their effects on the resulting label quality, measured based on agreement with a gold set.

Keywords: IR evaluation, relevance data gathering, crowdsourcing.

1 Introduction

The evaluation of a search engine's effectiveness typically requires sets of relevance labels, indicating the relevance of search results to a set of queries. However, the gathering of relevance labels is reportedly a time consuming and expensive process that is usually carried out by hired trained experts. For example, TREC[1] employs retired intelligence analysts as assessors.

Recently, crowdsourcing [5] has emerged as a feasible alternative to gather relevance data for the evaluation of search engines [2,1,4]. It promises to offer a solution to the scalability problem that hinders the traditional approaches, where either the size of the test collection or limitations on time and other resources are proving increasingly prohibitive. Crowdsourcing is an open call for contributions from members of the crowd to solve a problem or carry out human intelligence tasks (HITs), often in exchange for micro-payments, social recognition, or entertainment value. Crowdsourcing platforms, such as CrowdFlower[2] or Amazon's Mechanical Turk (AMT)[3] service, allow for anyone to create and publish HITs, and gather vast quantities of data from a large population within a short space of time and at a relatively low cost. However, crowdsourcing, and

[1] http://trec.nist.gov/
[2] http://crowdflower.com/
[3] https://www.mturk.com/

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 165-176, 2011.

more specifically crowdsourcing when monetary incentives are involved, is a solution with its own set of challenges [8,10]. Indeed, crowdsourcing has been widely criticized for its mixed quality output. Marsden, for example, argues that 90% of crowdsourcing contributions are rubbish [9]. On the other hand, several studies in relevance data collection concluded that crowdsourcing leads to reliable labels [1,4]. At the same time, works such as [13,8] provide evidence of cheating and random behavior among members of the crowd.

Clearly, the gathering of useful data requires not only technical capabilities, but also sound experimental design. This is especially important in crowdsourcing where the interplay of the various motivations and incentives affects the quality of the collected data [11,10]. The usefulness or the quality of the data will of course depend on the goals and the context of a given crowdsourcing experiment. For example, in IR evaluation, relevance labels contributed by non-experts have been shown to lead to different conclusions when comparing system performance [3]. Thus, when crowdsourcing relevance labels for the evaluation of IR systems, quality may be defined in terms of agreement with expert judges.

With the aim to obtain insights that can guide the design of crowdsourcing experiments, in this paper, we explore the relationship between the quality of the crowdsourced relevance labels and properties of the task design: the monetary reward it offers, the effort it demands, and the qualifications it requires. Specifically, we aim to answer the following research questions:

- Does quality pay? Or rather, can higher pay guarantee better quality results?
- Do more qualified workers produce higher quality labels?
- Are increased levels of required effort detrimental to output quality?

To answer these questions, we designed several batches of HITs on AMT where we varied pay, required effort and worker qualifications, and observed their effects on the quality of the gathered labels. Our analysis reveals intricate relationships between the examined properties, highlighting the need to study crowdsourcing experiments as complex ecosystems with implications for the design of HITs as well as of spam filters.

Next, we motivate our work and describe the data used in the experiments. Section 3 details the HIT design. Results and findings are given in section 4.

