

# The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior

Mark D. Smucker  
Department of Management Sciences  
University of Waterloo  
mark.smucker@uwaterloo.ca

Chandra Prakash Jethani  
David R. Cheriton School of Computer Science  
University of Waterloo  
cpjethan@cs.uwaterloo.ca

## ABSTRACT

There are considerable differences in remuneration and environment between crowd-sourced workers and the traditional laboratory study participant. If crowd-sourced participants are to be used for information retrieval user studies, we need to know if and to what extent their behavior on information retrieval tasks differs from the accepted standard of laboratory participants. With both crowd-sourced and laboratory participants, we conducted an experiment to measure their relevance judging behavior. We found that while only 30% of the crowd-sourced workers qualified for inclusion in the final group of participants, 100% of the laboratory participants qualified. Both groups have similar true positive rates, but the crowd-sourced participants had a significantly higher false positive rate and judged documents nearly twice as fast as the laboratory participants.

## 1. INTRODUCTION

Much of the existing information retrieval (IR) research on crowd-sourcing focuses on the use of crowd-sourced workers to provide relevance judgments [5, 2] and several researchers have developed methods for extracting better quality judgments from multiple workers than is possible from a single worker [1, 4]. Common to this work is the need to deal with workers who are attempting to earn money for work without actually doing the work — random crowd-sourced workers are not to be trusted.

In contrast, many IR user studies are traditionally designed with trust in the participant. We as researchers ask the study participants to “do your best” at the given task. We ask this of the participants because we often rely on the participants to identify for us documents that they find to be relevant for a given search task. When the participant determines what is relevant, and especially when the participant originates the search topic, we are left trusting the participant’s behavior and judgment.

Trust in the participant is only one of the many differences between crowd-sourcing and traditional laboratory environments. Remuneration also differs considerably. For example, McCreadie et al. paid crowd-sourced workers an effective hourly wage of between \$3.28 and \$6.06 [7] to judge documents for the TREC 2010 Blog Track, and Horton and Chilton have estimated the hourly reservation wage of crowd-sourced workers to be \$1.38 [3]. Many laboratory partic-

ipants are paid at rates around \$10 per hour. For timed tasks, in the lab we can eliminate distractors such as phones and instant messages. In contrast, a crowd-sourced worker may multi-task between doing the task and answering email. Laboratory studies allow the researcher to control many variables. Without this control, much larger samples are required to observe differences between experimental groups.

In this paper, we begin looking at the question of how crowd-sourced study participants behave compared to traditional, university recruited, laboratory study participants for IR tasks. In particular, we concern ourselves with the non-trivial, but relatively simple task of judging the relevance of documents to given search topics. Our goal here is the study of behavior rather than developing a new process to obtain a set of good relevance judgments from noisy workers. If crowd-sourced participants behave in ways that are different than laboratory participants for the task of judging document relevance, then we should expect other IR user studies to likewise be different given that judging document relevance is an inherent component of many IR studies.

Many user studies in IR involve some sort of search task. A researcher has many choices of how to measure the performance of participants on a search task. One possible way is to ask the participant to work for a fixed amount of time. The advantage to this is that the participant has no incentive to rush the task and do a poor job. The hope is that the participant works at their usual pace and usual quality. The disadvantage to a fixed time task is that the participant may not be motivated to perform at their maximum potential. Another possible way is to give the participant a task of fixed size such as finding 5 relevant documents. An advantage of this design is that the participant may work harder to finish sooner knowing the work is fixed. A disadvantage is that the participant may submit non-relevant documents as relevant simply to finish the task quickly. Most crowd-sourced tasks are of a fixed size. The faster a crowd-sourced worker works, the more the worker earns per hour.

To mimic a crowd-sourced environment, we designed a laboratory study that first had participants qualify for participation in a larger fixed-size task. The use of a qualification task is a feature of Amazon’s Mechanical Turk crowd-sourcing platform. Requesters of work on Mechanical Turk can create tasks (HITs) that only workers who have passed a qualification task are allowed to accept.

