# Breaking Bad - Understanding Behavior of Crowd Workers in Categorization Microtasks

Ujwal Gadiraju, Patrick Siehndel, and Besnik Fetahu
L3S Research Center
Appelstr. 9a
30167 Hanover, Germany
{gadiraju, siehndel, fetahu}@L3S.de

Ricardo Kawase
mobile.de
Marktplatz 1
14532 Europarc Dreilinden, Germany
rkawase@team.mobile.de

## ABSTRACT

Crowdsourcing systems are being widely used to overcome several challenges that require human intervention. While there is an increase in the adoption of the crowdsourcing paradigm as a solution, there are no established guidelines or tangible recommendations for task design with respect to key parameters such as *task length*, *monetary incentive* and *time required for task completion*. In this paper, we propose the tuning of these parameters based on our findings from extensive experiments and analysis of '*categorization*' tasks. We delve into the behavior of workers that consume categorization tasks to determine measures that can make task design more effective.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## Keywords

Crowdsourcing; Workers; Behavior; Categorization; Microtasks; Incentives; Task Length

## 1. INTRODUCTION

With the advent of the Internet age and its ubiquity, more and more people are turning towards standardized crowdsourcing platforms in order to service needs requiring large-scale human input. Amazon's Mechanical Turk[1] was the first such crowdsourcing platform to gain widespread popularity, followed by CrowdFlower[2]. Over the last decade there has been a considerable amount of research that has investigated means to improve the quality of results produced via crowdsourcing, and methods to measure performance metrics such as reliability or accuracy of workers in the crowd [5, 10]. However, not all task administrators are well-versed with using the existing platforms to their full potential. This is largely due to two major reasons; (i) there are either few or no concrete guidelines that are task specific and aid an administrator during the important phase of task design, and (ii) there are no existing principles based on which a task administrator can adjust important parameters such as length of a task, or incentive to be offered in order to obtain optimal results in the presence of any limiting constraints.

In this paper, we aim to take the first steps towards tackling the aforementioned challenges by studying the behavior of workers in a crowdsourcing paradigm with varying parameters (task length and incentive), through extensive experiments. We discuss our observations that can aid a task administrator with adjusting key parameters during the task design phase. Based on our study of their behavior, we model workers in the particular task type of '*categorization*'. We choose this task type since it is one of the most popularly crowdsourced task within the taxonomy of microtasks introduced in our previous work [3]. By relying on behavioral metrics for crowd workers, and investigating the behavior flow of workers within tasks, we establish the following guidelines to obtain optimal results from crowdsourced *categorization tasks*. A task administrator is recommended to design tasks with; (i) low to moderate monetary incentives (of the order of a few USD cents), (ii) shorter task lengths (of the order of a few minutes), and (iii) provide ample time to the workers for task completion (we recommend defining the minimum limit, but not the maximum).

## 2. RELATED LITERATURE

Earlier works have shown that task specific features of microtasks affect different types of microtasks differently [1, 3]. Here we study categorization tasks, discuss findings from earlier works that hold for this type of tasks, and present advances through our work.

### 2.1 Task Design and Quality of Results

Marshall et al. analyzed workers who took surveys on Amazon's Mechanical Turk and examined how the characteristics of the surveys influenced the reliability of the data produced [8]. We build on a similar premise and gather data from categorization tasks with varying settings, in order to conduct a meaningful analysis of worker behavior and arrive at sound insights for task design.

Mason et al. studied the effect of varying financial incentives on the performance of workers [9]. Authors conclude that increasing monetary incentives of microtasks attracts more workers to the tasks but does not improve the quality of the results produced. We study this effect in categoriza-

[1] `https://www.mturk.com/mturk/`
[2] `http://www.crowdflower.com/`


*HT '15,* September 1–4, 2015, Guzelyurt, Northern Cyprus.

DOI: http://dx.doi.org/10.1145/2700171.2791053.

tion microtasks, while additionally analyzing optimal incentives as per the length of a task. In contrast to our work in this paper, previous works have investigated methods to improve the quality of the results produced and the reliability of workers in crowdsourcing tasks in general. Oleson et al. present a method to achieve quality control for crowdsourcing, by providing training feedback to workers while relying on programmatic creation of gold data [10].

