An Exploratory Study of the Impact of Task Selection Strategies on Worker Performance in Crowdsourcing Microtasks

Huda Banuqitah[1,2], Mark Dunlop[1], Maysoon Abulkhair[2], Sotirios Terzis[1]
[1]University of Strathclyde
[2]King Abdulaziz University
(huda.banuqitah, mark.dunlop, sotirios.terzis)@strath.ac.uk, mabualkhair@kau.edu.sa

Abstract

In microtask crowdsourcing systems like Amazon Mechanical Turk (AMT) and Appen Figure Eight, workers often employ task selection strategies, completing sequences of tasks to maximize earnings. While previous literature has explored the effects of sequential tasks with varying complexities of the same type, there is a lack of knowledge on the consequences when multiple types of tasks with similar levels of difficulty are performed. This study examines the impact of sequences of three frequently employed task types, namely image classification, text classification, and surveys, on workers' engagement, accuracy, and perceived workloads. In addition, we analyze the influence of workers' personality qualities on their strategy for selecting tasks. Our study, which involved 558 participants using AMT, found that engaging in sequences of distinct task types had a detrimental effect on classification task engagement and accuracy. It also increases the perceived task load and the worker's frustration. Nevertheless, the precise order of tasks does not significantly impact these results. Moreover, we showed a slight association between personality traits and the workers' selection strategy for the tasks. The results offered valuable knowledge for designing an efficient and inclusive crowdsourcing platform.

Introduction

Crowdsourcing has transformed standard employment patterns by allowing activities to be delegated to a wide range of individuals through open requests(Gegenhuber, et al. 2022). This approach offers notable benefits, including rapid reaction times and the ability to tap into a large pool of workers (Lenart-Gansiniec 2021), making it crucial for activities such as annotation and knowledge acquisition in AI research(Dong, et al. 2023). However, there are still difficulties in maintaining worker concentration and guaranteeing a varied group of participants (Hornuf and Vrankar 2022). In CS platforms, task consumption often involves a self-selection process where workers choose projects based on their eligibility and preferences. Previous research has shown that worker performance improves with experience due to learning effects (Difallah, et al. 2014; Gadiraju and Dietze 2017). However, unsupervised environments can incentivize workers to prioritize throughput over contribution quality, potentially compromising the integrity of the results (Chandler, et al. 2014).

Worker engagement and performance may be influenced by various factors including task complexity, perceived workload, and monetary incentives (Shi, et al. 2021). Personality traits can also play a role, with self-efficacy regulating the relationship between traits like Conscientiousness and engagement (Shi, et al. 2021).

Understanding worker engagement is crucial for the development of CS communities (Troll, et al. 2016), necessitating exploration of worker behavior across different task types. While previous studies have explored multitasking preferences (Lascau, et al. 2019), gaps remain in understanding the underlying factors driving such behaviors. Addressing these gaps is essential for designing more effective crowdsourcing platforms and improving task outcomes.

This study addresses these gaps by employing both quantitative and qualitative approaches to examine workers' contributions across multiple task types on Amazon Mechanical Turk. Understanding how these choices affect their perceived effort and productivity could help to improve work creation space and efficiency.

Specifically, we investigate the following research questions:

RQ1: How does multiple-tasking behavior affect workers engagement, performance and their perceived workload in CS platforms?

RQ2: Do the workers' personality traits correlate to their accuracy and engagement in CS different tasks?

RQ3: Does the order of the selected tasks affect the accuracy and engagement of the CS workers?

The significance behind (RQ1) is to identify the driving force or incentive behind a certain behavior. Crowd workers frequently perform multiple tasks, since they undertake numerous activities simultaneously to optimize their earnings.

Nevertheless, the consequences of this behavior on their level of involvement, precision, and perceived amount of effort are not thoroughly comprehended. Through the examination of the impact of multitasking on these outcomes, we may discern possible disadvantages and advantages. This feedback can assist in the creation of jobs that uphold exceptional contributions and sustain worker motivation.

The rationale behind (RQ2) is Personality traits have an impact on how individuals' approach and carry out activities. gaining a knowledge of these characteristics in CS platforms can aid in forecasting worker effectiveness and involvement. The analysis of personality traits and their impact on job performance can devise techniques to align workers with activities that align with their individual strengths, so potentially enhancing overall efficiency and pleasure.

The importance of research question 3 (RQ3) is to understand the underlying motivational factors. The system in which employees are identified and selected for activities may affect their engagement and performance due to their cognitive labor and variability between tasks outcomes. Examining the potential influence of task order might offer valuable insights on how to arrange task sequences to maximize outcomes. This is important for platforms that provide a range of tasks categories, since this can inform the creation of more efficient task processes.

To conclude, comprehending these aspects is important for designing CS platforms that not only improve workers satisfaction and performance but also guarantee the caliber and dependability of the resulted data. Our aim is to enhance the creation of more streamlined and productive crowdsourcing systems by investigating these research inquiries.

The paper proceeds with a review of related research, followed by a description of our study methodology, results, and discussion. Finally, we conclude by summarizing our findings and outlining directions for future research.

Related Work

Some studies investigate the crowd worker's nature, their profiles, and their engagement. Other studies determine some behavior approaches of doing multiple tasks of different complexity in crowdsourcing platforms. Our study aims to understand their tendency of working on multiple different task types of same difficulty level, and the effect of such behavior over their engagement, perceived tasks load and performance. Moreover, determine the effect of workers' personality on this tendency.