As a SIGIR 2011 Crowdsourcing Workshop Challenge grantee, we did our work with CrowdFlower. As we worked with the CrowdFlower platform, it became clear that it would not be easy to do a qualification task. Instead, we choose to

| Number | Topic Title                  | Relevant |
|--------|------------------------------|----------|
| 310    | Radio Waves and Brain Cancer | 65       |
| 336    | Black Bear Attacks           | 42       |
| 362    | Human Smuggling              | 175      |
| 367    | Piracy                       | 95       |
| 383    | Mental Illness Drugs         | 137      |
| 426    | Law Enforcement, Dogs        | 177      |
| 427    | UV Damage, Eyes              | 58       |
| 436    | Railway Accidents            | 356      |

**Table 1: Topics used in the study and the number of NIST relevant documents for each topic.**

utilize CrowdFlower’s quality control system of *gold questions*. A gold question is a question to which the answer is already known. If a worker’s accuracy as measured by the gold questions drops below 70%, that worker cannot accept any further tasks.

We contend that the performance we obtained from our laboratory participants should be considered a “gold standard” for the typical university controlled laboratory study that involves students. The students are assumed to be of good character, are being paid at a reasonable level, and are working under supervision without distractions. Crowd-sourced workers are well known to contain many who are scammers trying to get paid without work, are being paid a low wage, and are working in their own uncontrolled environments.

We measured both the crowd-sourced and laboratory participants on the judgments they made as well as the time it took them to make these judgments. Next we describe our experiments in more detail, and then we present and discuss the results.

## 2. MATERIALS AND METHODS

We conducted two experiments. The first was a laboratory-based study at a university with 18 participants. The second, was via CrowdFlower and ran on Amazon Mechanical Turk and had 202 crowd-sourced participants. Both studies received ethics approval from our university’s Office of Research Ethics.

We utilized 8 topics from the 2005 TREC Robust track, which used the AQUAINT collection of newswire documents. Table 1 shows the 8 topics. Topics 383 and 436 were used for training and qualification purposes while the remaining 6 topics were used for testing the performance of the participants.

### 2.1 Laboratory Experiment

In this experiment, each participant judged the relevance of documents for two search topics. Figure 1 shows the user interface for judging documents. The study utilized a tutorial and a qualification test before allowing participants to continue with the study and judge documents for the two search topics.

We provided instructions on how to judge the relevance of documents at the start of the tutorial. In previous experiments, we have seen some evidence that a few participants will not carefully read instructions. To try and prevent this skimming of instructions, we placed a simple quiz about the instructions at their end. Participants could not proceed

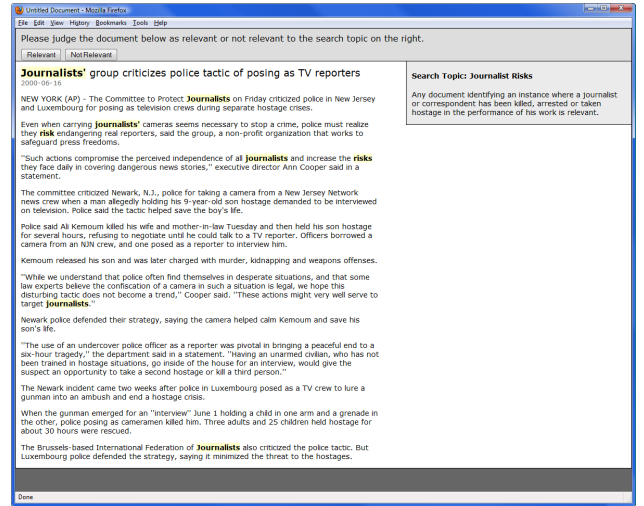


**Figure 1: This screenshot shows the user interface (UI) for judging a document used in both experiments.**

with the study until they answered all quiz questions correctly.

The tutorial involved practice judging the relevance of 10 documents and the qualification test required participants to achieve 70% accuracy on the relevance judgments for 20 documents. For both the 10 and 20 document sets, the participants judged a 50/50 mix of relevant and non-relevant documents. Both the tutorial and qualification task used topics 383 and 436. We paid participants \$7 for completing the tutorial and qualification task.

