Performance of Paid and Volunteer Image Labeling in Citizen Science — A Retrospective Analysis

Kutub Gandhi,[1] **Sofia Eleni Spatharioti,**[2] **Scott Eustis,**[3] **Sara Wylie,**[1] **Seth Cooper,**[1]
[1] Northeastern University
[2] Microsoft Research
[3] Healthy Gulf
gandhi.ku@northeastern.edu, sspatharioti@microsoft.com, scott@healthygulf.org, s.wylie@northeastern.edu, se.cooper@northeastern.edu

Abstract

Citizen science projects that rely on human computation can attempt to solicit volunteers or use paid microwork platforms such as Amazon Mechanical Turk. To better understand these approaches, this paper analyzes crowdsourced image label data sourced from an environmental justice project looking at wetland loss off the coast of Louisiana. This retrospective analysis identifies key differences between the two populations: while Mechanical Turk workers are accessible, cost-efficient, and rate more images than volunteers (on average), their labels are of lower quality, whereas volunteers can achieve high accuracy with comparably few votes. Volunteer organizations can also interface with the educational or outreach goals of an organization in ways that the limited context of microwork prevents.

1 Introduction

Citizen science can be a powerful approach for finding potential solutions to difficult problems. Crowdsourcing solutions in citizen science projects can often be more accessible but just as effective as consulting with subject matter experts, or even replace experts in cases where none exist.

There are two general approaches a scientific organization can take for crowdsourcing in their citizen science projects: engaging volunteers in public citizen science projects, or using paid labor through microtask platforms. Volunteers often join citizen science projects due to an interest in seeing the project succeed or out of a more general desire for learning and challenge (Tinati et al. 2017; Haywood 2016). Paid workers however often join as part of the "gig economy", working on small projects for low pay during free time or unemployment.

There may be differences between how these two populations interact with crowdsourcing tasks, however. These differences may be especially important to organizations looking to spend resources on one of the two approaches — either in terms of the monetary cost of hiring workers, or the monetary, social, and temporal costs of building up a volunteer network.

We noticed systematic differences in the quality of labels with workers from Mechanical Turk and volunteers while conducting an analysis of image label data provided by Healthy Gulf (Healthy Gulf 2022) — a partner organization which has spent the last two years utilizing both their volunteer network and Amazon Mechanical Turk to identify land loss and climate change related damage in aerial imagery of the Louisiana coast. In order to better understand the differences between these populations in the context of an image labeling task, we analyzed various metrics regarding the two populations. This analysis was retrospective, insofar as it was conducted after Healthy Gulf had completed their crowdsourcing project, rather than as part of a singular, controlled experiment designed to test specific hypotheses. See section 3.7 for further discussion.

We found that workers completed more images per participant than the volunteer population (on average), though a feature displaying volunteer progress roughly doubled the quantity of images volunteers labeled.

Despite providing fewer labels on average though, volunteer labels were of higher quality: they were generally more accurate (determined as agreement with experts) than Mechanical Turk workers, regardless of whether the volunteers were members of Healthy Gulf or not. Additionally, of the many images where volunteers were confident on the label of an image (had high inter-participant agreement), it was highly likely that that label was correct. This beneficial property of volunteer labels did not as apply as simply or directly to Mechanical Turk workers: worker confidence did not always appear to correlate with accuracy and there were few high-confidence images.

Finally, we found that the large number of volunteer votes Healthy Gulf collected were likely unnecessary: votes beyond 5-10 volunteer votes provided diminishing benefits to accuracy.

Beyond simple image labeling, it is important to note that Healthy Gulf's volunteer network was beneficial to their other goals such as climate change advocacy and educational outreach — goals that cannot be reasonably furthered by Mechanical Turk workers due to the limited context of a Mechanical Turk task.