## 2.2 Worker Behavior - Influential Factors

Eickhoff et al. acknowledged the importance of understanding worker behavior in order to develop reliability metrics and design fraud-proof tasks [2]. Kazai et al. used behavioral observations to define the types of workers in the crowd [7]. By type-casting workers as either sloppy, spammer, incompetent, competent, or diligent, the authors expect their insights to help in designing tasks and attracting the best workers to a task. While the authors correlate these types to the personality traits of workers, we aim to unravel how the behavioral patterns of workers vary with changes in the task design in categorization microtasks. Eickhoff et al. additionally evaluated factors such as the size of microtask, interface used and composition of the crowd [1]. Based on this the authors suggest to design microtasks in a manner that discourages malicious workers. The authors acknowledge that there can be varying effects based on the type of crowdsourced tasks. We thereby, take this further by evaluating the impact of task length and incentives on workers behavior of categorization tasks, while utilising behavioral metrics and task-related characteristics.

# 3. TASKS DESIGN

We aim to analyse the behavior of workers during the consumption of '*categorization*' tasks, under varying task conditions (with respect to length of the task and incentive offered). In order to do so, we deployed 9 different image categorization tasks on CrowdFlower catering to varying task settings. From each task that was deployed we collected responses from 100 distinct workers in the crowd.

We varied the parameters of length and monetary incentives, according to which the categorization tasks were deployed. We experimented with three different variations in the length of the task (20, 30, and 40 units[3]). At the same time, we considered 3 different monetary offers of 1, 2, and 3 USD cents.

## 3.1 Categorization Tasks

Each categorization task was formulated with very clear instructions and help-snippets, in order to avoid introducing bias or bad responses due to poor task design. Figure 1 shows an example of the categorization task that workers had to perform for each unit in the task. An image was presented and workers had to select the most suitable category in each Set (1-5) consisting of 10 different categories. Since the aim of the task was to assess performance related behavior of workers under varying task related circumstances, it was important to ensure that there was no ambiguity within the categories provided as options. Hence, we manually tailored each unit by choosing images that are comprehensible, explicit and unambiguous. In addition, we hand-picked unmistakeable categories for each set, such that the correct category is easy to match for any diligent worker. In order to reflect realistic categorization tasks, we added 9 additional categories apart from the correct one for each of the five sets of options. By doing so, we also minimize the chance of workers selecting the accurate options at random.

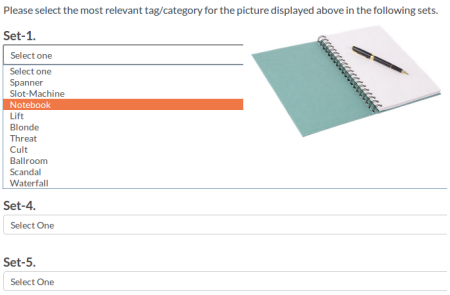


**Figure 1: Workers were asked to select the most suitable category corresponding to each image displayed in the unit within the task (image is overlayed here).**

For each unit, workers were required to select a category from the first set (Set-1), and the selection from the next 4 sets was made optional. By doing so we can measure the extra effort that a worker puts into the task to help the task administrator, as discussed in later sections. This design choice was motivated by the findings of Rogstadius et al., where the authors found that framing a task as helping others increases the intrinsic motivation of workers, and improves the quality of the responses produced [11].

Tasks were deployed non-concurrently such that at any point in time, there was no more than one task available to the workers for consumption. By doing so, we curtail the bias which may otherwise have crept in owing either to the consumption of the categorization tasks with higher incentive, or through workers getting too used to this particular task design. In addition, it is important to note that we randomized the order in which different workers received the units within a task.

## 3.2 Dataset

In total we collected 27,000 unit judgments with at least one tag provided (from mandatory Set-1). In 88% of the cases (23,767) workers provided answers for all sets (Set-1 through to Set-5). The average time to complete the task was 11.3 minutes for tasks with 20 units, 16.4 minutes for tasks with 30 units, and 18.6 minutes for tasks with 40 units. In total, 900 workers participated in our study. We present no further information regarding gender, age, etc. due to the anonymous identity of workers.

# 4. DATA ANALYSIS

First, we present some definitions which will be used hereafter in this paper. These definitions relate to the behavior of microtask workers in a crowd.