All participants passed the qualification test. The actual task consisted of making relevance judgments for documents from two of six topics. For each of two topics, a participant judged 40 documents selected randomly from the documents in the set of TREC relevance judgments such that each 40 documents was composed of 20 relevant and 20 non-relevant documents. The six topics were rotated across blocks of six participants such that each topic was judged by two of the six participants and each topic was once a first task and once a second task topic. We paid participants \$18 for completing this judging task, for a total of \$25. Excluding the tutorial and qualification task, each participant judged 80 documents at a cost of 31.3 cents per document. Including tutorial and qualification judgments, we paid 22.7 cents per judgment.

Many participants completed the study within an hour and all completed it within 2 hours. Our participants were mainly graduate students. We conducted the study in a quiet laboratory setting and supervised all work.

### 2.2 Crowd-Sourced Experiment

We utilized CrowdFlower to run the crowd-sourced experiment. CrowdFlower provides a convenient platform to allow users to run crowd-source jobs across a range of crowd-source worker pools. We ran all of our jobs on Amazon Mechanical Turk. One job briefly ran on Gambit by accident when CrowdFlower’s support attempted to help the job complete faster.

We created one job per topic for a total of 6 jobs. CrowdFlower workers can accept *assignments*, which on Mechan-

ical Turk is the equivalent of a HIT. Each assignment provided a set of instructions that included ethics and consent information. The instructions for how to judge relevance matched those of the laboratory study. While the laboratory study required the participants to take a quiz about the instructions, here the quiz was provided with its answers. In addition, while the laboratory tutorial required participants to view and judge documents for practice, we provided the same opportunity to the crowd-sourced participants but judging the practice documents was optional.

Each assignment consisted of 5 *units*. A unit is a set of questions tied to a row of data that one uploads to CrowdFlower when creating a job. For our jobs, each unit corresponded to a document to judge. We provided a link to an external website that first placed the participant on a page that asked them to click another link when they were ready to judge it. We did this to be sure that we could set cookies to track the participant as well as try and more accurately measure the time it took to judge an individual document. We were concerned that participants would open all 5 links in an assignment and then begin working on them. Unfortunately, it appears that some participants did this and also went ahead and clicked the link to judge a document for all 5 links before beginning to judge any of the 5 documents. To correct for the cases where the participant loaded multiple documents at once, we estimated the user’s time to judge a document as the interval from the time the participant submitted the judgment back to the previous recorded event.

On submission of a relevance judgment, we provided the participant with a “codeword” that was to then be entered into the unit’s multiple choice question. There were five codewords: Antelope, Butterfly, Cheetah, Dolphin, and Elephant. For each document, we randomly assigned a codeword for the correct and incorrect judgment. We wanted to collect judgment information with our website so as to be able to measure the time it took the participant to make the judgment. We also wanted to utilize CrowdFlower’s system of *gold* to end the participation of participants whose performance was below 70% accuracy, and thus the participants also needed to enter their judgment into the CrowdFlower form. By using our codeword system, we could identify participants who were not viewing the documents for they only had a 40% chance of selecting a plausible answer. In addition, participants not viewing and judging the document, only had a 20% chance of guessing the correct answer compared to binary relevance’s usual 50%.

We used a mix of 50% relevant and 50% non-relevant documents for each topic. We selected all documents marked relevant by the NIST assessors and then randomly selected an equal number of non-relevant documents. For each topic, we selected approximately 10% of the documents as gold based on the recommended amount in CrowdFlower’s documentation. A gold document is one on which the participant is judged. If the participant’s accuracy on gold drops below 70%, the participant may not accept further assignments from a job. We selected documents for which we had already verified their relevance by a consensus process [8], and then randomly selected the remaining documents. All gold was 50% relevant and 50% non-relevant. For topic 310, we added more gold when the job got stuck for having too many participants be rejected by the gold. In the end, topic

310 had 35 gold (18 non-relevant, 17 relevant, and 27% of units).

CrowdFlower shows one gold per assignment, and thus one out of five documents in an assignment were gold documents. Only after completion of our jobs did we discover that CrowdFlower recycles the gold if a worker has judged all the gold. Our website told the participant whenever a document had already been judged and provided the codeword to use. Thus, after judging 50% of a topic’s possible documents, the participants were effectively qualified for the remaining documents and could have taken the opportunity to lower their judging accuracy or even cheat.