Multiple Tasks Selection in Crowdsourcing

One of the vital causes of inattentiveness in workers of the crowd might be multiple tasking behavior. This behavior is expected in most traditional workplace settings (Saravanos, et al. 2021). Since crowdsourcing workers generally do not have a boss or experimenter to watch over them, some workers may integrate multiple activities with their work. Many studies have published different task types of different complexity in the platforms (Qiu, et al. 2020) in order to measure the worker engagement on each particular task, their result showed different engagement scores due to different of task complexity. Other researchers used different complexities of the tasks to measure their retention and performance by integrating small entertainment diversion and their result showed improvement of the workers output on different complexity(Dai, et al. 2015). Another research objective of exploring different task types is to understand the development of the workers and their interaction to the tasks in order to enhance the dynamic of marketplace (Jain, et al. 2017). To this end, research is still needed to address and investigate the workers behavior who prefer doing multiple different task types in the platform and the effect of that on their performance and their perceived task loads in addition to their engagement to such behavior.

Although they can be carried out separately, in actuality people frequently complete them in a chain, one after the other (Cai, et al. 2016). Task chaining or ordering has a significant impact on worker performance, according to other studies(Cai, et al. 2016; Newell and Ruths 2016). This is due to the fact that changes in mental processes during the transition between two related cognitive tasks can be observed.

According to related psychological research, when workers transition between jobs as opposed to performing a single activity, their ability to contribute tends to decline and they become more prone to making mistakes. This can be explained by the physical and psychological factors being reconfigured to suit the new task at hand (Monsell 2003; Wylie and Allport 2000). Those studies consider only the same microtask type ordering, while the study report here explores the effect of selecting multiple different task types of the same complexity on workers' engagement, perceived tasks load and performance and the role of their personality on this.

Worker Engagement and Perceived Tasks Loads

In online crowdsourcing contexts, interactivity and co-creative user experiences lead to a subjective psychological condition known as "solver engagement" (Huang, et al. 2019; Ihl, et al. 2020). Engagement essentially refers to a person's level of participation in a variety of activities (Troll, et al. 2019). Participants in online crowdsourcing groups build deep psychological links that encourage solver participation behavior and a sustained engagement to the community (Yang, et al. 2019). Its effects on community loyalty (Li, et al. 2020), value co-creation (Piyathasanan, et al. 2018), and

other outcomes have shown the significance of solver participation.

User retention and engagement are two primary indicators that are linked positively to workers' lifetime value. These two indices, on the other hand, correspond to different aspects of participants' behavior (Webster and Ahuja 2006). The retention of the workers focuses on the length of users' interest toward a product, or their loyalty, whereas the engagement of the workers stresses the intensity of user activity (O'Brien, et al. 2018). A wide extent of motivation components has been outlined and considered in crowdsourcing settings for worker engagement and retention. Money-related rewards are one of the significant regular motivating forces to utilize in a paid crowdsourcing environment to get a fast response from the workers. Different analysts have conducted many tests to get the impacts of money-related motivations on crowd work amount and quality (Difallah, et al. 2014; Ho, et al. 2015; Horton and Chilton 2010; Mason and Watts 2009). Crowd recruitment is always a challenge because workers with unique knowledge and abilities are needed for various tasks (Bhatti, et al. 2020). As a result, attracting, motivating, and retaining the right users is a fundamental challenge (Doan, et al. 2011). More recently, more attention has been paid to studying the workers' characteristics and how they correlate to the quality of their solutions. The perceived tasks load of workers doing individual task type have been studied in order to measure the effect of the task on the workers performance(Qiu, et al. 2021). However, linking the performance and engagement of workers doing multiple different tasks with different order selection within short time with their perceived tasks loads in CS has not been explored enough in the online crowdsourcing community.

Worker Personality Traits

According to (McCrae and Costa 2003), personality is defined as a person's consistent pattern of ideas, feelings, and behaviors. These behaviors are impacted by biological and environmental variables, such as life experiences, which are generally shared across cultures. Over the past few decades, research on personality has increased in a variety of sectors. Personality psychology, a field in the social sciences that has been active for decades, is a result of the study of personality (Adamopoulos, et al. 2018). Many research studies have looked at the concept of personality to identify the underlying variables that influence it. As a result, various taxonomies of personality traits have been developed, and psychologists have developed coherent theories about the nature of personality and its ideas.

More recently, more attention has been paid to studying the workers' characteristics and how they correlate to the quality of their solutions. Personality traits have appeared to relate to client behavior within the setting of e-commerce (Huang and Yang 2010), social media (Gosling, et al. 2007), and web browsing and searching behaviors (Bachrach, et al. 2012), their impact on worker behavior in crowdsourcing platforms remains low.

According to the body of literature, personality profiles which are defined as a person's qualities that remain the same throughout their life (Leonidou, et al. 2019) are a crucial aspect that influences how much time a person spends in online communities (Sulaiman, et al. 2018). According to theory of (Vander Shee, et al. 2020), personality is a very important component in shaping behavior in online communities. Empirical data constantly demonstrate that individual variations have a direct impact on engagement habits, which supports this line of thinking (Schäper, et al. 2021). However, while existing literature has hinted at the importance of personality traits in online communities (Leonidou, et al. 2019); (Vander Shee, et al. 2020), there is a lack of research exploring their specific role in shaping worker dynamics in crowdsourcing environments.

