A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality

Kenji Hata, Ranjay Krishna, Li Fei-Fei, Michael S. Bernstein
Stanford University
{kenjihata, ranjaykrishna, feifeili, msb}@cs.stanford.edu

ABSTRACT

Microtask crowdsourcing is increasingly critical to the creation of extremely large datasets. As a result, crowd workers spend weeks or months repeating the exact same tasks, making it necessary to understand their behavior over these long periods of time. We utilize three large, longitudinal datasets of nine million annotations collected from Amazon Mechanical Turk to examine claims that workers fatigue or satisfice over these long periods, producing lower quality work. We find that, contrary to these claims, workers are extremely stable in their quality over the entire period. To understand whether workers set their quality based on the task's requirements for acceptance, we then perform an experiment where we vary the required quality for a large crowdsourcing task. Workers did not adjust their quality based on the acceptance threshold: workers who were above the threshold continued working at their usual quality level, and workers below the threshold self-selected themselves out of the task. Capitalizing on this consistency, we demonstrate that it is possible to predict workers' long-term quality using just a glimpse of their quality on the first five tasks.

Author Keywords

Crowdsourcing; microtasks; fatigue; consistent; satisficing

ACM Classification Keywords

H.5.3 Group and Organization Interfaces: Computer supported cooperative work

INTRODUCTION

Microtask crowdsourcing is gaining popularity among corporate and research communities as a means to leverage parallel human computation for extremely large problems [2, 3, 31, 33, 53]. These communities use crowd work to complete hundreds of thousands of tasks per day [38], from which new datasets with over 20 million annotations can be produced within a few months [26]. A crowdsourcing platform like Amazon's Mechanical Turk (AMT) is a marketplace subject to human factors that affect its performance, both in terms of speed and quality [12]. Prior studies found that work division in crowdsourcing follows a Pareto principle, where a small minority of workers usually completes a great majority of the work [36]. If such large crowdsourced projects are being completed by a small percentage of workers, then these workers spend hours, days, or weeks executing the exact same tasks. Consequently, we pose the question:

How does a worker's quality change over time?

Multiple arguments from previous literature in psychology suggest that quality should decrease over time. *Fatigue*, a temporary decline in cognitive or physical condition, can gradually result in performance drops over long periods of time [4, 29, 45]. Since the microtask paradigm in large scale crowdsourcing involves monotonous sequences of repetitive tasks, fatigue buildup can pose a potential problem to the quality of submitted work over time [9]. Furthermore, workers have been noted to be "satisficers" who, as they gain familiarity with the task and its acceptance thresholds, strive to do the minimal work possible to achieve these thresholds [8, 51].

To study these long term effects on crowd work, we analyze worker trends over three different real-world, large-scale datasets [26] collected from microtasks on AMT: image descriptions, question answering, and binary verifications. With microtasks comprising over 60% of the total crowd work and microtasks involving images being the most common type [17], these datasets cover a large percentage of the type of crowd work most commonly seen. Specifically, we use over 5 million image descriptions from 2674 workers over a 9 month span, 0.8 million question-answer pairs from 2179 workers over a 3 month span, and 2 million verifications from 3913 workers over a 3 month span. The average worker in the largest dataset worked for an average of 2 eight-hour work days while the top 1% of workers worked for nearly 45 eight-hour work days. Using these datasets, we look at temporal trends in the accuracy of annotations from workers, diversity of these annotations, and the speed of completion.

Contrary to our hypothesis that workers would exhibit glaring signs of fatigue via large declines in submission quality over time, we find that workers who complete large sets of microtasks maintain a consistent level of quality (measured as the percentage of correct annotations). Furthermore, as workers become more experienced on a task, they develop stable strategies that do not change, enabling them to com-

plete tasks faster. But are workers generally consistent or is this consistency simply a product of the task design?

We thus perform an experiment where we hire workers from AMT to complete large-scale tasks while randomly assigning them into different task designs. These designs were varied across two factors: the acceptance *threshold* with which we accept or reject work, and the *transparency* of that threshold. If workers manipulate their quality level strategically to avoid rejection, workers with a high (difficult) threshold would perform at a noticeably better level than the ones with a low threshold who can satisfice more aggressively. However, this effect might only be easily visible if workers have transparency into how they performed on the task.

By analyzing 676,628 annotations collected from 1134 workers in the experiment on AMT, we found that workers display consistent quality regardless of their assigned condition, and that lower-quality workers in the high threshold condition would often self-select out of tasks where they believe there is a high risk of rejection. Bolstered by this consistency, we ask: can we predict a worker's future quality months after they start working on a microtask?

If individual workers indeed sustain constant correctness over time, then, intuitively, any subset of a worker's submissions should be representative of their entire work. We demonstrate that a simple glimpse of a worker's quality in their first few tasks is a strong predictor of their long-term quality. Simply averaging the quality of work of a worker's first 5 completed tasks can predict that worker's quality during the final 10% of their completed tasks with an average error of 3.4%.

Long-term worker consistency suggests that paying attention to easy signals of good workers can be key to collecting a large dataset of high quality annotations [42, 48]. Once we have identified these workers, we can back off the gold-standard (attention check) questions to ensure good quality work, since work quality is unvarying [37]. We can also be more permissive about errors from workers known to be good, reducing the rejection risk that workers face and increasing worker retention [11, 35].

RELATED WORK
Our work is inspired by psychology, decision making, and workplace management literature that focuses on identifying the major factors that affect the quality of work produced. Specifically, we look at the effects of fatigue and satisficing in the workplace. We then study whether these problems transfer to the crowdsourcing domain. Next, we explore how our contributions are necessary to better understand the global ecosystem of crowdsourcing. Finally, we discuss the efficacy of existing worker quality improvement techniques.

Fatigue
Repeatedly completing the same task over a sustained period of time will induce fatigue, which increases reaction time, decreases production rate, and is linked to a rise in poor decision-making [29, 57]. The United States Air Force found that both the cognitive performance and physical conditions of its airmen continually deteriorated during the course of long, mandatory shifts [45]. However, unlike these mandatory, sustained shifts, crowdsourcing is generally opt-in for workers — there always exists the option for workers to break or find another task whenever they feel tired or bored [32, 34]. Nonetheless, previous work has shown that people cannot accurately gauge how long they need to rest after working continuously, resulting in incomplete recoveries and drops in task performance after breaks [16]. Ultimately, previous work in fatigue suggests that crowd workers who continuously complete tasks over sustained periods would result in significant decreases in work quality. We show that contrary to this literature, crowd workers remain consistent throughout their time on a specific task.