2 Experiment Setup and Data

For our experiments we chose the problem of building a test collection for the evaluation of focused retrieval approaches. This challenge is faced by the INEX Book Track, which has thus far struggled to meet this need by relying on its participants' voluntary efforts alone due to the scale of the problem [7]: "The estimated effort required of a participant of the INEX 2008 Book Track to judge a single topic was to spend 95 minutes a day for 33.3 days". Motivated by this need to scale up the Cranfield method for constructing test collections, our work explores the possibility of employing crowdsourcing methods to replace or compliment traditional modes of gathering relevance data. In particular, our

```
<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 from the INEX 2008 Book Track test set

goal is to study how reliable crowdsourcing is in terms of its output quality, what factors impact on the quality and how these influence each other.

Among the range of tasks investigated by the INEX Book Track, we chose the Focused Book Search (FBS) task, where users expect to be pointed directly to relevant book parts. In this task, systems return ranked lists of book parts (e.g., pages) estimated relevant to a topic. To evaluate the task, INEX collects relevance labels for a subset of the book pages retrieved by participating systems.

The data used in our experiments consists of the books, search topics, and relevance judgments of the INEX 2008 Book Track test collection. The corpus contains 50,239 out-of-copyright books (17 million pages), totaling 400GB. From the total of 86 available topics, we selected 8 topics (ID: 27, 31, 37, 39, 51, 57, 60 and 63) based on the number of available judgments in the INEX test collection. Figure 1 shows one of the topics. For our 8 topics, we have 470 judged books, of which 149 are relevant, and 4,490 judged pages, of which 1,109 are relevant. From these, we randomly picked 100 pages per topic ensuring a 40-60% ratio of relevant/irrelevant labels. These labels form the gold set for our AMT experiments, which are described next.

3 HIT Design

We designed several batches of HITs on AMT with the following objectives:

- To test the effects of various HIT properties (task parameters) on the quality of the resulting labels: pay, effort, and worker qualification.
- To gather relevance labels for the book pages in our gold set based on which we can measure the quality of a crowdsourcing experiment in terms of the agreement with the gold set (accuracy). For each book page, the following options were offered in the HIT: 'Relevant', 'Not relevant', 'Broken link', 'Don't know'. A free-text comment field was also provided. In addition, to reduce the attractiveness of the HITs to random clicking, we used a challenge-response test or 'captcha' asking workers to enter the last word printed on the scanned book page.
- To collect data about the workers' perceptions of the task, e.g., if they thought it too difficult or if it paid too little, etc.

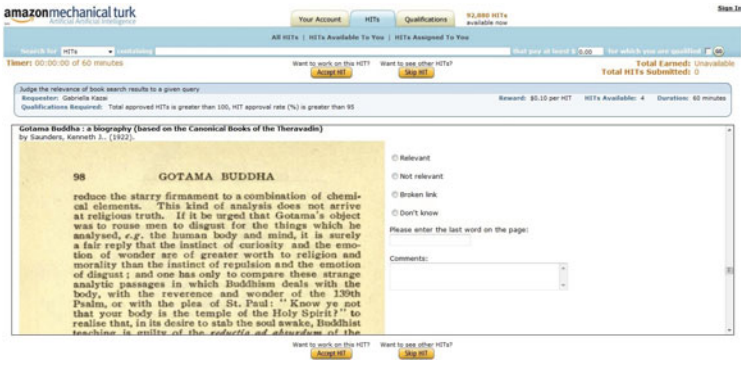


Fig. 2. Example HIT from the experiments on AMT

Figure 2 shows an example HIT. In order to display the book page images inside a HIT, we made use a web service call of the Book Search System[4] provided by the INEX Book Track.

3.1 Task Parameters

We experimented with the following task parameters:

- Pay: We experimented with two pay levels, paying $0.10 or $0.25 per HIT.
- Worker qualifications: This parameter was controlled through the use of AMT's worker selection criteria. We used two settings: 1) no required qualifications ('noQ'), and 2) restricting to workers with over 95% HIT approval rate with over 100 approved HITs ('yesQ').
- Effort: Effort was controlled via the number of book pages that workers were required to judge in a given HIT: 1) HITs containing 5 book pages and 2) HITs with 10 pages.

It is clear that some of these parameters are directly or indirectly related to each other. For example, one may expect higher pay for tasks that require more effort. We aim to uncover such relationships and explore their influences on each other and on the quality of the crowdsourced labels.

With this aim, we created 8 batches of HITs based on the different combinations of task parameter values, see Table 1. The naming of a batch combines the three task parameter values: pay ('10' or '25' cents), qualifications ('noQ' or 'yesQ'), and effort ('5' or '10' book pages per HIT). Each batch contained 800 book pages, 100 per topic. This resulted in 160 HITs when 5 pages were included per HIT, and 80 HITs when effort was 10 pages per HIT. We requested 3 workers per HIT, yielding either 240 or 480 assignments per batch and giving us 2,400 relevance labels per batch. The total number of collected labels is thus 19,200. Note that these numbers exclude rejected work. The totals including rejected HITs is shown in brackets in Table 1.