We collected judgments via both CrowdFlower’s system and our own website. We had difficulty matching our identification of the participant to CrowdFlower’s worker IDs. As a result of this difficulty, we use only the judgments that we collected via our website.

While CrowdFlower ceased the participation of participants with gold accuracies that dropped below 70%, after examining our data, it was clear that this was not a sufficient filter nor nearly equivalent to our laboratory study. All of our laboratory participants had to display 70% accuracy on 20 documents made up of 10 relevant and 10 non-relevant documents. In addition, for all laboratory participants, we measured their performance on a topic with 40 document judgments. To make the qualification of both groups more similar, we only retained crowd-sourced participants who obtained 70% on the “first 20” documents judged and who judged at least 40 documents for a topic. The “first 20” documents consisted of the first 10 relevant and first 10 non-relevant documents judged by the participant. Because CrowdFlower appears to deliver documents randomly to users, it is possible for a user to obtain a mix of 20 documents that does not have a precision of 0.50. If accuracy is to be used to qualify participants, it is important that the mix of documents be equally divided between relevant and non-relevant documents. For example, we saw a participant who judged all documents to be relevant, and this participant received a mix of 20 documents with a precision of 0.70. In addition to the filtering we applied to participants, CrowdFlower excludes workers they have found to be spammers or to provide low quality work.

For each job, we specified that each document was to be judged by a minimum of 10 qualified participants. We paid participants \$0.07 (1.4 cents per document) for each completed assignment. In total, we paid CrowdFlower \$313.14 for 10374 judgments from the participants who met our criteria, or 3.02 cents per judgment. CrowdFlower collected 22445 judgments from all participants.

## 2.3 Measuring Judging Behavior

We view the task of relevance judging to be one of making a classic signal detection yes/no decision. Established practice in signal detection research is to measure the performance of participants in terms of their true positive rate (hit rate) and their false positive rate (false-alarm rate). Accuracy is rarely a suitable measure unless the positive (relevant) documents and negative (non-relevant) documents are balanced, which they are in this study. The true positive rate is measured as:

$$TPR = \frac{|TP|}{|TP| + |FN|} \quad (1)$$

| Participant  | NIST Judgment            |                          |
|--------------|--------------------------|--------------------------|
|              | Relevant (Pos.)          | Non-Relevant (Neg.)      |
| Relevant     | $TP = \text{True Pos.}$  | $FP = \text{False Pos.}$ |
| Non-Relevant | $FN = \text{False Neg.}$ | $TN = \text{True Neg.}$  |

**Table 2: Confusion Matrix.** “Pos.” and “Neg.” stand for “Positive” and “Negative” respectively.

and the false positive rate as:

$$FPR = \frac{|FP|}{|FP| + |TN|} \quad (2)$$

and accuracy is:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \quad (3)$$

where  $TP$ ,  $FP$ ,  $TN$ , and  $FN$  are from Table 2.

In both experiments, we judge the participants against the judgments provided by NIST. While we know the NIST assessors make mistakes [8], here we are comparing two groups to a single standard and mistakes in the standard should on average equally affect the scores of both groups.

Signal detection theory says that an assessor’s relevance judging task may be modeled as two normal distributions and a criterion [6]. One distribution models the stimulus in the assessor’s mind for non-relevant documents and the other for relevant documents. The better the assessor can discriminate between non-relevant and relevant documents, the farther apart the two distributions are. The assessor selects a criterion and when the stimulus is above the criterion, the assessor judges a document relevant, otherwise non-relevant.

Given this model of the signal detection task, with a TPR and FPR, we can characterize the assessor’s ability to discriminate as:

$$d' = z(TPR) - z(FPR) \quad (4)$$

where the function  $z$  is the inverse of the normal distribution function and converts the TPR or FPR to a  $z$  score [6]. The  $d'$  measure is very useful because with it we can measure the assessor’s ability to discriminate independent of the assessor’s criterion. For example, assume we have two users A and B. User A has a TPR = 0.73 and a FPR of 0.35, and user B has a TPR = 0.89 and a FPR = 0.59. Both users have a  $d' = 1$ , in other words, both users have the same ability to discriminate between relevant and non-relevant documents. User A has a more conservative criterion than User B, but if the users were to use the same criterion, we’d find that they have the same TPR and FPR. Figure 2 shows curves of equal  $d'$  values.