**Ineligible Workers.** Crowdsourcing microtasks present the workers in the crowd with a task description and a set of instructions that the workers must follow, for successful

[3]On CrowdFlower a *unit* is the basic building block of a task. So, $n$ units in a task implies that a worker has to complete $n$ workflows of the same type.

Table 1: Consistency in the difficulty of units within a task across all configurations.

| Task Configuration (Units X USD cents) | 20x1 | 20x2 | 20x3 | 30x1 | 30x2 | 30x3 | 40x1 | 40x2 | 40x3 | Average Accuracy in % |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy in % | 92.90 ±1.25 | 90.95 ±2.42 | 91.40 ±1.67 | 90.90 ±2.19 | 88.97 ±1.94 | 87.14 ±2.28 | 90.00 ±2.33 | 85.95 ±2.47 | 88.03 ±3.72 | 89.58 ±2.25 |

task completion. Those workers who do not conform to the priorly stated pre-requisites, belong to this category.

**Tipping Point.** The first point (i.e., the unit index) at which a worker begins to provide unacceptable responses after having provided at least one acceptable response, is called the *tipping point* [4].

**Beaver Workers.** Some workers put in additional effort in order to help out the task administrator by answering optional questions. Such hard-workers are called *beavers*, and the additional effort is referred to as *extra effort*.

We find 9 *ineligible workers* who used browser-embedded translators in order to attempt these tasks[4]. Such workers may or may not provide valid responses, but their responses cannot be used by the task administrator since they do not satisfy the pre-requisites. We discard these workers from further analysis. Despite requiring to provide only one mandatory response from the 5 sets of categories, we observe that several workers (over 88%) go the extra mile by providing responses and identifying categories from additional sets.

## 4.1 Consistency of Units within a Task

In this paper, since we aim to study the behavior of workers as they proceed through a task, it is important to ensure that the difficulty in answering each unit accurately is consistent throughout the task. We ensure that each unit within a task can be easily categorized, and each set has only one category that directly relates to the corresponding image.

Table 1 presents the average accuracy that workers attain within each task configuration. In each task, workers received the units for completion in a randomized order. We observe that every configuration begets an accuracy around 90% with little standard deviation. This confirms that the task design does not introduce bias through the potential variance in the consistency of units comprising the task.

Due to the nature of the task design and setup, we expect workers to achieve an accuracy of 100% without any hindrance. However, owing to possible drifts in attention spans of workers or boredom induced by the repetitive nature of the categorization task, workers could commit mistakes inadvertently. We therefore decide to tolerate 10% incorrect responses from each worker with respect to each task. Based on the simplicity of these particular categorization tasks, we reason that any further incorrect responses should not be merely alluded to the inattentiveness of workers. Hence, we consider the following definitions.

**Bad Workers.** We define *bad workers* as those workers who answer 10% or more of the units within a categorization task incorrectly.

**Poor Starters.** We define *poor starters* as those bad workers whose first 2 responses within a categorization task are incorrect.

[4]This information can be extracted based on the results from CrowdFlower.

## 4.2 Tipping Point

As observed in our previous work [4], we find that several workers provide acceptable responses to begin with, before derailing towards poor accuracy. We thereby investigate this tendency of workers to trail off into providing inaccurate responses, and present our findings here.

Figure 2 presents a comparison between the number of poor starters and bad workers across the different tasks that we deployed. We observe that as the length of the task increases, the number of *bad workers* also increases. In addition, we find that an increase in monetary incentive also increases the number of *bad workers*. This reinforces the findings of [2], where the authors discuss that higher monetary incentives correlate positively with the number of malicious workers attracted to the task. As a consequence we also observe that for a fixed length of a task, the average tipping point of workers decreases with an increase in monetary incentive. This is due to the increase in the number of bad workers to a task with increasing incentive, since *bad workers* tend to 'tip' relatively early in a task. We do not observe a discernible trend in the case of *poor starters*.