Satisficing
Crowd workers are often regarded as "satisficers" who do the minimal work needed for their work to be accepted [8, 51]. Examples of satisficing in crowdsourcing occur during surveys [28] and when workers avoid the most difficult parts of a task [40]. Disguised attention checks in the instructions [44] or rate-limiting the presentation of the questions [21] improves the detection and prevention of satisficing. Previous studies of crowd workers' perspectives find that crowd workers believe themselves to be genuine workers, monitoring their own work and giving helpful feedback to requesters [41]. Workers have also been shown to respond well and produce high quality work if the task is designed to be effort-responsive [18]. However, workers often consider the cost-benefit of continuing to work on a particular task — if they feel that a task is too time-consuming relative to its reward, then they often drop out or compensate by satisficing (e.g. reducing quality) [41]. Prior work has shown that We observe that satisficing does occur, but it only affects a small portion of long-term workers. We also observe in our experiments that workers opt out of tasks where they feel they have a high risk of rejection.

The global crowdsourcing ecosystem
With the rapidly growing size of crowdsourcing projects, workers now have the opportunity to undertake large batches of tasks. As they progress through these tasks, questions arise and they often seek help by communicating with other workers or the task creator [39]. Furthermore, on external forums and in collectives, workers often share well-paying work opportunities, teach and learn from other workers, review requesters, and even consult with task creators to give constructive feedback [20, 39, 41, 49]. When considering this crowdsourcing ecosystem, crowd researchers often envision how more complex workflows can be integrated to make the overall system more efficient, fair, and allow for a wider range of tasks to be possible [24]. To continue the trend towards a more complex, but more powerful, crowdsourcing ecosystem, it is imperative that we study the long-term trends of how workers operate within it. Our paper seeks to identify trends that occur as workers continually complete tasks over a long period of time. We conclude that crowdsourcing workflows should design methods to identify good workers and provide them with the ability to complete tasks with a low threshold for acceptance as good workers work consistently hard regardless of the acceptance criteria.
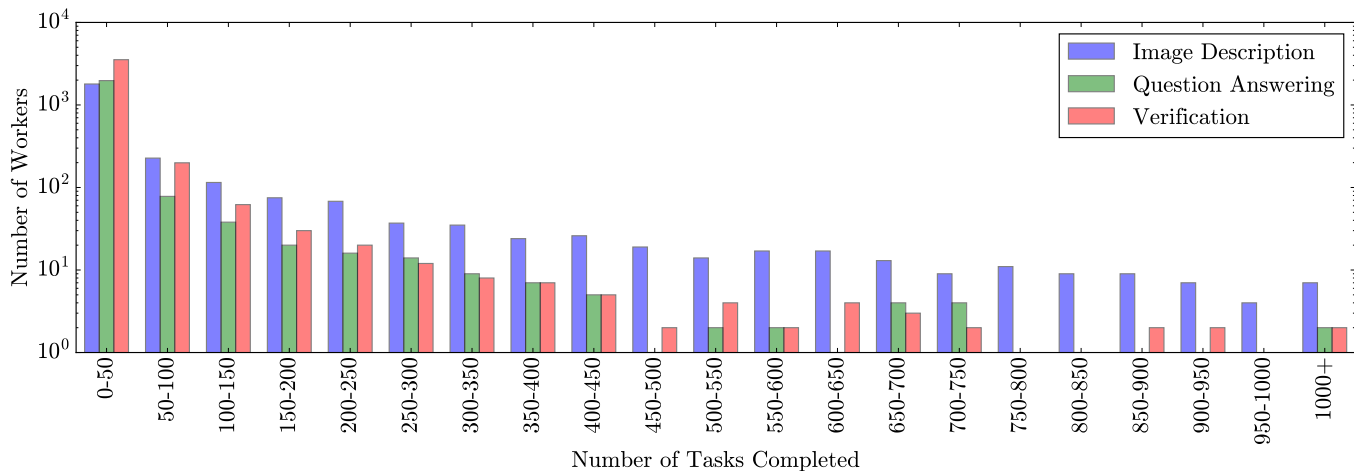
Figure 1: A distribution of the number of workers for each of the three datasets. A small number of persistent workers complete most of the work: the top 20% of workers completed roughly 90% of all tasks.

Improving crowdsourcing quality

External checks such as verifiable gold standards, requiring explanations, and majority voting are standard practice for reducing bad answers and quality control [6,23]. Other methods directly estimate worker quality to improve these external checks [19, 56]. Giving external feedback or having crowd workers internally reflect on their prior work also has been shown to yield better results [13]. Previous work directly targets the monotony of crowdsourcing, showing that by framing the task as more meaningful to workers (for example as a charitable cause), one obtains higher quality results [7]. However, this framing study only had workers do each task a few times and did not observe long-term trends. We, on the other hand, explore the changes in worker quality on microtasks that are repeated by workers over long periods of time.

ANALYSIS: LONG-TERM CROWDSOURCING TRENDS

In this section, we perform an analysis of worker behavior over time on large-scale datasets of three machine learning labeling tasks: image descriptions, question answering, and binary verification. We examine common trends, such as worker accuracy and annotation diversity over time. We then use our results to answer whether workers are fatiguing or displaying other decreases in effectiveness over time.

Data

We first describe the three datasets that we inspect. Each of the three tasks were priced such that workers could earn $6 per hour and were only available to workers with a 95% approval rating and who live in the United States. For the studies in this paper, workers were tracked by their AMT worker ID's. The tasks and interfaces used to collect the data are described in further detail in the Visual Genome paper [26].

Image descriptions. An image description is a phrase or sentence associated with a certain part of an image. To complete this task, a worker looks at an image, clicks and drags to select an area of the image, and then describes it using a short textual phrase (e.g., "The dog is jumping to catch the frisbee").

Each image description task requires a worker to create $5 - 10$ unique descriptions for one randomly selected image, averaging at least 5 words per description. Workers were asked to keep the descriptions factual and avoid submitting any speculative phrases or sentences. We estimate that each task takes around 4 minutes and we allotted 2.5 hours such that workers did not feel pressured for time. In total, 5,380,263 image descriptions were collected from 2674 workers over 9 months.

Question answers. Each question answering task asks a worker to write 7 questions and their corresponding answers per image for 2 different, randomly selected images. Workers were instructed to begin each sentence with one of the following questions: who, what, when, where, why and how [30]. Furthermore, to ensure diversity of question types, workers were asked to write a minimum of 4 of these question types. Workers were also instructed to be concise and unambiguous to avoid wordy and speculative questions. Each task takes around 4 minutes and we allotted 2.5 hours such that workers did not feel pressured for time. In total, 832,880 question-answer pairs were generated by 2179 workers over 3 months.

Binary verifications. Verification tasks were quality control tasks: given an image and a question-answer pair, workers were asked if the question was relevant to the image and if the answer accurately responded to the question. The majority decision of 3 workers was used to determine the accuracy of each question answering pair. For each verification task, a worker voted on 50 randomly-ordered question-answer pairs. Each task takes around 3 minutes and we allotted 1 hour such that workers did not feel pressured for time. In total, 2,498,640 votes were cast by 3913 workers over 3 months.