The Big-Five personality characteristic model (Rammstedt and John 2007), is now the most influential theory on personality (Personality traits dimensions were measured by used a standard 10 question test which are, the Openness to be inventive, autonomous, and interested in variation (high score) vs. practical, conforming, and interested in routine (low score) is referred to as Openness (low score). Conscientiousness is the inclination to be self-disciplined, act responsibly, be organized, cautious, and disciplined (high score) instead of being disorganized, thoughtless, and impulsive (low score). The inclination to be gregarious, fun-loving, and affectionate (high score) vs. retiring, solemn, and quiet (low score) is known as Extraversion. The ability to be empathetic, cooperative, trusting, and helpful is known as Agreeableness. Neuroticism is the inclination to be calm, secure, and self-satisfied (low score) vs. nervous, insecure, emotionally unstable, and self-pitying (high score).

Moreover, although that other studies have demonstrated that personality traits are crucial in establishing personal views and motivating personal activities in online crowdsourcing environments (Crone and Williams 2017; Mourelatos, et al. 2022), little research has been done to explore the influence of personality traits on workers' engagement and performance in multiple tasking behavior crowdsourcing platforms. (Kazai, et al. 2011) investigated worker types and personality traits in crowdsourcing relevance labels, shedding light on the relationship between personality traits and worker behavior in crowdsourcing environments. Further exploration of this relationship can contribute to a deeper understanding of worker dynamics in crowdsourcing platforms.

Other research showed that individuals who are less emotionally stable (as opposed to neurotic) are more inclined to choose crowdsourcing platforms for factual tasks (Zhang, et

al. 2017). It is fair to anticipate certain connections between workers' actions, such as their participation in different online crowdsourcing task types, and unique personalities.

As a result, in this study we are synthesizing these diverse perspectives and building upon prior research, our study aims to provide a comprehensive understanding of the role of personality traits in shaping worker behavior and tasks outcomes in crowdsourcing environments. Through this exploration, we seek to offer valuable insights for platform designers, task requesters, and researchers aiming to optimize worker engagement and performance in crowdsourcing platforms.

Experiment Setup

To answer the research questions and explore the crowdsourcing platforms and the worker features with their multiple tasking and selection ordering behavior on different task types, we ran our experiment on Amazon Mechanical Turk (AMT). The tasks employed in our experiments were text classification, image categorization and survey. To have similar task complexity between different task types, we followed the task complexity dimensions identified by (Aipe and Gadiraju 2018). These dimensions included the task completion time, the task batch size, and the task completion reward, those will be detailed later in the narrative. The tasks were implemented using AMT templates to have the same design style.

To support fair treatment of crowd-workers, we paid workers based on estimated time at typical minimum US wage of $8.50/hr (Hara, et al. 2018). The tasks were created to have an estimated medium level of difficulty, and the workloads were evaluated to make sure it didn't result in excessive stress. By employing metrics such as NASA-TLX and the Big Five Personality traits, we were able to evaluate perceived workload and personality features, ensuring that our findings are grounded in dependable facts. The study was conducted under departmental ethics approval.

The datasets used in this study were annotated by experts (Imran, et al. 2016) and (Khosla, et al. 2011), allowing us to compare worker labels to the ground truth standard to determine differences in performance.

As stated earlier, the study focused on the type of categorization work (text and images) as well as different classification levels (nine and ten classes).

Our first task set was based around a tweet dataset (Dataset 1) that consisted of tweets collected during a crisis or disaster (Imran, et al. 2016). This dataset contained roughly 52 million disaster-related texts and was collected from 19 different crises between 2013 and 2015.

The scheme of annotation was used to categorize each content of the tweet into one of nine categories: Injured or dead people; Missing, trapped, or found people; Displaced people and evacuations; Infrastructure and utilities damage buildings, roads, or interrupted services; Donation needs or offers or volunteering services; Caution and advice; Sympathy and emotional support; Other useful information; and finally, Not related or irrelevant.

Our second dataset was the Stanford Dogs image collection containing 120 annotated photographs (Dataset 2) (Khosla, et al. 2011), which was created utilizing images and annotation from ImageNet. It was initially gathered to aid with fine-grain image categorization. We choose ten categories from the dog breeds to avoid long tasks resulting and task abandonment which are: Labrador, Golden Retriever, Yorkshire Terrier, German Shepherd, French Bulldog, Standard Poodle, Beagle, Doberman, Boxer and Pug.

The third category of CS task we wished to investigate was surveys around learning experiences over the 2019-2022 pandemic with related global lock-down restrictions.

To assess the design and the difficulty levels of those tasks based on tasks' time completion and to give us a sense of the probable response pattern, a pilot study on our university colleagues was conducted of 10 sample sizes. The result showed that the average time of the workers to complete a task with same batch size (10) and was used to estimate the reward per minute (Hara, et al. 2018)in the main study.

In our experiment, we published one batch of HITs for each specific task type we mentioned earlier, and the workers were free to choose them - from the interface of AMT' workers dashboard with the following titles:

Tweet classification: workers are asked to read ten tweet sentences (size of batch) and then select the proper label for each tweet.

Images classification: workers were shown ten different dog images (size of batch) and asked to select the proper breeds.

Survey tasks: the workers will answer a multiple-choice question regarding the E-Learning experience during Covid-19.