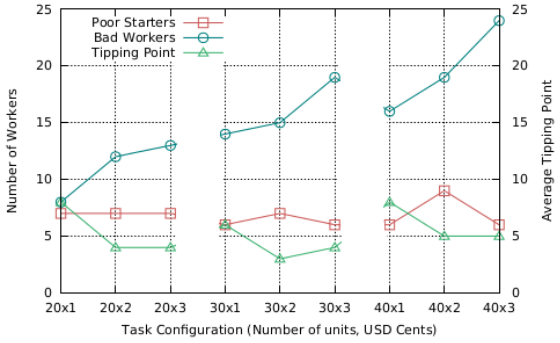


**Figure 2: Number of *Poor Starters* and *Bad Workers*, and the average *Tipping Point* of workers across different tasks.**

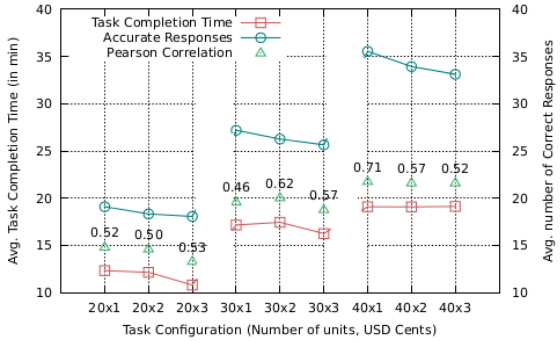


**Figure 3: Correlation between the average task completion time (scaled on the y-axis) and average accuracy of workers for all tasks (scaled on the y2-axis).**

## 4.3 Completion Time vs Worker Accuracy

We computed the average time that workers take to complete the tasks and their corresponding average accuracy, for all task configurations (different task lengths and varying incentives). Figure 3 presents our findings. We see that for a fixed length of the task, the average accuracy of workers increases with an increase in the task completion time, with high Pearson Correlation (see Figure 3).

We notice that this observation holds across the tasks with varying task lengths (20, 30, and 40 units). Our findings align with previous works where authors have shown that with an increase in monetary incentive, there are more workers who are attracted to a task, but the accuracy of the workers is not effected [9]. At the same time, an increase in the monetary incentive increases the number of malicious workers that are attracted to the task, since their priority is to attain immediate financial gains through quick task completion [6, 12].

## 4.4 Worker Behavior within a Task

We also studied the worker behavior specifically within each task, across all the task configurations, i.e., how a worker's accuracy evolves as one proceeds through the index of units; from the first till the last unit within the given task. Table 2 presents our findings of the correlations between the *unit index* (UI) and *accuracy* of a worker, the accuracy of a worker and the amount of *extra effort*, and finally the *unit index* and the amount of *extra effort*. The correlation is measured using Pearson's **r**. Note that the unit index represents how far along a worker is within a task during the task consumption.

**Table 2: Evolution of the *accuracy* and *extra effort* of workers through the course of a task, across different task configurations.**

| Task Configuration (Units X USD cents) | Pearson's r (UI,Accuracy) | Pearson's r (Extra,Accuracy) | Pearson's r (UI,Extra) |
|---|---|---|---|
| **20x1** | 0.25 | 0.22 | -0.58 |
| **20x2** | -0.42 | 0.14 | -0.57 |
| **20x3** | -0.49 | 0.68 | -0.69 |
| **30x1** | -0.64 | 0.71 | -0.71 |
| **30x2** | -0.27 | 0.46 | -0.64 |
| **30x3** | -0.61 | 0.80 | -0.66 |
| **40x1** | -0.68 | 0.80 | -0.74 |
| **40x2** | -0.48 | 0.70 | -0.80 |
| **40x3** | -0.75 | 0.71 | -0.75 |

The general trend we observe, is that as a worker proceeds from the first unit to the last unit of a task, the worker's accuracy decreases. Here the accuracy is computed only on the compulsory responses from Set-1 for each unit, and the optional sets are not considered. From the Table 2, in column 2 we can see that the negative correlation grows stronger with increasing length of the task.

Next we study the relationship between the *extra effort* that workers exert and their accuracy within the task (once again we consider only the first set which is compulsory for each unit, while computing a worker's accuracy). We find that workers that exert more extra effort tend to project higher accuracies within the tasks (see column 3). We can see that this positive correlation grows stronger as the length of the task increases. We reason that this is due to the fact that it is more taxing to exert *extra effort* in longer tasks. If workers, still go the extra mile and do so, it indicates their genuine attempt to provide the most suitable responses.