Overall. Figure 1 shows the distribution of how many tasks workers completed over the span of the data collection period, while Table 1 outlines the total number of annotations and tasks completed. The top 20% of workers who completed the most tasks did 91.8%, 90.9%, and 88.9% of the total work in each of the three datasets respectively. These distributions are similar to the standard Pareto 80-20 rule [36], clearly demon-

	Annotations	Tasks	Workers
Descriptions	5,380,263	605,443	2674
Question Answering	830,625	54,587	2179
Verification	2,689,350	53,787	3913

Table 1: The number of workers, tasks, and annotations collected for image descriptions, question answering, and verifications.
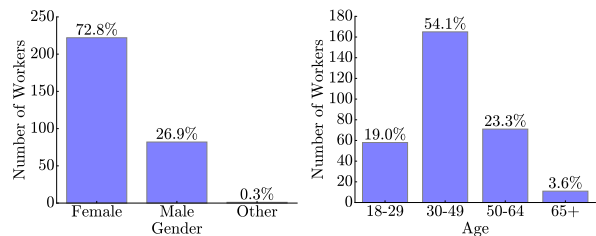

Figure 2: Self reported gender (left) and age distribution (right) of 298 workers who completed at least 100 of the image description, question answer, or binary verification tasks.

strating that a small, but persistent minority of workers completes an extremely large number of similar tasks. We noticed that workers in the top 1% each completed approximately 1% of the respective datasets each, with 5455 image description tasks, 758 question answering tasks, and 1018 verification tasks completed on average. If each of these workers in the top 1% took 4 minutes for image descriptions and question answering tasks and 3 minutes for verification tasks, the estimated average work time equates to 45, 6.2 and 6.2 eight-hour work days for each task respectively. This sheer workload demonstrates that workers may work for very extended periods of time on the same task. Additionally, workers, on average, completed at least one task per week for 6 weeks. By the final week of the data collection, about 10% of the workers remained working on the tasks, suggesting that our study captures the entire lifetime of many of these workers.

We focus our attention on workers who completed at least 100 tasks during the span of the data collection. The completion time for 100 tasks is approximately 6.7 hours for image description and question answering tasks and 5.0 hours for verification tasks. We find that 657, 128, and 177 workers completed 100 of the image description, question answering, and verification tasks respectively. The median worker in each task type completed 349, 220, and 181 tasks, which translates to 23.2, 14.6, and 6.0 hours of continuous work. These workers also produced 94.5%, 70.5%, and 66.3% of each of the total annotations. These worker pools are relatively unique: there are 61 shared workers between image descriptions and QA, 69 shared workers between image description and verification, 42 shared workers between question answering and verifications, and 25 shared workers between all three tasks.

We reached out to the 815 unique workers who had worked on at least 100 tasks and asked them to complete a survey. After collecting 305 responses, we found the gender distribution to be 72.8% female, 26.9% male, and 0.3% other (Figure 2). Furthermore, we found that workers with ages 30-49 were the majority at 54.1% of the long-term worker population. Ages 18-29, 50-64, and 65+ respectively comprised 19.0%, 23.3% and 3.6% of the long-term worker population. Compared to the distributions in previously gathered demographics on AMT [11, 17, 47], the gender and age distribution of all workers closely aligns with these other previously gathered distributions [26]. However, the distribution of long-term workers is skewed towards older and female workers.

Workers are consistent over long periods
We analyzed worker accuracy and annotation diversity over the entire period of time that they worked on these tasks. Because workers performed different numbers of tasks, we nor-

malize time data to percentages of their total *lifetime*, which we define as the period from when a worker starts the task until they stop working on that task. For example, if one worker completed 200 tasks and another completed 400 tasks, then the halfway point in their respective lifetimes would be when they completed 100 and 200 tasks.

Annotation accuracy. A straightforward metric of quality is the percentage of microtasks that are correct. To determine accuracy for an image description or question answering task, we computed the percentage of descriptions or question-answer pairs deemed true by a majority vote made by other workers. However, to use this majority vote in a metric, we need to first validate that this verification process is repeatable and accurate. Since the ground truth of verification tasks is unknown at such a large scale, we need a method to estimate the accuracy of each verification decision. We believe that comparing a worker's vote against the majority decision is a good approximation of accuracy. To test accuracy, we randomly sampled a set of $1,000$ descriptions and image answers and manually compared our own verifications against the majority vote, which resulted in a 98.2% match. To test repeatability, we randomly sampled a set of $15,000$ descriptions and question answers to be sent back to be voted on by 3 new workers 6 months after the initial dataset was collected. Ultimately, we found a 99.3% similarity between the majority decision of this new verification process with the original decision reported in the dataset [26]. The result of this test indicates that the majority decision is both accurate and repeatable, making it a good standard to compare against.

We find that workers change very little over time (Figures 3 and 4). When considering those who did at least 100 image description tasks, people on average started at 97.9 ± 12.1% accuracy and ended at 96.6 ± 9.1%, averaging an absolute change of 3.3 ± 5.6%. Workers who did at least 100 question answering tasks started with an average of 88.4 ± 6.3% and ended at 87.5 ± 6.0%, resulting in an absolute change of 3.1 ± 3.3%. For the verification task, workers agreed with the majority on average 88.1 ± 3.6% at the start and 89.0 ± 4.0% at the end, resulting in an absolute change of 3.1 ± 3.4%.

Annotation diversity. Accuracy captures clearly correct or incorrect outcomes, but how about subtler signals of effort level? Since each image description or question answering task produces multiple phrases or questions, we examine the linguistic similarity of these phrases and questions over time.
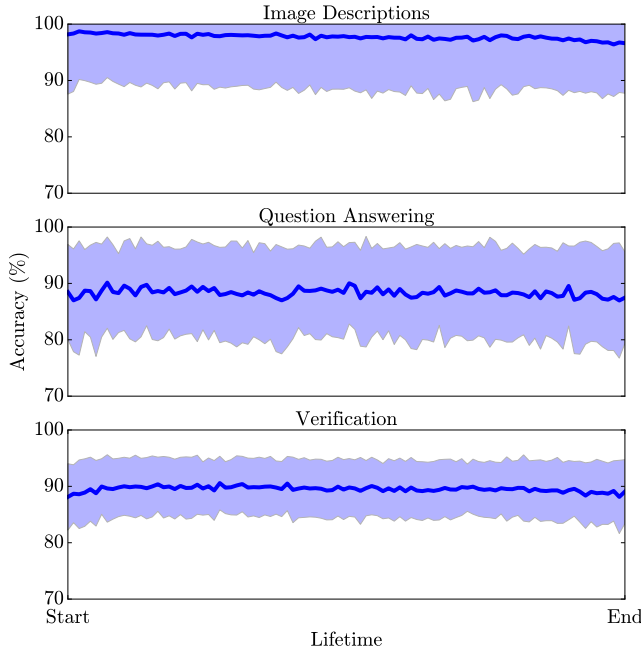
Figure 3: The average accuracy over the lifetime of each worker who completed over 100 tasks in each of the three datasets. The top row shows accuracy for image descriptions, the middle row shows accuracy for question answering, and the bottom row shows accuracy for the verification dataset.