The workers could select any task type they were interested in performing and to avoid memory bias, we allow the workers to participate in only one time of each tasks type. After finishing the task, the workers were required to complete a short post-task Qualtrics survey and copy the code given back to AMT to encourage completion of the survey. We set up the number of assignments in each batch (number of workers assigned for each task type) to be 250 workers. we pay every worker $1.50 for each 10 minutes task.

Evaluation Metrics

We evaluated use of our system on three grounds:

Accuracy. Workers' predicted labels were compared to the labels in gold standard, this measure was utilized to calculate their score. Accuracy is expressed as the measure of correct predictions rate to the errors rate, such accuracy is equal to 1.0 - error rate (Sammut, et al. 2011).

Perceived Workload. The workers were asked to fill in a questionnaire on the Qualtrics platform about some demographic information at the end of each task and their perceived workload using a standard six subjective questions of NASA-TLX (Hart and Staveland 1988). This test also allowed us to assess the impact of batch design on task difficulty.

User Engagement. The User Engagement Scale (UES) is a tool that has been used in a variety of digital domains to quantify UE. We assess worker involvement in this study using a standardized short questionnaire UES (O'Brien, et al. 2018).

Additionally, as part of the questionnaire, we included short scales questions (Rammstedt and John 2007) to collect personality traits information on the workers. Moreover, a feedback textbox was included to give the workers space for their opinion about the task.

Multiple logistic regression was used to test the if any different demographic features relate to multiple tasking behavior, this is done by predicting class membership or probability of class membership for a dependent variable based on multiple explanatory variables (Kwak and Clayton 2002). This has been used in our study to predict the probability of membership of each category of dependent variables (e.g., gender, age) to independent variables (e.g., task types, multiple tasking behavior).

As data was not normally distributed, we used Mann-Whitney test (MacFarland and Yates 2016) to assess if any significance different in the personality traits of the workers, engagement level and accuracy between each pair of task types from groups of those who behave to do multiple tasks and the groups of workers who did one tasks and also between any pair of different ordering of multiple tasking . The Kruskal test (McKight and Najab 2010) was used to assess significant differences between the three task types on engagement and personality traits.

In this paper we will focus on analysis of workers' perceived workload tasks, their engagement level and their performance relating to working on multiple tasks within short period of time in addition to the effect of their personality traits in such behavior.

Results and Discussion

After collecting the result from the platform, we analyzed the outputs of the workers. Since the survey task didn't have a quality control factor, we didn't exclude workers unless they didn't complete the whole document. In the classification tasks, we rejected the worker assignments who solved the patch of ten labeling tasks with less than 0.3 accuracy.

This was mentioned to the workers in the consent form before starting the task. Based on that, remaining workers data were, 139 workers out of 250 were approved in text-based classification tasks while 169 workers out of 250 were approved in image-based classification task and all 250 approved in survey task study. As a result, 558 worker results were included in our study. Since tasks were released with an approximate interval of 3 minutes between each task. Every assignment was intended to have a duration of around 10 minutes. Although no restriction was in place, we analyzed submissions and confirm that the workers who performed multiple tasks had an average time of approximately 3 minutes between tasks. The repeated release of tasks and their brief duration indicate that workers most likely concentrated on these activities one after another without engaging in irrelevant tasks, but this cannot be fully guaranteed.

The Effect of Multiple Tasking Behavior on Workers' Workload, Engagement and Performance- RQ1

To answer RQ1, we analyze the behavior of those doing multiple tasks and the effect of that on performance, perceived tasks load, by using different approaches. We clustered the workers into groups based on whether they did a single or multiple tasks regardless of the selection order of the tasks: (image and tweet classification), (tweet classification and survey), (image classification, survey), (all three tasks), and (only one task).

To ensure that our study was adequately powered to detect meaningful effects, we conducted a power analysis based on the sample sizes used on these clustered. Power analysis was performed using a medium effect size (Cohen's $d = 0.5$) and large effect size (Cohen's $d = 0.8$) and a significance level (α) of 0.05. Our sample sizes are as follows: 82 participants for the Image and Tweet Classification task, 63 participants for the Tweet Classification and Survey task, 60 participants for the Image Classification and Analysis task, and 33 participants for the All Three Tasks condition. These results suggest that our sample sizes in the first three groups (82, 63, 60) provide adequate power to detect medium effect sizes, while the power for the group of 33 participants was lower but still likely to detect a large effect. While this may limit the robustness of the conclusions drawn from this group, it may provide valuable insights, especially if additional studies or larger sample sizes are included in future studies This power analysis supports our adequate sample size for most groups in our study, thus increasing the reliability of our results and addressing potential concerns about our adequate sample size.

As seen in Table 1, workers who did all three tasks felt that the tasks were more physically and temporally demanding than those who did any of the pairs of tasks ($F(21, 62) = 330.5$, $p < 0.01$) and ($F(21, 62) = 347.1$, $p < 0.01$), respectively. Moreover, the workers that did all three tasks reported higher frustration than those who did image and tweet classification tasks ($F(21, 32) = 502.5$, $p = 0.001$), and more

	Sample Size	Mental Demanding	Physical Demanding	Effort Demanding	Performance Demanding	Temporal Demanding	Frustration Demanding
Image & tweet tasks	82	4.03	3.66	4.47	4.94	4.19	3.50
Tweet & survey	63	3.85	1.77	4.23	6.00	3.00	1.85
Image & survey	60	3.59	3.59	4.12	5.24	4.06	3.47
Three tasks	33	4.95	5.19	5.48	5.43	5.85	5.24

Table 1: Workers' workloads in different tasks

than those who did image classification and survey (F (13, 17) =260.5, p=0.007). Although the frustration was also higher than the tweet and survey task, the different was not statistically significant.