We can also compute the assessor’s criterion  $c$  from the TPR and FPR:

$$c = -\frac{1}{2}(z(TPR) + z(FPR)) \quad (5)$$

A negative criterion represents a liberal judging behavior where the assessor is willing to make false positive mistakes to avoid missing relevant documents. A positive criterion represents a conservative judging behavior where the assessor misses relevant documents in an attempt to keep the false positive rate low.

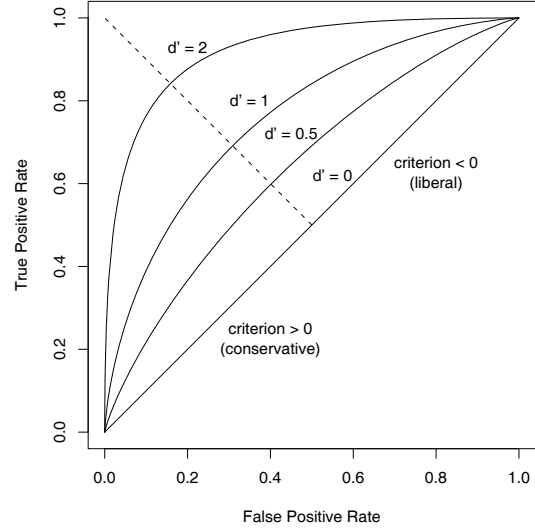


**Figure 2: Example  $d'$  curves.**

For both the computation of  $d'$  and  $c$ , a false positive or true positive rate of 0 or 1 will result in infinities. Rates of 0 and 1 are most often caused by these rates being estimated based on small samples. To better estimate the rates and avoid infinities, we employ a standard correction of adding a pseudo-document to the count of documents judged. Thus, the estimated TPR (eTPR) is:

$$eTPR = \frac{|TP| + 0.5}{|TP| + |FN| + 1} \quad (6)$$

and the estimated FPR is:

$$eFPR = \frac{|FP| + 0.5}{|FP| + |TN| + 1} \quad (7)$$

We use the estimated rates for all calculations.

### 3. RESULTS AND DISCUSSION

Across the six topics, 61 unique crowd-sourced participants contributed judgments with at least 8 participants per topic. Table 3 shows the number of participants per topic and the number of retained participants meeting the study criteria. The laboratory study had 18 participants with 6 participants per topic.

The largest difference between the two groups is that while on average 84% of crowd-source participants do not qualify for inclusion in the final set of participants for a given topic, all of the laboratory participants qualified. The 84% figure is per topic and overstates the rejection rate for the study. CrowdFlower recorded judgments from 202 unique participants across the 6 topics, and we retained 61 of these participants for the study. As such, we retained 30% of the participants and only rejected 70% of them. The higher value of 84% is caused by participants failing to be accepted on all topics for which they attempted participation. In future work, we plan to look at changing the criteria such that if participant qualifies for any topic, then that participant

| Topic   | CrowdFlower  |          |          |            | Study Criteria |            |
|---------|--------------|----------|----------|------------|----------------|------------|
|         | Participants | Retained | Rejected | % Rejected | Retained       | % Retained |
| 310     | 96           | 15       | 81       | 84%        | 11             | 11%        |
| 336     | 46           | 11       | 35       | 76%        | 8              | 17%        |
| 362     | 100          | 20       | 80       | 80%        | 19             | 19%        |
| 367     | 134          | 17       | 117      | 87%        | 13             | 10%        |
| 426     | 99           | 21       | 78       | 79%        | 23             | 23%        |
| 427     | 61           | 13       | 48       | 79%        | 8              | 13%        |
| Average | 89           | 16       | 73       | 81%        | 14             | 16%        |

**Table 3: Number of crowd-sourced participants, the number of participants rejected at some point by CrowdFlower given low gold accuracy, and the number of participants included in the study based on the study’s criteria for inclusion.**

will be qualified for all topics. The results in the current paper may present the crowd-sourced participants as being better than they really are.