This work contributes insights into the different populations of image labelers, which we hope will be beneficial to any organization deciding between these two crowdsourcing approaches. We must, however, note two important limitations to our contributions: Firstly, the crowdsourcing task was an image labeling task, and so conclusions may not be

generalizable to other task types. Secondly, we provide descriptive statistics and discussion of the dataset, but the retrospective nature of this analysis prevents us from conducting inferential statistical tests.

2 Background

Citizen Science The usage of citizen science as an approach in environmental projects is well established. For example, Munro, Schnoebelen, and Erle (2013) found that volunteer image labeling for Hurricane Sandy data was extremely effective: The volunteers only disagreed with experts on a small fraction of images (though experts were only asked to look at a subset of high difficulty images) and were able to achieve high accuracy with only ~ 5 separate volunteer guesses per image. Buytaert et al. (2014) discuss a litany of case study projects taking place around the planet, including hydrology projects out of Peru, Ethiopia, Nepal, and Kyrgyzstan. They claim that these projects serve as a testament to the power of wider public participation in solving the hydrological problems of these areas, as opposed to "traditional, external sources of information."

Dickel et al. (2019) illustrate how broad citizen science disciplines can often be, categorizing citizen science projects as falling into three major categories: (1) emancipatory (the most common), focusing on assembling the public towards a common goal; (2) entrepreneurial, taking the form of an enterprise built to foster innovation; and (3) science communication, educational projects meant to bring an issue to public light.

Volunteers Citizen science volunteers have complex motivations. Beyond a simple desire to see a project succeed, many are motivated by the sense of community a project might bring, the challenge and learning that comes with the puzzle-oriented structure of most citizen science tasks, and the sense of recognition or achievement that comes with successfully completing a task (Tinati et al. 2017; Haywood 2016). These motivations can be accommodated with careful design, such as enabling discussion via forums, allowing for community leadership, providing communication with the science team behind each project, and providing context before and feedback after each task (Tinati et al. 2015).

Citizen science volunteers follow patterns of increasing engagement with projects; having designers focus on engaging volunteers and encouraging them to take an active part in their physical or virtual communities can lead to subsets of volunteers who take on complex leadership and organizing roles (Hendricks, Meyer, and Wilson 2022). Engaging a volunteer with the scientific process has benefits beyond the project itself: Lewandowski and Oberhauser (2017) describe how a successful butterfly conservation project caused volunteers to not only stay engaged with that specific project, but caused them to become increasingly engaged with the scientific process and citizen science as a whole.

Despite the societal benefits of community engagement, there are ethical concerns organizations must be cognizant of. Beyond the obvious issue that volunteering is unpaid labor (Hendricks, Meyer, and Wilson 2022), volunteers overwhelmingly identify as white and educated — causing many citizen science projects to target and benefit already wealthy populations (Allf et al. 2022). Organizations ought to extend their efforts towards underserved populations while recruiting diverse voices into their ranks. As our study did not gather demographic data directly, we are unable to assess participant diversity of our populations.

Mechanical Turk workers Many tasks feasible for citizen science volunteers are also feasible for workers on platforms such as Amazon Mechanical Turk. These workers complete short individual tasks for low pay during free time or unemployment. Mechanical Turk workers are generally as diverse as the internet using population (Paolacci and Chandler 2014), have an average level of scientific knowledge (Cooper and Farid 2016), and fill out surveys reasonably honestly and conscientiously (Paolacci and Chandler 2014). These characteristics, along with inbuilt verification of worker behavior, allow for Mechanical Turk to be a useful platform for subject pool accessibility and rapid iteration (Mason and Suri 2012).

While Mechanical Turk is defined by the fact that workers are paid for tasks, Kaufmann, Schulze, and Veit (2011) found that workers have more motivations than money, including autonomy in how they perform their tasks and variety in the tasks themselves. Designing around these motivations and providing context about the meaningfulness of a task is useful: workers who found their tasks meaningful worked for longer, while workers who thought of their tasks as useless contributed a lower quality of responses (Chandler and Kapelner 2013).