[4] `http://www.booksearch.org.uk/`

Table 1. Batches of HITs with different task parameter settings

Batch	Pay	Qualif.	Effort	HITs	Assignments	Judged pages	Cost
10-noQ-5	$0.10	no	5 pages	160	480 (608)	2,400 (3,040)	$52.80
10-noQ-10	$0.10	no	10 pages	80	240 (460)	2,400 (4,600)	$26.40
10-yesQ-5	$0.10	yes	5 pages	160	480 (722)	2,400 (3,610)	$52.80
10-yesQ-10	$0.10	yes	10 pages	80	240 (358)	2,400 (3,580)	$26.40
25-noQ-5	$0.25	no	5 pages	160	480 (592)	2,400 (2,960)	$132.00
25-noQ-10	$0.25	no	10 pages	80	240 (299)	2,400 (2,990)	$66.00
25-yesQ-5	$0.25	yes	5 pages	160	480 (480)	2,400 (2,400)	$132.00
25-yesQ-10	$0.25	yes	10 pages	80	240 (304)	2,400 (3,040)	$66.00

3.2 Workers' Perception of the Task

With the aim to examine the use of self-reported measures as possible indicators of quality, we also gathered various feedback from the workers. For example, we may expect that more knowledgeable workers would produce higher quality labels [3]. Workers could provide feedback on the following aspects:

- Familiarity: We asked workers to rate their familiarity with the subject of the topic for which relevance labels were sought in a 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 collected 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 if the task was in their opinion 'boring', 'ok', or 'interesting'.
- Pay: We solicited workers' opinions if they thought they were being paid 'too little', 'ok', or 'too much', or if pay did not matter to them.

4 Results and Findings

In this section, we present our findings with the aim of answering our original research questions. Since all book pages included in the HITs have known labels, we use accuracy as a measure of crowdsourced label quality, calculated as the ratio of the total number of correct labels and the total number of labels in a given batch or subset. A relevance label submitted by a worker is correct if it matches the label (relevant/irrelevant) in our gold set for the given book page. We note that 45 of the 800 book pages are actually 'missing' from the corpus in the sense that their page image, fetched from the Book Search System, fails to load in the HIT. In the case of these known missing pages, we accept 'Broken link' as the correct answer.

Table 2 (rows 1-8) shows the calculated accuracy levels for the 8 batches over three filter sets: 1) all the collected data, including rejected work and unusable labels: when workers clicked 'Don't know' or 'Broken link' or when no label was given (all data), 2) the subset that excludes rejected work (no spam), and 3) the fully cleaned subset that excludes rejected work and pages with unusable

Table 2. Number of unique workers, average time spent per book page, and accuracy of the crowdsourced relevance labels across the different batches or subsets, corresponding to different task parameter combinations

	Batch/Subset	All Gathered Labels			No Spam			Cleaned		
		#Wkrs	Time	Acc.	#Wkrs	Time	Acc.	#Wkrs	Time	Acc.
1.	10-noQ-5	70	42	59.74%	66	51	61.79%	65	48	66.38%
2.	10-noQ-10	69	26	34.98%	63	42	59.29%	61	40	67.26%
3.	10-yesQ-5	66	42	60.78%	62	58	62.62%	62	58	66.97%
4.	10-yesQ-10	35	42	59.22%	33	59	59.75%	33	58	62.90%
5.	25-noQ-5	71	51	52.03%	68	61	54.79%	66	59	57.48%
6.	25-noQ-10	58	41	52.34%	54	49	63.50%	54	48	72.56%
7.	25-yesQ-5	43	61	71.50%	43	61	71.50%	43	61	73.58%
8.	25-yesQ-10	54	33	67.04%	48	38	69.83%	48	38	74.01%
9.	$0.10	199	37	52.18%	186	52	60.86%	183	51	65.85%
10.	$0.25	197	46	60.22%	184	52	64.91%	182	51	69.36%
11.	$0.01	99	33	45.59%	92	50	59.52%	90	49	65.01%
12.	$0.02	130	42	60.30%	122	54	62.21%	121	53	66.67%
13.	$0.025	105	37	59.75%	95	43	66.67%	95	42	73.31%
14.	$0.05	108	56	60.75%	105	61	63.15%	103	60	65.63%
15.	$0.10 noQ	121	32	44.83%	113	46	60.54%	110	44	66.81%
16.	$0.10 yesQ	90	42	60.00%	84	58	61.19%	84	58	64.93%
17.	$0.25 noQ	121	46	52.18%	114	55	59.15%	112	53	64.70%
18.	$0.25 yesQ	92	45	69.01%	86	49	70.67%	86	49	73.79%
19.	noQ	225	38	48.05%	211	51	59.84%	207	49	65.75%
20.	yesQ	155	43	63.88%	148	54	65.93%	148	54	69.40%
21.	5 Pages	225	48	60.50%	215	58	62.68%	212	56	66.14%
22.	10 Pages	189	34	51.60%	175	47	63.09%	173	46	69.15%

labels (cleaned). HITs were rejected if the worker completing the HIT filled in, on average over all their HITs, less than 30% of the captcha fields and spent less than 20 seconds judging a page.