We think the large percentage of crowd-source participants who did not qualify were participants trying to earn money without doing the required work. We suspect that these participants could obtain the required accuracy to qualify if they truly attempted the task.

Table 4 shows the judging behavior of both the crowd-sourced and laboratory participants. Pairs of numbers in Table 4 are bold if there is a statistically significant difference ( $p < 0.05$ ) between the measure’s value for the crowd-source vs. the laboratory participants. We measure statistical significance with a two-sided Student’s t-test for the per-topic measures. For the averages across the six topics, we use a paired, two-sided Student’s t-test with the pairs being the topics.

Both groups have true positive rates that are quite similar for all but topic 310. On the other hand, the crowd-sourced participants have a much higher false positive rate than the laboratory participants. While not significant at the 0.05 level, the laboratory participants appear to be better able to discriminate between relevant and non-relevant documents compared to the crowd-sourced participants ( $d'$  of 2.2 vs. 1.9 with a p-value of 0.08). This apparently better discrimination ability though did not result in a statistically significant difference in accuracy. The laboratory participants were more conservative in their judgments with a criterion of 0.51 vs. the crowd-source participant’s 0.14 ( $p < 0.01$ ). The difference in criterion though is coming largely from differences in the false positive rate and not a correspondingly large difference in the true positive rate.

While crowd-sources participants with gold accuracy of less than 70% were filtered out by CrowdFlower, we still have crowd-source participants who have an accuracy of less than 70% in the final set of participants. Of the crowd-sourced 61 participants, 14 (22%) had final accuracies of less than 70% on at least one topic. The minimum accuracy was 54% for a participant with 70 judgments. Of the 18 laboratory participants, 4 (23%) had final accuracies of less than 70% on at least one of the two topics they completed. The minimum accuracy was 63%. This low-accuracy laboratory participant was likely not guessing, for a one-sided binomial test gives a p-value of 0.08 for the rate not being equal to 50%.

Our results are very similar to ones we have reported for NIST assessors compared to a different set of laboratory participants [8] than those in this study. The results are similar in that both groups have similar true positive rates but have

very different false positive rates. In addition, in our previous study we found laboratory participants to be close the neutral in their criterion while the NIST assessors were more conservative. Interestingly, here the laboratory participants have a low false positive rate and are conservative while in our other work it was the NIST assessors. While the topics were the same in both this paper and [8], the documents were not. In our other work, the documents were all highly ranked documents while in this paper the documents have been randomly selected from the pool of NIST judged documents. Another difference between the studies is that we put the laboratory participants here through a more involved tutorial and administered a qualification test. It may be that the true positive rate is limited by the amount of time participants can give to studying a document while the false positive rate can be affected by the training participants receive.

In terms of the time it takes participants to judge documents, the crowd-sourced participants judged documents nearly twice as fast as the laboratory participants (15 vs. 27 seconds,  $p = 0.01$ ).

In summary, the two groups of participants behaved differently. The biggest difference between the groups is the large fraction of crowd-sourced participants that must have their participation in the study ended early for failure to conscientiously perform the assigned tasks. The differences between the retained crowd-sourced participants and the laboratory participants were firstly the rate at which the two groups work and secondly the false positive rate. We cannot conclusively say that the crowd-sourced environment caused these differences as the two groups were not trained and qualified in exactly the same manner. In future work, we will try to make the crowd-source process better match that of the laboratory study with a qualification separate from the actual task of judging documents.

## 4. CONCLUSION

We conducted two experiments where in each experiment participants judged the relevance of a set of documents. One experiment had crowd-sourced participants while the other had university students and was conducted in a laboratory setting. A large fraction of crowd-source workers did not qualify for inclusion in the final set of participants while all of the laboratory participants did qualify. Judging behavior was similar between the two groups except that crowd-sourced participants had a higher false positive rate