The workers' engagement level in doing only image classification task was higher than the other two individual tasks as shown in Figure 1. The workers' feedback explains this, as they reported that they found the images interesting. We also combined all workers who carried out all three tasks and compared them to those who carried out two tasks.

Moreover, the results show that there were statistically significant differences (p=0.0083) between the groups in terms of engagement and perceived workload. The workers who did (tweet & survey) tasks, were more engaged than those who did the (tweet & image), and those who did all three tasks with (F (13, 32) =330.0, p=0.001) and (F (13, 21) =210.5, p=0.004), respectively.

Based on the accuracy of the workers in the two classification tasks, see Figure 2 and Figure 3, around (20%) of workers who did only one classification task (in red color) had higher accuracy rate with average rate than those who participate on more than one task.

The result indicates that the survey task has heavier task load, this aligns with the outcomes of (Yang, et al. 2016), that some tasks, like a survey with numbers of radio buttons, may be difficult because of their structural complexity. As a result, doing other tasks in addition to survey tasks will increase the perceived tasks load which in turn negatively impacts workers' performance (Sweller 1988), or an effect of sepicalization in selectinng tasks (Mason and Watts 2009).

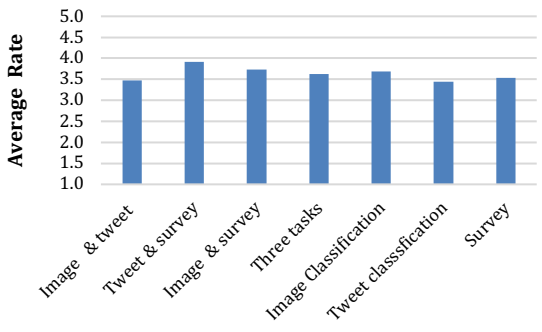

Figure 1: The Engagement level for different tasks.

The Correlation Between Workers' Personality Traits and Their Task-Workload, Performance and Engagement Levels. -RQ2

The study analyzed the relationship between CS workers' personality traits and their task workload, performance, and engagement levels. The results showed as in Table 2 that, in tweet classification and survey tasks pair, the workers had higher levels of Openness with an average rate (3.85), Extraversion (3.62), and Conscientiousness (4.23) compared to those who did all tasks or other pair of tasks. workers who did all three tasks exhibited elevated Neuroticism ratings, namely 3.0. In the task of tweet classification, there was a negative relationship between higher levels of Extraversion and accuracy, with a coefficient of -0.29 and a p-value of 0.13. On the other hand, Agreeableness and Conscientiousness had slight positive effects on accuracy, with coefficients of 0.16 and 0.17, respectively, and p-values of 0.48 and 0.44. However, these correlations were not statistically significant. Neuroticism had a coefficient of -0.06 with a p-value of 0.709, indicating a negative relationship.

Moreover, Openness had a coefficient of 0.048 with a p-value of 0.853, suggesting a positive relationship. However, neither of these relationships were statistically significant. In the task of classifying images, a higher level of Conscientiousness was found to possibly lead to an increase in accuracy (coefficient = 0.3158, p = 0.160). However, higher Extraversion levels (coefficient = -0.2603, p = 0.199), Agreeableness (coefficient = 0.0586, p = 0.770), Neuroticism (coefficient = -0.0662, p = 0.655), and Openness (coefficient = 0.1758, p = 0.437) did not showed statistically significant correlations with accuracy. In general, even though personality traits appeared to have an effect on accuracy, those outcomes did not attain statistically significant, which suggest further analyses using larger samples size or another analysis methods.

For further information and based on the worker's, personality traits, we analyze workers' Demographic features for those who did multiple tasking and their correlation to their personality. The result showed that the male workers outnumbered female ones, with around (72%) male participants. The distribution of ages is similar in all tasks with the majority in the age group 25 to 34 (above 50%) and around (50%) with a university degree or higher.

	Openness	Extraversion	Agreeableness	Consciousness	Neuroticism
Image & tweet	2.96	2.81	3.33	3.36	2.67
Tweet & survey	3.85	2.58	3.62	4.23	2.12
Image & survey	3.50	3.21	3.62	3.50	2.47
Three tasks	3.09	2.88	3.12	3.21	3.0

Table 2: Workers' personality traits in different tasks

Around (70%) of the workers were employed for wages, so the CS platform is not their only source of income. In line with previous studies (e.g. (Martin, et al. 2017)), the majority of workers (60%) participated from home and (50%) characterized themselves as having an intermediate level of 3-4 years of experience of crowdsourcing work.