As it can be seen, we obtain accuracy levels mostly in the region of 52-74%, with the exception of `10-noQ-10` resulting in only 35% agreement with the gold set labels (over all collected data). This is perhaps not surprising as this batch presents the 'hardest deal' where workers needed to judge 10 pages per HIT for only $0.10 payment, while this batch was also open to all workers (noQ). The best accuracy is 74% obtained for the `25-yesQ-10` batch (cleaned data), similar to inter-assessor agreement levels at TREC. It is worthwhile to note how the perception of quality changes depending on how the collected data is filtered. For example, batch `25-noQ-10` ranks 6th when quality is calculated over all collected data, but it ranks 3rd when rejected work is removed. In the next sections, we will report findings based on all three filter sets and attempt to explain the differences between them.

4.1 Does Quality Pay?

Our first research question regards whether pay affects the quality of the collected relevance labels. To answer this question, we grouped the collected labels into two

bins based on the amount of pay per HIT, see rows 9-10 in Table 2. We see that increasing pay from \$0.10 to \$0.25 per HIT leads to improved label quality in all our data sets (all data, no spam, cleaned). A two sample t-test confirms that pay leads to a significant difference in accuracy per HIT in the unfiltered and no spam sets ($p < 0.01$, two-tailed). Unlike previous reports, e.g., [10], this finding suggests that pay does impact quality: encouraging better work. At the same time, we observe a reduced benefit to the increase in pay as we filter out spam and unusable labels: the increase in accuracy gained by paying more drops from 115% (all data) to 107% (no spam) and then to 105% (cleaned). So, what does this mean?

Looking at the differences between the three filter sets, we see that more unusable and spam labels are contributed when pay is lower: in total 4,293 incorrect labels (29% of all collected labels) are removed during our filtering from the \$0.10 set, while this is only 17% (1,990 incorrect labels) in the \$0.25 set. This is also reflected in the relative increase in accuracy between the unfiltered and the cleaned data sets: 126% improvement for the \$0.10 HITs while only 115% for the \$0.25 HITs. This advocates that pay also impacts quality indirectly, where higher pay leads to more HITs with usable labels and less spam. This is also supported by the average time that workers spent on judging a page: 37 seconds in the lower pay condition vs. 46 seconds in the higher pay batches (all data). Interestingly, after excluding rejected work, the average time is 52 seconds per page for both pay levels. However, as we will see later, time spent per page is influenced by a range of aspects, e.g., expertise, captcha, which complicate its use for indicating label quality.

Comparing quality between different pay per label sets (instead of pay per HIT), see rows 11-14 in Table 2, confirms the same trend: quality increases with pay. However, we can also observe evidence of a diminishing return effect, whereby the rate of increase in pay is matched by a slowing (or dropping) rate of increase in quality. This is especially clear on the cleaned data set, where accuracy first increases from 65% (\$0.01 per page) to 67% (\$0.02 per page) and then to 73% (\$0.25 per page) but then it drops back down to 66% (\$0.05 per page). The difference in accuracy across the three filter sets indicates that the lowest paid condition is most inducing of spamming. Interestingly, the best accuracy is achieved by the subset that has the second highest level of spam and unusable labels (\$0.025 per page). As very little (detected) spam or unusable labels are contributed in the highest paid condition, we may reason that the drop in accuracy is indeed due to the phenomena of diminishing return.

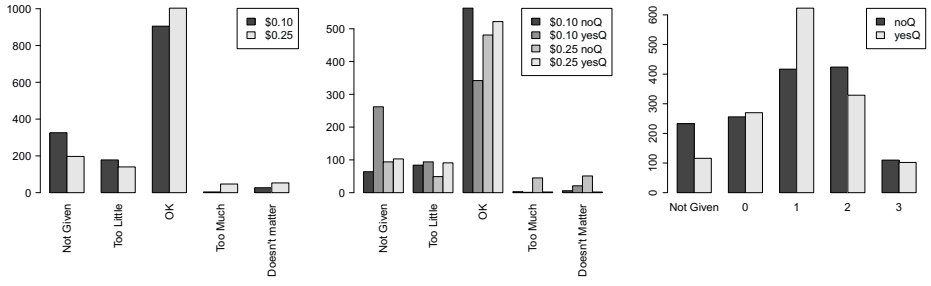
To investigate whether pay affects different groups of workers differently, we now focus on the subsets of labels contributed in the 'noQ' and 'yesQ' conditions within the two pay levels, see rows 15-18 in Table 2. We find differences in label quality for both qualified and non-qualified workers: accuracy of 60% (yesQ) and 45% (noQ) when paid \$0.10 per HIT vs. 69% (yesQ) and 52% (noQ) when receiving \$0.25 per HIT (all data). This suggests that pay effects both groups equally: higher pay encourages better work regardless of the AMT qualifications. However, after removing spam and unusable labels, we find that the level of pay shows no effect on the accuracy of non-qualified workers (around 66%). After