As N-grams have often been used in language processing for gauging similarity between documents [10], we construct a metric of syntax diversity for a set of annotations as follows:

$$\text{diversity} = \frac{\text{number of unique N-grams}}{\text{number of total N-grams}}. \qquad (1)$$

As the annotation set increasingly contains different words and ordering of words, this diversity metric approaches 1 because the number of unique N-grams will approach the total possible N-grams. Conversely, if the annotation set contains increasingly similar annotations, many N-grams will be redundant, making this diversity metric approach 0. To account for workers reusing similar sentence structure in consecutive tasks, we track the number of unique N-grams versus total N-grams in sequential pairs of tasks.

Figure 5 illustrates that the percentage of unique bigrams decreases slightly over time. In the image description task, the percent of unique bigrams decreases on average from 82.4% to 78.4% between the start and end of a worker's lifetime. Since there are 4.2 bigrams on average per phrase, a worker writes approximately 42 total bigrams per task. Thus, a decrease in 4.0% results in a loss of 1.7 unique bigrams per task. In the question answering task, the percent of unique bigrams decreases on average from 60.7% to 54.0%. As there are on average 3.4 bigrams per question, this 6.7% decrease would cost a loss of 3.2 distinct bigrams per task. Ultimately, these results show that over the course of a worker's lifetime, only a small fraction of diversity is lost, as less than a sentence or question's contribution of bigrams is lost.
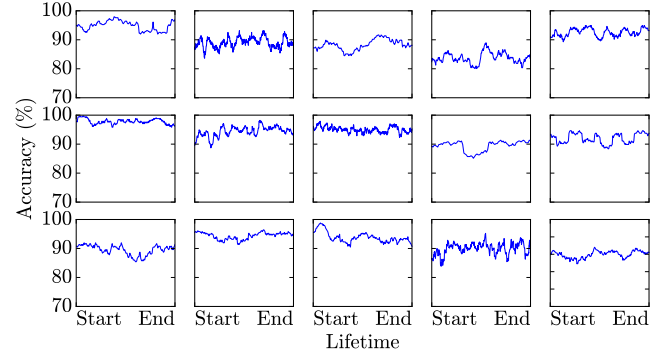

Figure 4: A selection of individual workers' accuracy over time during the question answering task. Each worker remains relatively constant throughout his or her entire lifetime.

A majority of workers stay constant during their lifetime. However, a few workers decrease to an extremely low N-gram diversity, despite writing factually correct image descriptions and questions. This behavior describes a "satisficing" worker, as they repeatedly write the same types of sentences or questions that generalize to almost every image. Figure 6 demonstrates how a satisficing worker's phrase diversity decreases from image-specific descriptions submitted in early-lifetime tasks to generic, repeated sentences submitted in late-lifetime tasks. To determine the percentage of total workers who are satisficing workers, we first compute the average diversity of submissions for each worker. We then set a threshold equal the difference between the maximum and mean of these diversities, labeling workers below the mean by this threshold as satisficers. We find that approximately 7% and 2% of workers satisfice in the image description and question answering datasets respectively.

Annotation speed. We recorded the time it takes on average for workers to complete a single verification. We removed 2.4% of the data points deemed as outliers from this computation, as workers will infrequently take longer times during a break or while reading the instructions. We defined outliers for each task of 50 verifications as times that outside 3 standard deviations of the mean time for those 50 verifications. Overall, Figure 7 demonstrates that workers indeed get faster over time. Initially, workers start off taking 4.5 seconds per verification task, but end up averaging under 3.4 seconds per task, resulting in an approximate 25% speedup. Although no time data was recorded for either the image descriptions or question answering tasks, we believe that they would also exhibit similar speedups over time due to practice effects [43] and similarities in the correctness and diversity metrics.

Discussion

No significant fatigue effects are exhibited by in long-term workers. Workers do not appear to suffer from long-term fatigue effects. With an insignificant average accuracy drop of on average 1.5% for workers across their lifetime, we find that workers demonstrate little change in their submission quality. Instead of suffering from fatigue, workers may be opting out or breaking whenever they feel tired [9]. Furthermore, this
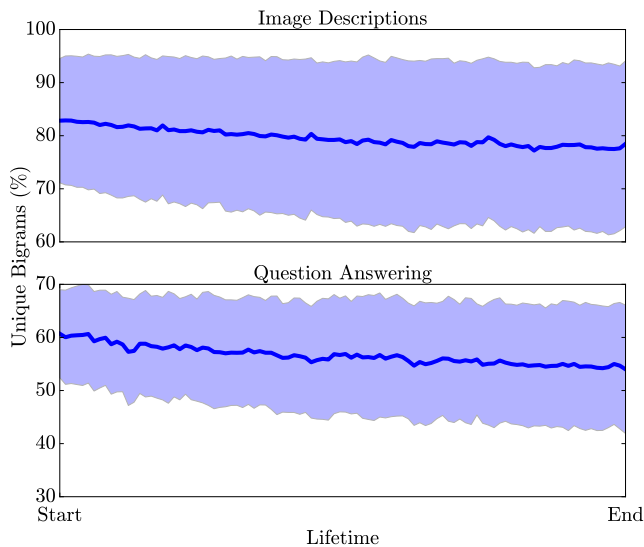
Figure 5: On average, workers who repeatedly completed the image description (top row) or question answering (bottom row) tasks gave descriptions or questions with increasingly similar syntactic structures.

finding agrees with previous literature that cumulative fatigue is not a major factor in quality drop [45].

Accuracy is constant within a task type, but varies across different task types. We attribute the similarity between the average accuracy of the question answering and verification tasks to their sequential relationship in the crowdsourcing pipeline. If the question-answer pairs are ambiguous or speculative, then the majority vote often becomes split, resulting in accuracy loss for both the question answering and verification tasks. Additionally, we notice the average accuracy for image descriptions is noticeably higher than the average accuracy for either the question answering or verification datasets. We believe this discrepancy stems from the question answering task's instructions that ask workers to write at four distinct types of *W* questions (e.g. "why", "what", "when"). Some question types such as "why" or "when" are often ambiguous for many images (e.g. "why is the man angry?"). Such questions are often marked as incorrect by other workers in the verification task. Furthermore, we also attribute the disparity between unique bigram percentage for the image description and question answering tasks to the question answering task's instructions that asked workers to begin each question with one of the 7 question types.