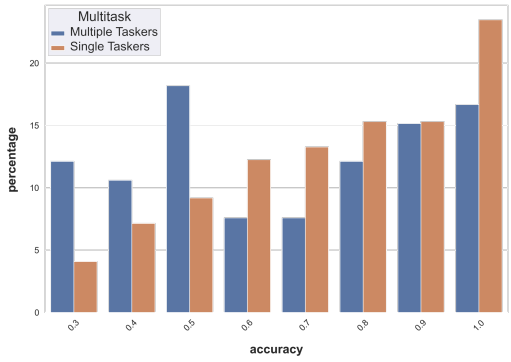

Figure 2: The percentage of workers with a certain accuracy rate in the image classification task

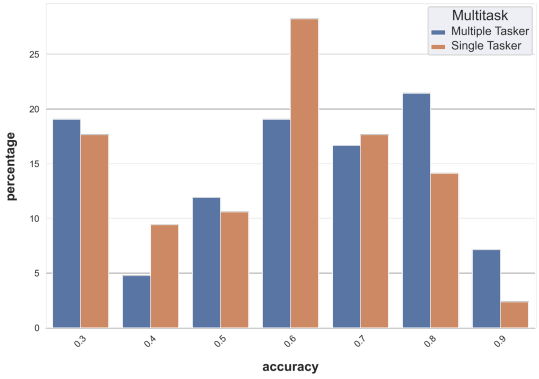

Figure 3: The percentage of workers with a certain accuracy rate in the tweet classification task

Selection Ordering of the Tasks – RQ3

Furthermore, in the second approach of analyzing multiple tasking behavior and to answer the RQ3, we consider the ordering of the tasks' selection for the workers who did two and three tasks.

We found that there were six groups of ordering selection of the tasks for the workers who did two tasks which are:

1. Image classification then survey task.
2. Image followed then tweet classification.
3. Survey then image classification task.
4. Survey then tweets classification task.
5. Tweets classification then survey task.
6. Tweets then image classification task.

While the groups who did all three tasks were:

1. Survey –image classification – tweet classification.
2. Survey– tweet classification – image classification.
3. Tweets classification–image classification – survey. There weren't any statistically significant differences in engagement levels, accuracy rate, or perceived workload for any of the task orders.

For the qualitative information analysis of the workers, feedback, Thematic analysis was used by Following a process defined by (Clarke and Braun 2017) to explain and present the raw data in a high constructional theme. Four major themes emerged from the thematic analysis of survey tasks, dog image categorization, tweet classification, and user feedback: suggestions, difficulties, time constraints, and positive attitude. The visually appealing dog image classification exercise elicited positive engagement from the participants, indicating a general enjoyment and appreciation for the tasks. To increase clarity and engagement in tweet task, crowd-workers recommend certain changes, such adding more examples and also there were other noteworthy issues that were observed, such as difficulties telling apart comparable items, which added to the task's complexity and confusion. Time restrictions were a major worry specifically in survey task, as many participants thought there wasn't enough time allotted, which made them feel hurried. Together, these themes highlight areas where task design can be improved, especially in terms of clarity and time management, to enhance task efficacy and user experience.

Overall Discussion

The objective of our study was to investigate the influence of task selection strategies on worker performance, engagement, and perceived burden in microtask crowdsourcing environments. Through the analysis of three frequently employed task categories, namely image classification, text classification, and surveys, we have determined that multitasking typically results in increased perceived workload

and dissatisfaction, while simultaneously diminishing engagement and accuracy. These findings indicate that the mental effort required to switch between different types of jobs can lead to reduced performance and happiness among workers.

The examination of personality traits indicated that attributes such as Neuroticism had a detrimental impact on task performance, whereas Conscientiousness and Openness were linked to enhanced involvement. This confirmed the possibility of utilizing personality traits to more effectively workers with tasks that correspond to their strengths, so improving overall engagement and performance.

The task order did not have an important impact on worker performance or engagement, indicating that the organization of activities may be less crucial than the design characteristics and complexity of the tasks. This discovery is consistent with recent progress in the field of crowdsourcing tooling, namely in the advancement of context-aware tailored task suggestion systems (Wang, et al. 2021). Crowdsourcing systems can improve overall efficiency and satisfaction by aligning tasks with worker skills and preferences through careful consideration of task design and complexity. Our findings confirmed the necessity of employing more advanced algorithms of task allocation that consider not only the complexity of tasks but also adjust to the distinctive traits of individual workers. This approach has the potential to enhance the effectiveness and personalization of crowdsourcing environments.

Limitations and Future Work

Additional studies on other voluntary crowdsourcing platforms are needed to generalize our results regarding the effect of multiple tasks selection on worker engagement, performance, and perceived workload. Publishing tasks at different times of the day may affect the samples of workers in terms of demographic characteristics, and thus influence other behaviors.

These studies can also figure out if the observed results demonstrate a consistent bias for workers involved in commercial crowdsourcing platforms. This study also poses the research question of how to develop crowdsourcing platforms that encourage workers to engage more with single tasks with high accuracy rather than multiple tasks with lower accuracy. Further study needed to be done by publishing different task types of different complexities' levels to study the effect of that on engagement level, performance, and perceived workload.

Conclusion

Crowdsourcing has transformed conventional employment paradigms by delegating jobs to a wide group of persons through open solicitations. This technique provides several benefits, including quick response times and the ability to access a huge sample size. These features make it extremely beneficial for tasks such as annotation and knowledge capture in AI research (Gadiraju, et al. 2014) . Gaining insight into the variables that influence worker performance and engagement in this field is essential for optimizing task structure and enhancing overall results.

This study offers valuable insights into the impact of task selection tactics on worker performance, engagement, and perceived burden in crowdsourcing contexts. Through the analysis of three frequently employed task categories, image classification, text classification, and surveys. We have determined that multitasking typically results in increased perceived workload and dissatisfaction, while also having a detrimental effect on engagement and accuracy. These findings align with previous work, which found that task complexity and the transition between different types of tasks can reduce performance and increase errors (Monsell 2003; Wylie and Allport 2000).