a more detailed look at the data, we found that this was in fact a result of labels contributed by possibly unethical workers who escaped our spam filter or by workers who mistook the task for OCR correction (instead of relevance assessment). We found that in the '$0.25-noQ' batches a higher number of HITs (115) were submitted by workers who filled in over 30% of captcha fields, but labeled most pages (> 80%) as relevant (a less likely label given our 40/60 ratio of relevant/irrelevant pages), compared with 5 HITs in the '$0.10-noQ' subset and 27 and 40 HITs in the '$0.10-yesQ' and '$0.25-yesQ' subsets, respectively. Of course, with an adapted spam filter, such HITs can again be removed from the cleaned data. Overall, thus, we conclude that higher pay induces two different behavior in workers: on the one hand it encourages better work, especially for qualified workers, but at the same time it also attracts more unethical workers, especially when workers are not pre-filtered.

Next, we analyze workers' feedback about the amount of pay offered per HIT, see Figure 3a (no spam set). Unsurprisingly, we find that workers were more satisfied with higher pay, but we also see that the majority of the responses indicated that workers were content ('pay is ok') with the offered pay in both pay categories: 63% in the $0.10 batches and 70% in the $0.25 batches. This confirms that we estimated the minimum level of pay correctly at $0.01 per page, but it also shows that workers accept a relatively wide spectrum of pay for the same work: from $0.05 down to $0.01 per judgment. Only 2% of the responses indicated that workers did not care about pay in the low paid batches, while this was 4% in the higher paid batches. Among the subset of HITs with 'too much pay' or 'pay does not matter' responses, workers indicated high (66% of responses) or moderate (31%) interest in the task. This ratio is very different for workers who selected 'pay is ok' as their feedback: only 20% of the responses expressed high interest and 71% indicated moderate interest. Oddly, when workers were unsatisfied with their pay ('pay is too little'), 30% of the responses registered high interest, and 60% moderate interest. This highlights the duality of pay and interest, where pay becomes secondary when interest is high, e.g., as demonstrated by the ESP game [12]. A chi-square test also confirmed that pay and interest are significantly related ($p < 0.01$).

When broken down by qualifications, we found that 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 3b. 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.

Looking at accuracy for the different feedback categories on pay, we find no difference between accuracy for the 'pay is ok' and 'pay is too little' sets: 70% for both. Accuracy drops slightly for 'pay does not matter' responses (63%) or when no feedback was given (60%). The lowest accuracy is obtained in the set where workers indicated that pay was 'too much' (40%). This highlights the possible use of these fields as weak signals to filter workers.



a) b) c)

Fig. 3. 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') in noQ and yesQ batches (c).

4.2 Does Worker Qualification Affect Quality?

To answer this question we looked at the quality of labels in two subsets of HITs: those that restricted participation by requiring workers to have over 95% acceptance rate and more than 100 HITs (yesQ), and those where no qualifications were required (noQ), see rows 19-20 in Table 2. We see that pre-selecting workers leads to better accuracy in all our filter sets, e.g., 48% for noQ vs. 64% for yesQ (all data). Moreover, a two sample t-test shows that qualification leads to a significant difference in accuracy per HIT for each of the three filter sets ($p < 0.01$, two-tailed). The differences between the reported accuracy levels between the unfiltered and filtered sets clearly show that non-qualified workers contribute more spam and unusable labels. However, in the cleaned data set we only observe a small benefit to pre-selecting workers: 66% for noQ vs. 69% for yesQ. This is of course a result of our filters which lead to 41% of the collected labels being thrown away from the noQ set (incl. 78% incorrect labels), compared with 33% in the yesQ set (incl. 47% incorrect labels). The difference between the percentages of incorrect labels in the discarded data suggests that different filtering methods may be more appropriate for different types of workers.

Looking at the reported levels of familiarity with the topic by qualified and non-qualified workers, we see that non-qualified workers tend to be more confident and report higher familiarity than qualified workers (not statistically significant), see Figure 3c.