Experience translates to efficiency. Workers retain constant accuracy, and slightly reduce the complexity of their writing style. Combined, these findings suggest that workers find a general strategy that leads to acceptance and stick with it. Studies of practice effects suggest that a practiced strategy helps to increase worker throughput according to a power law [43]. This power law shape is clearly evident in the average verification speed, confirming that practice plays a crucial role in the worker speedup.
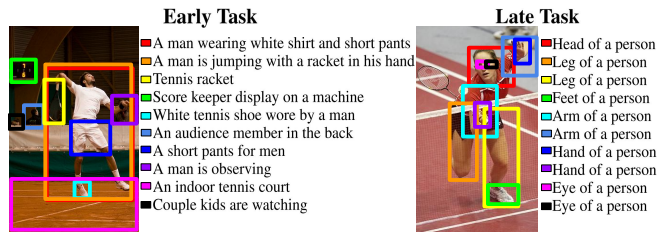

Figure 6: Image descriptions written by a satisficing worker on a task completed near the start of their lifetime (left) and their last completed task (right). Despite the images being visually similar, the phrases submitted in the last task are much less diverse than the ones submitted in the earlier task.

Overall findings. From an analysis of the three datasets, we found that fatigue effects are not significantly visible and that severe satisficing behavior only affects a very small proportion of workers. On average, workers maintain a similar quality of work over time, but also get more efficient as they gain experience with the task.

EXPERIMENT: WHY ARE WORKERS CONSISTENT?

Examining the image descriptions, question answering, and the verification datasets, we find that worker's performance on a given microtask remains consistent — even if they do the task for multiple months. However, mere observation of this consistency does not give true insight into the reasons for its existence. Thus, we seek to answer the following question: do crowd workers satisfice according to the minimum quality necessary to get paid, or are they consistent regardless of this minimum quality?

To answer this question, we perform an experiment where we vary the quality threshold of work and the threshold's visibility. If workers are stable, we would expect them to either submit work that is above or below the threshold, irrespective of what the threshold is. However, if workers satisfice according to the minimum quality expected, they would adjust the quality of their work based on set threshold [28, 51].

If workers indeed satisfice, then the knowledge of this threshold and their own performance should make it easier to perfect satisficing strategies. Therefore, to adequately study the effects of satisficing, we vary the visibility of the threshold to workers as well. In one condition, we display workers' current quality scores and the minimum quality score to be accepted, while the other condition only displays whether submitted work was accepted or rejected. To sum up, we vary the threshold and the transparency of this threshold to determine how crowd workers react to the same task, but with different acceptability criteria.

Task

To study why workers are consistent, we designed a task where workers are presented with a series of randomly ordered 58 binary verification questions. Each verification requires them to determine if an image description and its associated image part are correct. For example, in Figure 8,
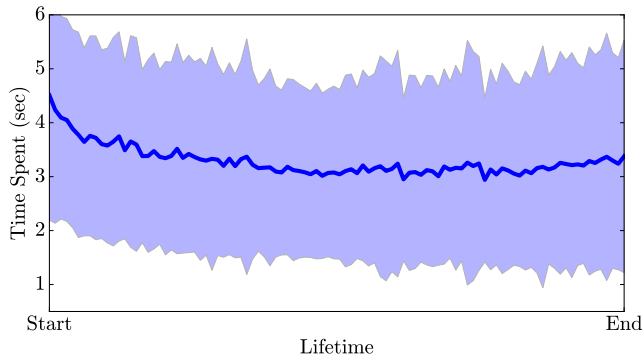
Figure 7: As workers gain familiarity with a task, they become faster. Verification tasks speed up by 25% from novice to experienced workers.

the zebras have stripes

Figure 8: An example binary verification task where workers are asked to determine if the phrase "the zebras have stripes" is a factually correct description of the image region surrounded within the red box. There were 58 verification questions in each task.

workers must decide if "the zebras have stripes" is a good description of a particular part of the image. They are asked to base their response based solely on the content of the image and the semantics of the sentence. To keep the task simple, we asked workers to ignore whether the box was perfectly surrounding the image area being described. The tasks were priced such that workers could earn $6 per hour and were available to workers with a 95% approval rating and who lived in the United States. Each task took approximately 4 minutes to complete and were given 2.5 hours to complete the task to ensure workers were not pressured for time.

We placed 3 attention checks in each task. Attention checks are gold-standard verification questions whose answers were already known. Attention checks were randomly placed within the series of 58 verifications to gauge how well a worker performed on the given task. To avoid workers from incorrectly marking an attention check due to subjective interpretation of the description, we manually marked these attention checks correct or incorrect. Examples of attention checks are shown in Figure 9. Incorrect attention checks were completely mismatched from their image; for example "A very tall sailboat" was used as a incorrect attention check matched to an image of a lady wearing a white dress. We created a total of 11,290 unique attention checks to prevent workers from simply memorizing the attention checks.

Even though these attention checks were designed to be obviously correct or incorrect, we ensured that we do not reject a worker's submission based off a single, careless mistake or an unexpected ambiguous attention check. After completing a task, each worker's submission is immediately accepted or rejected based on a rating, which is calculated as the percentage of the last 30 attention checks correctly labeled. If a worker's rating falls below the threshold of acceptable quality, their task is rejected. However, to ensure fair payment, even if a worker's rating is below the threshold, their task is accepted if they get all the attention checks in the current task correct. This enables workers who are below the threshold to perform carefully and improve their rating as they continue to do more tasks.

Experiment Setup

Our goal is to vary the acceptance threshold to see how it impacts worker quality over time. We performed a between-subjects 2×2 study where we varied *threshold* and *transparency*. We ran an initial study with a different set of 100 workers to estimate how people performed on this verification task. We found that workers get a mean accuracy of $94 \pm 10\%$ with a median accuracy of 95.5%. We chose the thresholds such that the high threshold condition asked workers to perform above the median and the low threshold was below $2\times$ the standard deviation, allowing workers plenty of room to make mistakes. The high threshold factor level was set at 96% while the low threshold factor level was set at 70%. Workers in the high threshold level could only incorrectly label at most 1 out of 30 of the previous attention checks to avoid rejection, while workers in the low threshold level could error on 8 out of the past 30 attention checks.

We used two levels of transparency: high and low. In the high factor level, workers were able to see their current rating at the beginning of every task and were also alerted of how their rating changed after submitting each task. Meanwhile, in the low factor level, workers did not see their rating, nor did they know what their assigned threshold was.

We recruited workers from AMT for the study and randomized them between conditions. We measured workers' accuracy and total number of completed tasks under these four conditions.