Our analysis of personality traits revealed that characteristics such as Neuroticism had a negative effect on task performance, whereas Conscientiousness and Openness were associated with improved engagement. This study builds upon prior research by demonstrating that personality traits can be leveraged to match workers with tasks that line with their strengths, resulting in enhanced engagement and performance (Kazai, et al. 2011) (Vander Shee, et al. 2020).

Furthermore, our research discovered that the sequence in which tasks were carried out did not have a major effect on the performance or involvement of workers. This implies that the arrangement of tasks may be less crucial compared to the characteristics and intricacy of the tasks themselves. According to (Cai, et al. 2016), in order to improve the effectiveness of crowdsourcing, it is crucial to prioritize the design and complexity of tasks rather than the order in which they are offered. Our findings indicate that multitasking can increase cognitive load and reduce accuracy, with personality traits playing a significant role in worker performance.

In conclusion, we have proposed a few CS task designs guideline for the requesters who have more than one task to be published: Publish the tasks based on your priority since most of the workers pick the tasks sequentially as their published. This study is the first known attempt to investigate the influence of task selection techniques on worker performance across numerous unique task types of comparable difficulty.

References

Adamopoulos, P.; Ghose, A.; Todri, V. 2018. The impact of user personality traits on word of mouth: Text-mining social media platforms. Information Systems Research 29(3):612-640.

Aipe, A.; Gadiraju, U. 2018. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. *In* Proceedings of the 29th on Hypertext and Social Media, Pp. 115-122.

Bachrach, Y.; Kosinski, M.; Graepel, T.; Kohli, P.; Stillwell, D. 2012. Personality and patterns of Facebook usage. In Proceedings of the 4th Annual ACM Web Science Conference, Pp. 24–32.

Bhatti, S. S.; Gao, X.; Chen, G. 2020. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey. Journal of Systems and Software, 167:110611.

Cai, C. J.; Iqbal, S. T.; Teevan, J. 2016. Chain reactions: The impact of order on microtask chains. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 3143-3154.

Chandler, J.; Mueller, P.; Paolacci, G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behavior research methods, 46(1):112-130.

Clarke, V.; Braun, V. 2017 . Thematic analysis. The journal of positive psychology, 12(3), 297-298

Crone, D. L.; Williams, L.A. 2017. Crowdsourcing participants for psychological research in Australia: A test of microworkers. Australian Journal of Psychology, 69(1):39-47.

Dai, P.; Rzeszotarski, J. M.; Paritosh, P.; Chi, E. H. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 628-638.

Difallah, D.; Catasta, M.; Demartini, G.; Cudré-Mauroux, P. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2014. Vol. 2.

Doan, A.; Ramakrishnan, R.; Halevy, A. Y. 2011. Crowdsourcing systems on the world-wide web. Communications of the ACM 54(4):86-96.

Dong, J.; Kang, Y.; Liu, J.; Sun, C.; Fan, S.; Jin, H.; Liu, X. 2023. Human-centred design on crowdsourcing annotation towards improving active learning model performance. Journal of Information Science:01655515231204802.

Gadiraju, U.; Dietze, S. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 2017, pp. 105-114.

Gadiraju, U.; Kawase, R.; Dietze, S. 2014. A taxonomy of microtasks on the web. Proceedings of the 25th ACM conference on Hypertext and social media, 2014, pp. 218-223.

Gegenhuber, T.; Schuessler, E.; Reischauer, G.; Thäter, L. 2022. Building collective institutional infrastructures for decent platform work: The development of a crowdwork agreement in Germany. In Organizing for societal grand challenges. Pp. 43-68: Emerald Publishing Limited.

Gosling, S. D.; Gaddis, S.; Vazire, S. 2007. Personality impressions based on facebook profiles. Icwsm 7:1-4.

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; Bigham, J. P. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Pp. Paper 449. Montreal QC, Canada: Association for Computing Machinery.

Hart, S. G.; Staveland, L. E 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology. Pp. 139-183: Elsevier.

Ho, C. J.; Slivkins, A.; Suri, S.; Vaughan, J. W. 2015. Incentivizing high quality crowdwork. Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 419-429.

Hornuf, L.; Vrankar, D. 2022. Hourly wages in crowdworking: A meta-analysis. Business & Information Systems Engineering 64(5):553-573.

Horton, J. J.; Chilton, L. B. 2010. The labor economics of paid crowdsourcing. Proceedings of the 11th ACM conference on Electronic commerce, 2010, pp. 209-218.

Huang, J. H.; Yang, Y. C. 2010. The relationship between personality traits and online shopping motivations. Social Behavior and Personality: an international journal 38(5):673-679.

Huang, Y.; Jasin, S.; Manchanda, P. 2019. “Level up”: Leveraging skill and engagement to maximize player game-play in online video games. Information Systems Research 30(3):927-947.

Ihl, A.; Strunk, K. S.; Fiedler, M. 2020. The mediated effects of social support in professional online communities on crowdworker engagement in micro-task crowdworking. Computers in Human Behavior 113:106482.

Imran, M.; Mitra, P.; Castillo, C. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894.

Jain, A.; Sarma, A. D.; Parameswaran, A.; Widom, J. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. arXiv preprint arXiv:1701.06207.

Sammut, C.; Webb, G. I. 2011. Encyclopedia of machine learning. Springer Science & Business Media.

Kazai, G.; Kamps, J.; Milic-Frayling, N. 2011. Worker types and personality traits in crowdsourcing relevance labels. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 1941-1944.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; & Li, F. F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), 2011. Vol. 2. Citeseer.