When calculating accuracy for each familiarity rating, we find that the most self confident workers in fact have the lowest accuracy (45% on all data, 58% on no spam), while accuracy for lower levels of familiarity ranges between 52-61% (all data) or 63-67% (no spam). When broken down by worker qualifications, we see that it is those non-qualified workers who rated their knowledge on the topic highly who do worst (accuracy of 33% on all data and 55% on no spam), while qualified 'expert' workers achieve 60-61% accuracy (all data and no spam). In general, we see a negative correlation between self-reported expertise and accuracy for non-qualified workers (accuracy drops as familiarity increases: 56% (familiarity level 0), 55% (familiarity 1), 42% (familiarity 2) and 33% (familiarity 3) on all data),

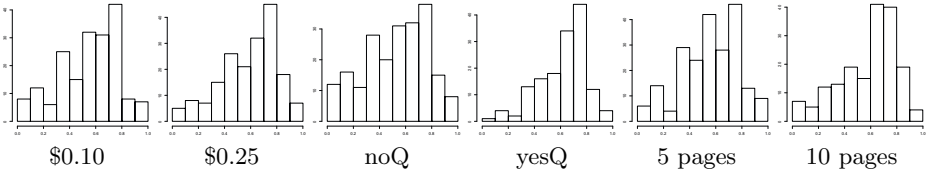


Fig. 4. Distribution of workers over accuracy (no spam set)

while for qualified workers accuracy varies less across familiarity levels (63-67% on all data and 65-70 on no spam (familiarity levels 0-2)). This indicates that more veteran workers on AMT are better at gaging their own level of expertise while new workers may be more prone to effects of satisficing [6]. It is interesting to note that best accuracy is obtained when workers reported only minimal knowledge of the topic (familiarity 0) for both qualified and non-qualified workers (on all our filter sets). This is somewhat counterintuitive as we would expect that more knowledge of a topic would lead to better accuracy [3]. A possible explanation is that these less confident workers may take more care in completing the task, while over confident workers may be prone to mistakes. This is especially true for non-qualified workers who may care less about the quality of their work as they do not yet have a track record to protect. This is also partly supported by the observation that more knowledgeable workers took on average less time to judge a page (no spam set): 41 seconds (familiarity of 2) vs. 53 seconds (familiarity of 0). However, it should be noted that the time alone cannot give a true picture of workers' commitment to the task, since workers claiming to have more extensive knowledge have in fact only filled in on average 55% of the captcha fields, compared with 72-88% in HITs where lower familiarity levels were reported.

4.3 Does Effort Affect Quality?

To answer this question, we grouped results from batches that asked workers to label 5 or 10 book pages, resp., see rows 21-22 in Table 2. Looking at the unfiltered data set, we may conclude that effort does matter and better results can be produced in conditions when workers are not overloaded (61% accuracy for HITs with 5 pages vs. 52% for HITs with 10 pages). Indeed, we find that effort leads to a significant difference in accuracy per HIT in the unfiltered set ($p < 0.01$, two-tailed, two sample t-test). However, accuracy in the cleaned set suggests the opposite (5 pages: 66% vs. 10 pages: 69%). That said, workers do seem to be less motivated to do well on tasks that require more effort and as a result the collected data contains more unusable and spam labels. This is corroborated by the finding that workers spent on average longer judging a page in the low effort HITs: 48 seconds vs. 34 seconds (all data).

The self reported difficulty revealed that non-qualified workers who found the task difficult performed the worst (35% all data, 50% no spam), compared with 48-50% (all data) and 60-61% (no spam) accuracy when the task was reported as not difficult. Qualified workers achieved a more consistent accuracy regardless whether they found the task easy or difficult (68-72%, no spam).

5 Conclusions

In this paper, in the context of crowdsourcing relevance labels, we investigated three basic parameters of crowdsourcing experiment design (pay, effort and worker qualifications) and their influence on the quality of the output, measured by accuracy. Unlike in [10], our findings show that pay does matter: higher pay encourages better work while lower pay results in increased levels of unusable and spam labels. Looking at pay per label (rather than per HIT), we found evidence of diminishing return where the rate of increase in quality flattens out or even drops as pay increases. This was partly due to the higher pay attracting more sophisticated unethical workers who escaped our simple filter. From this, it is clear that experiment designers need to find the right balance between too low pay that results in sloppy work and too high pay that attracts unethical workers. Estimating reward per unit of effort promises a suitable method for this. In addition, higher pay should also be balanced with better quality control mechanisms built into the design (e.g., pre-filtering workers, captcha, or training [8]) and more robust spam filters.