Data Collected

By the end of the study, 1,134 workers completed 11,666 tasks. In total, 676,628 binary verification questions were answered, of which 34,998 were attention checks. Table 2 shows the breakdown of the number of workers who completed at least 1 task. Not all workers who accepted tasks completed them. In the high threshold condition, 106 and 116 workers did not complete any tasks in the high and low transparency conditions respectively. Similarly, 137 and 138
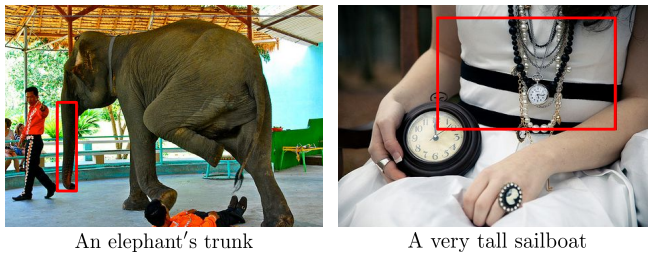
An elephant's trunk A very tall sailboat

Figure 9: Examples of attention checks placed in our binary verification tasks. Each attention check was designed such that they were easily identified as correct or incorrect. "An elephant's trunk" (left) is a positive attention check while "A very tall sailboat" (right) is an incorrect attention check. We rated worker's quality by measuring how well they performed on these attention checks.

workers did not complete tasks in the low threshold. This resulted in 29 and 28 more people in the low threshold that completed tasks. Workers completed on average a total of 576 verifications each.

Results

On average, the accuracy of the work submitted by workers in all four conditions remained consistent (Figure 10). In the low threshold factor level, workers averaged a rating of $93.6 \pm 4.7\%$ and 93.3 ± 5.8 in the high and low transparency factor levels. Meanwhile, when the threshold was high, workers in the low transparency factor level averaged $94.2 \pm 4.4\%$ while the workers in the high transparency factor level averaged $95.2 \pm 4.0\%$. Overall, the high transparency factor level had a smaller standard deviation throughout the course of workers' lifetimes. We conducted a two-way ANOVA using the two factors as independent variables on all workers who performed more than 5 tasks. The ANOVA found that there was no significant effect of threshold (F(1, 665)=0.55, p=0.45) or transparency (F(1, 665)=2.29, p=0.13), and no interaction effect (F(1, 665)=0.24, p=0.62). Thus, worker accuracy was unaffected by the accuracy requirement of the task.

Unlike accuracy, worker retention was influenced by our manipulation. By the 50th task, less than 10% of the initial worker population continued to complete tasks. This result is consistent with our observations with the Visual Genome datasets and from previous literature that explains that a small percentage of workers complete most of the crowdsourced work [36]. We also observe that workers in the high threshold and high transparency condition have a sharper dropout rate in the beginning. To measure the effects of the four conditions on dropout, we analyzed the logarithm of the number of tasks completed per condition using an ANOVA. (Log-transforming the data ensured that it was normally distributed and thus amenable to ANOVA analysis.) The ANOVA found that there was a significant effect of transparency (F(1, 665)=279.87, p<0.001) and threshold (F(1, 665)=88.61, p<0.001), and also a significant interaction effect (F(1, 665)=76.23, p<0.001). A post hoc Tukey test [54] showed that the (1) high transparency and high threshold con-

| Threshold | High: 96 | | Low: 70 | |
Transparency	High	Low	High	Low
# workers with 0 tasks	106	116	137	138
# workers with >1 tasks	267	267	300	300
# tasks	2702	2630	3209	3125
# verifications	5076	7890	9627	9375

Table 2: Data collected from the verification experiment. A total of $1,134$ workers were divided up into four conditions, with a high or low threshold and transparency.

dition had significantly less retention than the (2) low transparency and high threshold condition ($p < .05$).

Discussion

Workers are consistent in their quality level. With this experiment, we are now ready to answer whether workers are consistent or satisficing to an acceptance threshold. Given that workers' quality was consistent throughout all the four conditions, evidence suggests that workers were consistent, regardless of the threshold at which requesters accept their work. In the low threshold and high transparency condition, workers are aware that their work will be accepted if their rating is above 70%, and still perform with an average rating of 94%. Workers are risk-averse, and seek to avoid harms to their acceptance rate [41]. Once they find a strategy that allows their work to be accepted, they stick to that strategy throughout their lifetime [42]. This result is consistent with the earlier observational data analysis.

Workers minimize risk by opting out of tasks above their natural accuracy level. If workers do not adjust their quality level in response to task difficulty, the only other possibility is that workers self-select out of tasks they cannot complete effectively. Our data supports this hypothesis: workers in the high transparency and high threshold condition did statistically fewer tasks on average. The workers self-selected out of the task when they had a higher chance of rejection. Out of 267 workers in the high transparency and high threshold condition, 200 workers workers stopped working once their rating dropped below the 96% threshold. Meanwhile, in the high transparency and low threshold condition, out of the 300 workers who completed our tasks, almost all of them continued working even if their rating dropped below the 70% threshold, often bringing their rating back up to above 96%.

This study illustrates that workers are consistent over very long periods of hundreds of tasks. They quickly develop a strategy to complete the task within the first few tasks and stick with it throughout their lifetime. If their work is approved, they continue to complete the task using the same strategy. If their strategy begins to fail, instead of adapting, they self-select themselves out of the task.

PREDICTING FROM SMALL GLIMPSES

The longitudinal analysis in the first section and the experimental analysis in the second section found that crowd worker quality remains consistent regardless of how many tasks the worker completes and regardless of the required acceptance criteria. Bolstered by this result, this section demonstrates