Lascau, L.; Gould, S. J.; Cox, A. L.; Karmannaya, E.; Brumby, D. P. 2019. Monotasking or multitasking.

Lenart-Gansiniec, R. 2021. The benefits of crowdsourcing in science: Systematic literature review. Reading: Academic Conferences International Limited.

Leonidou, L, C; Kvasova, O; Christodoulides, P; Tokar, S. 2019. Personality traits, consumer animosity, and foreign product avoidance: the moderating role of individual cultural characteristics. Journal of International Marketing 27(2):76-96.

Li, M. W.; Teng, H. Y.; Chen, C. Y. 2020. Unlocking the customer engagement-brand loyalty relationship in tourism social media: The roles of brand attachment and customer trust. Journal of Hospitality and Tourism Management 44:184-192.

MacFarland, T. W.; Yates, J. M. 2016. Mann–whitney u test. *In* Introduction to nonparametric statistics for the biological sciences using R. Pp. 103-132: Springer.

Martin, D.; Carpendale, S.; Gupta, N.; Hoßfeld, T.; Naderi, B.; Redi, J.; Wechsung, I. 2017. Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. In Evaluation in the crowd. crowdsourcing and human-centered experiments, Pp. 27-69: Springer.

Mason, W.; Watts, D. J. 2009. Financial incentives and the" performance of crowds". Proceedings of the ACM SIGKDD workshop on human computation, 2009, pp. 77-85.

McCrae, R. R.; Costa, P. T. 2003. Personality in adulthood: A five-factor theory perspective: Guilford Press.

McKight, P. E., & Najab, J. 2010. Kruskal-wallis test. The corsini encyclopedia of psychology,1-1.

Monsell, S. 2003. Task switching. Trends in cognitive sciences, 7(3): 134-140.

Mourelatos, E.; Giannakopoulos, N.; Tzagarakis, M. 2022. Personality traits and performance in online labour markets. Behaviour & Information Technology, 41(3):468-484.

Newell, E.; & Ruths, D. 2016. How one microtask affects another. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 3155-3166.

O'Brien, H. L.; Cairns, P.; Hall, M. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. International Journal of Human-Computer Studies, 112:28-39.

Piyathasanan, B.; Mathies, C.; Patterson, P.; de Ruyter, K. 2018. Continued value creation in crowdsourcing from creative process engagement. Journal of Services Marketing.

Qiu, S.; Bozzon,A.; Birk,M.; Gadiraju,A. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing.

Qiu, S.; Gadiraju, U.; Bozzon,A. 2020. Improving worker engagement through conversational microtask crowdsourcing. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1-12.

Rammstedt, B.; John, O. P. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality 41(1):203-212.

Saravanos1,A. Zervoudakis1,S.; Zheng1,D.; Stott,N.; Hawryluk,B.; Delfino,D. 2021.The Hidden Cost of Using Amazon Mechanical Turk for Research. arXiv:2101.04459.

Schäper, T.; Foege, JN.; Nüesch, S.; Schäfer, S.2021. Determinants of idea sharing in crowdsourcing: evidence from the automotive industry. R&D Management, 51(1):101-113.

Shi, X.; Evans, R.; Pan, W.; Shan, W. 2021. Understanding the effects of personality traits on solver engagement in crowdsourcing communities: a moderated mediation investigation. Information Technology & People.

Kwak, C.; Clayton-Matthews, A. 2002. Multinomial logistic regression. Nursing research, 51(6), 404-410.

Sulaiman, A.; Jaafar, N. I.; Tamjidyamcholo, A. 2018.Influence of personality traits on Facebook engagement and their effects on socialization behavior and satisfaction with university life. Information, Communication & Society 21(10):1506-1521.

Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. Cognitive science 12(2):257-285.

Troll, J.; Blohm, I.; Leimeister, J. M. 2016. Revealing the impact of the crowdsourcing experience on the engagement process.

Troll, J.; Blohm, I.; Leimeister, J. M. 2019. Why incorporating a platform-intermediary can increase crowdsourcees' engagement. Business & Information Systems Engineering, 61(4):433-450.

Vander Schee, B. A.; Peltier, J.; Dahl, A. J. 2020. Antecedent consumer factors, consequential branding outcomes and measures of online consumer engagement: current research and future directions. Journal of Research in Interactive Marketing.

Wang, J.;Yang, Y.; Wang, S.; Chen, C.; Wang, D., & Wang, Q. 2021. Context-aware personalized crowdtesting task recommendation. IEEE Transactions on Software Engineering, 48(8):3131-3144.

Webster, J.; Ahuja, J. S. 2006. Enhancing the design of web navigation systems: The influence of user disorientation on engagement and performance. Mis Quarterly:661-678.

Wylie, G.; Allport, A.2000 . Task switching and the measurement of "switch costs". Psychological research, 63:212-233.

Yang, J.; Redi, J.; Demartini, G.; Bozzon, A. 2016 Modeling task complexity in crowdsourcing. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2016. Vol. 4.

Yang, M.; Ren, Y.; Adomavicius, G. 2019. Understanding user-generated content and customer engagement on Facebook business pages. Information Systems Research, 30(3):839-855.

Zhang, Y.; Sun, Y.; Kim, Y. 2017. The influence of individual differences on consumer's selection of online sources for health information. Computers in Human Behavior, 67:303-312.