When comparing different groups of workers, we found that more qualified workers produce better quality work. This may be a consequence of the 'reputation' system in AMT, where these workers strive to maintain their qualification levels, e.g., HIT approval rate, as this allows them to have access to a wider range of tasks. However, quality cannot be guaranteed just by pre-filtering workers as more sophisticated unethical workers can easily build a false reputation. At the same time, there are plenty of ethical non-qualified workers who aim to build up their reputation and may thus be more diligent when completing a task. To take advantage of both groups requires different HIT designs to engage the right workers as well as adapted spam filters to process the output. Both pay and qualifications lead to significant differences in output quality.

With respect to effort, we found that while higher effort induces more spam, it also leads to slightly better quality after spam removal than low effort HITs, though this is not statistically significant.

Figure 4 shows the distribution of workers over agreement, for the various subsets of the gathered labels (no spam set), confirming the above observations.

In summary, our findings highlight a network of influences between the different task parameters and the output quality. We found that all investigated task parameters had influence on the output quality, both directly and indirectly, e.g., higher pay encouraging better work while also attracting more sophisticated spammers. We conclude that increasing pay, reducing effort, and introducing qualification requirements can all help in reducing undesirable behavior. However, due to the interplay of the parameters and their influence on each other, on the HIT design, and the output, each such decision needs to be balanced overall, e.g., increased pay may call for additional quality control elements. Our analysis of the collected data also revealed the need for a deeper understanding of the observed variables, e.g., time spent judging a page, to aid in the detection of sloppy or unethical workers. In addition, we found that self-reported data, e.g.,

familiarity, perceived difficulty of the task, and satisfaction with pay, can also help experimenters in filtering out sub-quality data.

Our future work will explore these relationships in more detail. We also plan to expand the set of task parameters to include factors such as clarity, emotion, aesthetics, pre-task qualification tests, seeding, etc. Our ultimate goal is to provide experimenters a framework to guide the design of their crowdsourcing tasks to maximize quality.

Acknowledgments

Many thanks to Jaap Kamps for helpful comments on the camera ready version of this paper.

References

1. 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)
2. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. SIGIR Forum 42(2), 9–15 (2008)
3. 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 2008: Proceedings of the 31st Annual International ACM SIGIR Conference, pp. 667–674 (2008)
4. Grady, C., Lease, M.: Crowdsoucring 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)
5. Howe, J.: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Publishing Group, New York (2008)
6. 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 (CrowdConf 2010) (2010)
7. Kazai, G., Koolen, M., Doucet, A., Landoni, M.: Overview of the INEX 2009 book track. In: Geva, S., Kamps, J., Trotman, A. (eds.) INEX 2009. LNCS, vol. 6203, pp. 145–159. Springer, Heidelberg (2010)
8. Le, J., Edmonds, A., Hester, V., Biewald, L.: Ensuring quality in crowdsourced search relevance evaluation. In: SIGIR Workshop on Crowdsourcing for Search Evaluation, pp. 17–20 (2010)
9. Marsden, P.: Crowdsourcing. Contagious Magazine 18, 24–28 (2009)
10. Mason, W., Watts, D.J.: Financial incentives and the "performance of crowds". In: HCOMP 2009: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 77–85 (2009)
11. Quinn, A.J., Bederson, B.B.: A taxonomy of distributed human computation. Technical Report HCIL-2009-23, University of Maryland (2009)
12. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004, pp. 319–326 (2004)
13. Zhu, D., Carterette, B.: An analysis of assessor behavior in crowdsourced preference judgments. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (2010)