(a) Threshold: low (70) and transparency: low, accuracy: 93.3 ± 5.8%

(b) Threshold: low (70) and transparency: high, accuracy: 93.6 ± 4.7%

(c) Threshold: high (96) and transparency: low, accuracy: 94.2 ± 4.4%

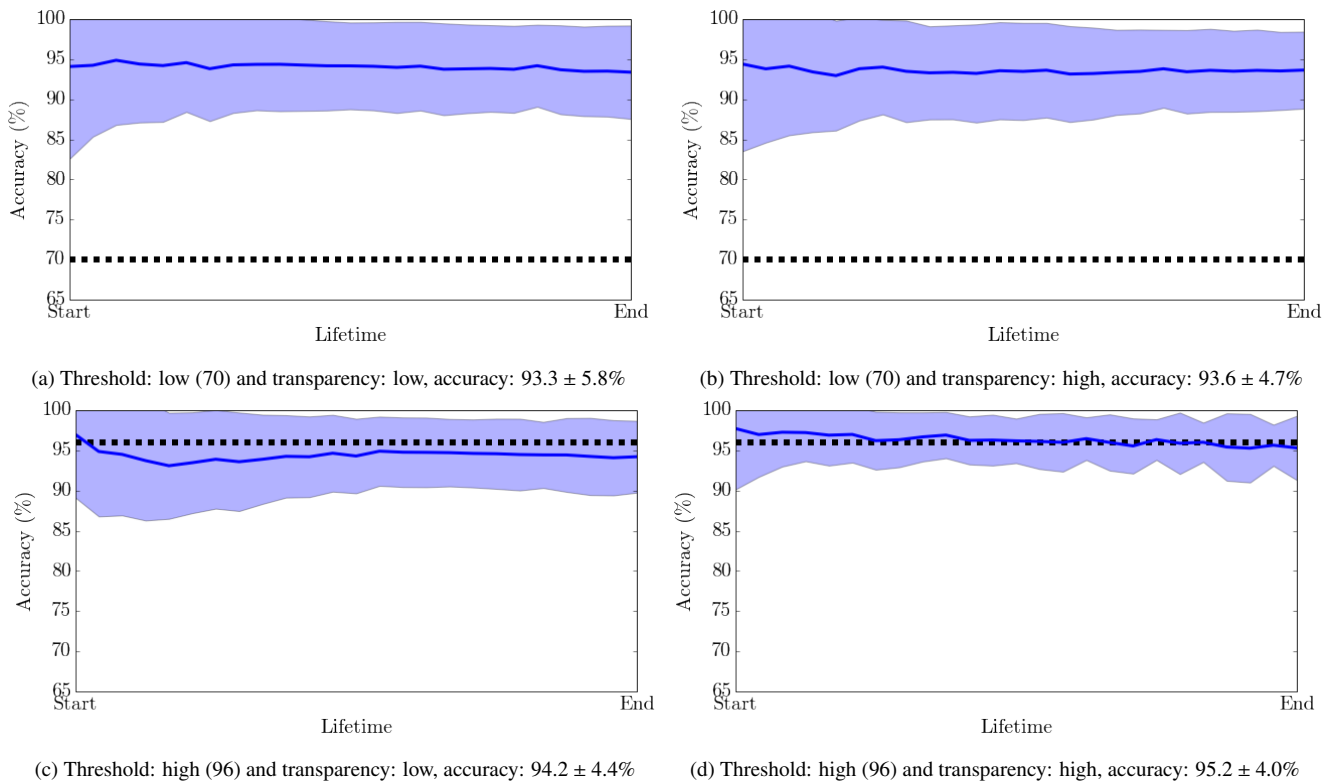(d) Threshold: high (96) and transparency: high, accuracy: 95.2 ± 4.0%

Figure 10: Worker accuracy was unaffected by the threshold level and by the visibility of the threshold. The dotted black line indicates the threshold that the workers were supposed to adhere to.

the efficacy of predicting a worker's future quality by observing a small glimpse of their initial work. The ability to predict a worker's quality on future tasks can help requesters identify good workers and improve the quality of data collected.

Experimental Setup

To create a prediction model, we use the question answering dataset. Our aim is to predict a worker's quality on the task towards the end of their lifetime. Since workers' individual quality on every single task can be noisy, we estimate a worker's future quality as the average of their accuracy on the last 10% of their tasks in their lifetime.

We allow our model to use between the first 5 and the first 200 tasks completed by a worker to estimate their future quality. Therefore, we only test our model on workers who have completed at least 200 tasks. As a baseline, we calculate *the average of all workers'* performances on their last n tasks. We use this value as our guess for each individual worker's future quality. This model assumes a worker does as well as the average worker does on their final tasks.

Besides the baseline, we use two separate models to estimate a worker's future quality: *average* and *sigmoid* models. The average model is a simple model that uses the average of the worker's n tasks as the estimate for all future quality predictions. For example, if a worker averages 90% accuracy on their first five tasks, the average model would predict that the worker will continue to perform at a 90% accuracy. How-

ever, if the worker's quality on their last 10% of tasks is 85%, then the prediction error would be 5%. The *sigmoid* model attempts to represent a worker's quality as a sigmoid curve with 4 parameters to adjust for the offset of the curve. We use a sigmoid model because we find that many workers display a very brief learning curve over the first few tasks and remain consistent thereafter. The initial adjustment and future consistency closely resembles a sigmoid curve.

Results

The *average of all workers'* accuracy is 87.8%. Using this value as a baseline model for quality yields an error of 6.9%. We plot the error of the baseline as a dotted line in Figure 11. The *average* model performs better: even for only a glimpse of $n = 5$ tasks, its error is 3.4%. After seeing a worker's first $n = 200$ tasks, the model gets slightly better and has a prediction error of 3.0%. The *sigmoid* model outperforms the baseline but underperforms the average model and achieves an error of 3.7% for $n = 5$. As the model incorporates more tasks, it becomes the most accurate, managing an error rate of 1.4% after seeing $n = 200$ tasks. Furthermore, the model's standard deviation of the error also decreases from 3.4% to 0.7% as n increases.

Discussion

Even a glimpse of five tasks can predict a worker's future quality. Since workers are consistent over time, both the *average* and the *sigmoid* models are able to model workers' qual-
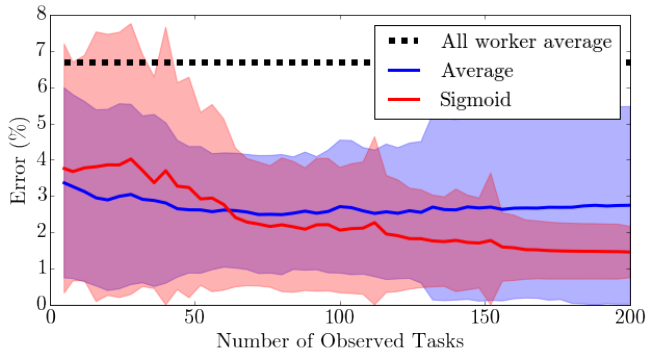
Figure 11: It is possible to model a worker's future quality by observing only a small glimpse of their initial work. Our *all workers' average* baseline assumes that all workers perform similarly and manages an error in individual worker quality prediction of 6.9%. Meanwhile, by just observing the first 5 tasks, our *average* and *sigmoid* models achieve 3.4% and 3.7% prediction error respectively. As we observe more hits, the *sigmoid* model is able to represent workers better than the *average* model.

ity with very little error. When workers initially start doing work, a simple *average* model is a good choice for a model to estimate how well the worker might perform in the future. However, as the worker completes more and more tasks, the *sigmoid* model is able to capture the initial adjustment a worker makes when starting a task. By utilizing such models, requesters can estimate which workers are most likely to produce good work and can easily qualify good workers for long-term work.

IMPLICATIONS FOR CROWDSOURCING

Encouraging diversity. The consistent accuracy and constant diversity of worker output over time makes sense from a practical perspective: workers are often acclimating to a certain style of completing work [41] and often adopt a particular strategy to get paid. However, this formulaic approach might run counter to a requester's desire to have richly diverse responses. Checks to increase diversity, such as enforcing a high threshold for diversity, should be employed without fear of worker quality as we have observed that quality does not significantly change with varying acceptance thresholds. Therefore, designing tasks that promote diversity without effecting the annotation quality is a ripe area for future research.

Worker retention. Additional experience affects completion speeds but does not translate to higher quality data. Much work has been done to retain workers [9, 11, 35], but, as shown, retention does not equate to increases in worker quality — just more work completed. Further work should be conducted to not only retain a worker pool, but also examine methods of identifying good workers [22] and more direct interventions for training poorly performing workers [13, 23].