Finally, we investigate the longevity of workers exerting extra effort within a task, i.e., we study whether workers continue to exert the same effort as they progress through the units within a task. We find that as workers proceed through towards task completion they exert lesser extra effort (see column 4). As the length of the tasks increases, this negative correlation between the unit index and the extra effort from workers grows stronger.

## 4.5 Scrutiny of Extra Responses

We take a closer look at the extra responses obtained from the workers through the optional categorization sets (Set-2 through to Set-5) for all the units within tasks, across varying configurations. Table 3 presents the aggregated responses from the 9 task configurations, with respect to each of the 5 sets of options for category selection.

**Table 3: Responses of workers with respect to each set of categories, aggregated across varying task configurations (% Wrong and % Correct are w.r.t. non-skipped responses.)**

| Response Type | Set-1 | Set-2 | Set-3 | Set-4 | Set-5 |
|---|---|---|---|---|---|
| **% Skipped** | - | 10.30 | 10.86 | 11.19 | 11.47 |
| **% Wrong Responses** | 9.23 | 9.52 | 10.92 | 11.58 | 13.31 |
| **% Correct Responses** | 90.77 | 90.48 | 89.08 | 88.42 | 86.69 |

Since the first set was made compulsory, no workers were allowed to skip Set-1. We find that the percentage of correct responses with respect to the extra responses received, gradually decreases from Set-1 to Set-5. We observe that workers tend to skip more optional sets as they proceed from Set-2 to Set-5. This is understandable, considering that workers may find it tedious to exhibit altruistic extra effort throughout the course of a task. Interestingly, of those responses which are provided by workers from Set-2 to Set-5, the percentage of wrong answers gradually increases.

## 4.6 Workers Breaking Bad

In addition to the method presented in previous work [4] to measure the tipping point of workers, we adopt another approach to assess the tipping point of *bad workers* by adjusting it for honest mistakes from workers. We define this relatively less aggressive measure as the *adjusted tipping point* (ATP). Since workers may lose attentiveness or get bored in repetitive tasks, honest workers may stumble at certain units. However, due to the ease of the task as mentioned earlier we do not expect workers to trail towards providing poor responses consecutively.

**Adjusted Tipping Point.** Workers that consecutively respond to at least 10% of the units within the task incorrectly, are said to have an *adjusted tipping point*. The index of the first unit at which the worker provided the first string of 10% or more incorrect answers is the ATP of the worker.

**Breaking Worker/Breaker.** A bad worker who exhibits an Adjusted Tipping Point is said to be a breaker.

For example, consider a task with 30 units, and a worker who responds to units 7,8, and 9 incorrectly. The given worker thus consecutively provided incorrect responses to 10% of the questions. The worker is hence said to be a *breaker* and the worker's ATP is 7. In case workers depict multiple strings of inaccurate responses, the ATP is consid-

Table 4: Types of workers and average ATP of *breakers*, with respect to varying task configurations.

| Task Configuration (Units X USD cents) | Perfect Workers | Poor Starters | Bad Workers | Breakers | Avg. ATP |
|---|---|---|---|---|---|
| **20x1** | 85 | 7 | 8 | 1 | 1 |
| **20x2** | 72 | 7 | 12 | 5 | 1 |
| **20x3** | 76 | 7 | 13 | 6 | 3 |
| **30x1** | 68 | 6 | 14 | 8 | 4 |
| **30x2** | 63 | 7 | 15 | 8 | 2 |
| **30x3** | 62 | 6 | 19 | 13 | 4 |
| **40x1** | 67 | 6 | 16 | 10 | 4 |
| **40x2** | 57 | 9 | 19 | 10 | 4 |
| **40x3** | 51 | 6 | 24 | 18 | 9 |

Table 5: Correlation between acceptable answers and workers' trust score.

| Task Configuration | 20x1 | 20x2 | 20x3 | 30x1 | 30x2 | 30x3 | 40x1 | 40x2 | 40x3 |
|---|---|---|---|---|---|---|---|---|---|
| **Correlation** | 0.85 | 0.82 | 0.66 | 0.76 | 0.74 | 0.71 | 0.67 | 0.78 | 0.55 |

ered to be the index of the first unit of the first occurrence of such a string of responses.