| Topic | True Positive Rate |             |         | False Positive Rate |             |         | $d'$  |     |         | Criterion $c$ |             |         |
|-------|--------------------|-------------|---------|---------------------|-------------|---------|-------|-----|---------|---------------|-------------|---------|
|       | Crowd              | Lab         | p-value | Crowd               | Lab         | p-value | Crowd | Lab | p-value | Crowd         | Lab         | p-value |
| 310   | <b>0.67</b>        | <b>0.42</b> | < 0.01  | <b>0.15</b>         | <b>0.02</b> | 0.03    | 1.7   | 1.8 | 0.81    | <b>0.38</b>   | <b>1.09</b> | < 0.001 |
| 336   | 0.72               | 0.63        | 0.79    | <b>0.12</b>         | <b>0.04</b> | 0.03    | 2.0   | 2.3 | 0.45    | 0.26          | 0.68        | 0.19    |
| 362   | 0.87               | 0.83        | 0.08    | 0.18                | 0.09        | 0.09    | 2.4   | 2.6 | 0.52    | -0.16         | 0.22        | 0.10    |
| 367   | 0.74               | 0.78        | 0.12    | 0.15                | 0.06        | 0.11    | 2.0   | 2.5 | 0.09    | 0.30          | 0.47        | 0.28    |
| 426   | 0.84               | 0.77        | 0.21    | 0.17                | 0.10        | 0.18    | 2.4   | 2.2 | 0.54    | 0.04          | 0.30        | 0.21    |
| 427   | 0.69               | 0.68        | 0.92    | <b>0.27</b>         | <b>0.12</b> | 0.01    | 1.2   | 1.9 | 0.08    | <b>0.04</b>   | <b>0.31</b> | 0.35    |
| All   | 0.75               | 0.69        | 0.15    | <b>0.17</b>         | <b>0.07</b> | < 0.001 | 1.9   | 2.2 | 0.08    | <b>0.14</b>   | <b>0.51</b> | < 0.01  |

| Topic | Accuracy |      |         | Seconds per Judgment |           |         |
|-------|----------|------|---------|----------------------|-----------|---------|
|       | Crowd    | Lab  | p-value | Crowd                | Lab       | p-value |
| 310   | 0.76     | 0.71 | 0.19    | <b>15</b>            | <b>37</b> | 0.04    |
| 336   | 0.80     | 0.81 | 0.90    | <b>6</b>             | <b>24</b> | < 0.001 |
| 362   | 0.85     | 0.89 | 0.32    | 20                   | 28        | 0.32    |
| 367   | 0.80     | 0.88 | 0.07    | 18                   | 24        | 0.51    |
| 426   | 0.84     | 0.85 | 0.73    | 15                   | 20        | 0.46    |
| 427   | 0.71     | 0.80 | 0.21    | 18                   | 27        | 0.39    |
| All   | 0.80     | 0.82 | 0.24    | <b>15</b>            | <b>27</b> | 0.01    |

Table 4: Judging behavior results. Pairs in bold are statistically significant differences ( $p < 0.05$ ).

and judged documents at a rate nearly twice as fast as the laboratory participants.

## 5. ACKNOWLEDGMENTS

Special thanks to Alex Sorokin and Vaughn Hester for their help with CrowdFlower. This work was supported in part by CrowdFlower, in part by NSERC, in part by Amazon, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

## 6. REFERENCES

- [1] O. Alonso and S. Mizzaro. 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*, pages 15–16, July 2009.
- [2] V. Carvalho, M. Lease, and E. Yilmaz. Crowdsourcing for search evaluation. *ACM SIGIR Forum*, 44(2):17–22, December 2010.
- [3] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, 2010.
- [4] H. J. Jung and M. Lease. Improving Consensus Accuracy via Z-score and Weighted Voting. In *Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI*, 2011. Poster.
- [5] M. Lease, V. Carvalho, and E. Yilmaz, editors. *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*. Geneva, Switzerland, July 2010.
- [6] N. Macmillan and C. Creelman. *Detection theory: a user's guide*. Lawrence Erlbaum Associates, 2005.
- [7] R. McCreddie, C. Macdonald, and I. Ounis. Crowdsourcing blog track top news judgments at TREC. In *WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, 2011.
- [8] M. D. Smucker and C. P. Jethani. Measuring assessor accuracy: A comparison of NIST assessors and user study participants. In *SIGIR'11*. ACM, 2011.