Additionally, other studies have shown that the motivation of workers is the predominant factor in the development of fatigue, rather than the total time worked [1]. Although crowd-

sourcing can be intrinsically motivated [55], the microtask paradigm found in the majority of crowdsourcing tasks favors a structure that is efficient [15, 27] for workers rather than being interesting for them [7]. Future tasks should consider building continuity in their workflow design for both individual worker efficiency [32] and overall throughput and retention [9].

Person-centric versus process-centric crowdsourcing. Attaining high quality judgments from crowd workers is often seen as a challenge [46, 50, 52]. This challenge has catalyzed studies suggesting quality control measures that address the problem of noisy or low quality work [14, 23, 40]. Many of these investigations study various quality-control measures as standalone intervention strategies. While we explored process-centric measures like varying the acceptance or transparency threshold, previous work has experimented with varying financial incentives [42]. All the results support the conclusion that process-centric strategies do not have significant difference in the quality of work submitted. While we agree that such process focused strategies are important to explore, our data reinforces that person-centric strategies (like utilizing worker approval ratings or worker quality on initial tasks) may be more effective [42, 48] because they identify a worker's (consistent) quality early on.

Limitations. Our analysis solely focuses on data labeling microtasks, and we have not yet studied whether our findings translate over to more complex tasks, such as designing an advertisement or editing an essay [3, 25]. Furthermore, we focus on weeks-to-months crowd worker behavior based on datasets collected over a few months, but there exist some crowdsourcing tasks [5] that have persisted far longer than our study. Thus, we leave the analysis of crowd worker behavior spanning multiple years to future work.

CONCLUSION

Microtask crowdsourcing is rapidly being adopted to generate large datasets with millions of labels. Under the Pareto principle, a small minority of workers complete a great majority of the work. In this paper, we studied how the quality of workers' submissions change over extended periods of time as they complete thousands of tasks. Contrary to previous literature on fatigue and satisficing, we found that workers are extremely consistent throughout their lifetime of submitting work. They adopt a particular strategy for completing tasks and continue to use that strategy without change. To understand how workers settle upon their strategy, we conducted an experiment where we vary the required quality for large crowdsourcing tasks. We found that workers do not satisfice and consistently perform at their usual quality level. If their natural quality level is below the acceptance threshold, workers tend to opt out from completing further tasks. Due to this consistency, we demonstrated that brief glimpses of just the first five tasks can predict a worker's long-term quality. We argue that such consistent worker behavior must be utilized to develop new crowdsourcing strategies that find good workers and collect unvarying high quality annotations.

REFERENCES

1. Debby GJ Beckers, Dimitri van der Linden, Peter GW Smulders, Michiel AJ Kompier, Marc JPM van Veldhoven, and Nico W van Yperen. 2004. Working overtime hours: relations with fatigue, work motivation, and the quality of work. *Journal of Occupational and Environmental Medicine* 46, 12 (2004), 1282–1289.

2. Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 33–42.

3. Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

4. Maarten AS Boksem and Mattie Tops. 2008. Mental fatigue: costs and benefits. *Brain research reviews* 59, 1 (2008), 125–139.

5. Jonathan Warren Brelig and Jared Morton Schrieber. 2013. System and method for automated retail product accounting. (Jan. 30 2013). US Patent App. 13/754,664.

6. Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 286–295.

7. Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.

8. Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. Risks and rewards of crowdsourcing marketplaces. In *Handbook of human computation*. Springer, 377–392.

9. Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with Micro-Diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 628–638.

10. Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267, 5199 (1995), 843.

11. Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

12. Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 617–617.

13. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.

14. Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2399–2402.

15. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 203–212.

16. Robert A. Henning, Steven L Sauter, Gavriel Salvendy, and Edward F Krieg Jr. 1989. Microbreak length, performance, and stress in a data entry task. *Ergonomics* 32, 7 (1989), 855–864.

17. Paul Hitlin. 2016. Research in the Crowdsourcing Age, a Case Study. (July 2016). **http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/**

18. Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 419–429.

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

20. Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620.

21. Adam Kapelner and Dana Chandler. 2010. Preventing Satisficing in online surveys. In *Proceedings of CrowdConf*.

22. David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, 284–291.

23. Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.

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

25. Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

26. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016b. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* (2016).
http://arxiv.org/abs/1602.07332

27. Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016a. Embracing Error to Enable Rapid Crowdsourcing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3167–3179.

28. Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5, 3 (1991), 213–236.

29. Gerald P Krueger. 1989. Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress* 3, 2 (1989), 129–141.

30. Raymond Kuhn and Erik Neveu. 2013. *Political journalism: New challenges, new practices*. Routledge.

31. Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 23–34.

32. Walter S Lasecki, Adam Marcus, Jeffrey M Rzeszotarski, and Jeffrey P Bigham. 2014. *Using microtask continuity to improve crowdsourcing*. Technical Report. Tech. rep.

33. Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.

34. Walter S Lasecki, Jeffrey M Rzeszotarski, Adam Marcus, and Jeffrey P Bigham. 2015. The Effects of Sequence and Delay on Crowd Work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1375–1378.

35. Edith Law, Ming Yin, Kevin Chen Joslin Goh, Michael Terry, and Krzysztof Z Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4098–4110.

36. Greg Little. 2009. How many turkers are there. (Dec 2009).
http://groups.csail.mit.edu/uid/deneme/?p=502

37. Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of the NAACL-HLT 2016*. Association for Computational Linguistics, 897âĂŞ–906.

38. Adam Marcus and Aditya Parameswaran. 2015. *Crowdsourced data management: industry and academic perspectives*. Foundations and Trends in Databases.

39. David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 224–235.

40. Winter Mason and Duncan J Watts. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11, 2 (2010), 100–108.

41. Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and WorkersâĂŹ Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2271–2282.

42. Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing Person-and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

43. Allen Newell and Paul S Rosenbloom. 1981. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition* 1 (1981), 1–55.

44. Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.

45. Layne P Perelli. 1980. *Fatigue Stressors in Simulated Long-duration Flight. Effects on Performance, Information Processing, Subjective Fatigue, and Physiological Cost*. Technical Report. DTIC Document.

46. Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.

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

48. Jeffrey Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 55–62.

49. Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and others. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1621–1630.

50. Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 275–284.

51. Herbert A Simon. 1972. Theories of bounded rationality. *Decision and organization* 1, 1 (1972), 161–176.

52. Alexander Sorokin and David Forsyth. 2008. Utility data annotation with amazon mechanical turk. *Urbana* 51, 61 (2008), 820.

53. Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817* (2015).

54. John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.

55. Luis Von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (2006), 92–94.

56. Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*. 2035–2043.

57. Stanley Wyatt, James N Langdon, and others. 1937. Fatigue and Boredom in Repetitive Work. *Industrial Health Research Board Report. Medical Research Council* 77 (1937).