Table 4 presents our findings with respect to the distribution of the different kinds of workers and the ATP of *breakers*. We note that as the length of the task increases, the number of *perfect workers* decreases while the number of *bad workers* and *breakers* increases. We find no significant fluctuation in the number of *poor starters* across varying task configurations. An interesting observation is that with an increase in monetary incentive for a fixed length of the task, we see that the ATP of *breakers* increases to a higher unit index. This tells us that although malicious workers may be prone to getting attracted to tasks with higher rewards (as can be observed from the number of *bad workers* in each task), higher incentives can delay the adjusted tipping point. With an increase in monetary incentive for a fixed length of the task, we do not observe a significant trend with respect to the ATP of *breakers*. However, across different task lengths we observe that with an increase in task length the ATP of workers increases as well.

## 4.7 Can we trust the 'trust-score'?

CrowdFlower additionally provides a trust-score for the workers. This trust score, a value between 0 and 1, represents the accuracy of a worker in a job. Here, we draw a comparison between our findings and the trust-score provided by the platform.

Ideally, all workers could achieve a perfect score, given the simplicity of the tailored units. However, in total we identified 299 workers who did not manage to complete the task perfectly. Under our relaxed definition of *bad workers* which tolerates 10% of inaccuracy, we identified 56 bad workers. Table 5 shows the correlation between the trust-score and the number of correct answers given by the workers. In all cases we see a strong correlation, meaning that workers with a higher trust-score provided by the platform, indeed tend to perform better.

The average trust-score of all workers in the dataset is 0.64. Considering only the trust-scores of bad workers, the average is significantly lower, 0.39. However, we found out that 22 workers (out of the 56 bad workers) have trust-scores above 0.64. This indicates that although the platform associates them with high trust scores, these workers end up providing unacceptable responses, which based on our setup can only be attributed to malicious intent.

## 4.8 Caveats and Limitations

We ensured that each worker participated in only one of the deployed tasks by leveraging worker IDs. In this way, we eliminated the influence of some users who could perform all the tasks, additionally excluding the influence of learning and familiarity with the job, that a worker might bring from one task to another. With more experiments that stretch the limits of length of the task as well as monetary incentives offered, we can propose boundaries for these parameters. This will form a part of our imminent future work.

# 5. DISCUSSION AND CONCLUSIONS

We find that *bad workers* are attracted to tasks with relatively high monetary incentives when compared to those with relatively low incentives, despite being of the same length and requiring the same amount of effort for task completion. This shows that it is of prime importance for a requester to fine-tune the incentive offered, in order to obtain optimal results. From our study we find that it is safer to err on the lower side of the monetary incentive offered for a task, to attain more accurate responses from the crowd. Since we establish that the *task completion time* of a worker is strongly and positively correlated to the worker's accuracy, adequate time must be provided to the crowd for task completion. For optimal results, it is therefore safer to err on the higher side of the time required.

We establish that the accuracy of workers decreases as they proceed in a task, more so towards the end of longer tasks. This shows that a task administrator can profit by splitting a relatively long task into shorter ones before deploying it to the crowd. Our findings suggest that for extracting ideal output from a crowd, it is safer to err on the shorter side with respect to the length of a task.

By giving workers an option to provide additional work through their *extra effort*, we can gather more information regarding the worker and deduce the nature of the worker. We find that workers that exert more extra effort tend to perform with higher accuracies within the tasks. Thus, by identifying and treating these different types of workers accordingly, one can improve the effectiveness of categorization tasks. Finally, this paper sets important ground work for future work. Through further experiments we plan to quantify the limits and guidelines presented in this work.

**Acknowledgements.** This work has been partially funded by the European Commission within the 7th Framework Programme (Grant Agreement no: 600908).

## 6. REFERENCES


[1] C. Eickhoff and A. de Vries. How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*, pages 11–14, 2011.

[2] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[3] U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223. ACM, 2014.

[4] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of CHI'15, CHI Conference on Human Factors in Computing Systems*, 2015.

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

[6] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk. In *AMCIS*, 2011.

[7] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

[8] C. C. Marshall and F. M. Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 234–243, New York, NY, USA, 2013. ACM.

[9] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[10] D. Oleson, A. Sorokin, G. P. Laughlin, V. Hester, J. Le, and L. Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11:11, 2011.

[11] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*, 2011.

[12] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM, 2010.