Unlike with volunteers, task designers must be cognizant of the possibility of workers who maliciously seek payment without contributing. Gadiraju et al. (2015) describe various kinds of malicious workers, such as those who input seemingly correct but ultimately useless responses to tasks. They go on to recommend techniques to reduce the harm malicious workers can cause to projects; recommendations include slipping questions in specifically to validate worker responses or ensure they are paying attention. Eickhoff and de Vries (2013) further recommend structuring a task in ways that discourage automation while filtering potential workers "by origin or through a recruitment step" to avoid potentially malicious workers (though filtering by prior acceptance rates was ineffective).

While Mechanical Turk workers can abuse citizen science projects, the potential for abuse can run the other way as well when projects take advantage of the low pay traditionally afforded to workers. We urge organizations to pay a livable wage and discuss the ethics of our payment scheme in section 3.4.

Workers and Volunteers Previous work assessing citizen science volunteer and Mechanical Turk workers show nuance regarding the differences between the populations.

With simple tasks, Mao et al. (2013) found that paying workers for time spent led to similar quality responses as volunteers (though accuracy could be traded for speed by switching to a "pay per task" model). Krause and Kizilcec (2015) found similar results, though they gave volunteers an image labeling *game* while giving workers a direct image

labeling task.

Differences between response quality started to appear with higher difficulty tasks, however. When Krause and Kizilcec (2015) attempted to use workers and volunteers for a complex website annotation task, they noticed that volunteer responses were of higher quality than worker responses. Similarly, Sarkar and Cooper (2018) found higher quality responses with volunteers who played a human computation focused video game, as opposed to workers who were tasked with doing the same.

Extending Prior Work This paper attempts to provide further evidence of potential differences between these two populations. Our task is notably distinct from previous work in the arguably more specialized image labeling task. Whereas image labeling in some other works (Mao et al. 2013; Krause and Kizilcec 2015) referred to associating or typing nouns that were represented in a photograph, our image labeling task required participants to identify subtle visual patterns that are related to wetland loss or restoration. Furthermore, we use a "recruitment cost" fixed payment scheme as opposed to paying per task / paying per time like Mao et al. (2013) and Krause and Kizilcec (2015). Further discussion of payment is in section 3.4.

3 Methodology

3.1 Citizen Science Projects

In this work, we focused on two citizen science projects created in collaboration with Healthy Gulf, a non-profit organization focused on protecting the US Gulf coast's natural resources. Healthy Gulf's goal for both projects was to raise awareness towards Louisiana's rapidly deteriorating wetlands. To this end, Healthy Gulf aimed to engage the crowd in identifying six different wetland loss or restoration patterns by looking at near-infrared aerial photographs. 387 locations of interest were selected for the projects, in areas of the lower Barataria watershed on the West Bank of Jefferson Parish, and the East and West Banks of Plaquemines Parish in Louisiana. The six main patterns chosen were Shoreline Erosion, Shipping, Oil & Gas, Agriculture, Restoration and Sea Level Rise. Imagery was acquired using the Louisiana Department of Natural Resources's SONRIS tool.

Participants would be assigned to one of the six patterns at random every time they visited the project page. A brief tutorial about the specific pattern would first be presented. Effectively, this resulted in 2,322 images in need of classification per project (387 locations $\times$ 6 patterns). Both projects focused on the same points of interest and the same six patterns; however, the near infrared imagery was collected at different points in time. The first project looked at imagery from 2016, while the second focused on 2008 imagery. These two image sets are designed to be generally comparable, though changes in image capture technology may result in some differences. Example images are shown in Figure 1. The task presented participants with a single image at a time and asked them if the pattern was present or not (Figure 2).

2016 (primary dataset)

2008 (alternate dataset)

Figure 1: Examples of images from the two datasets used. Imagery publicly available through the Strategic Online Natural Resources Information System (SONRIS).

3.2 Datasets Analyzed

This paper examines the analysis of image labels associated with the 2016 imagery, sourced from Mechanical Turk workers and citizen science volunteers, along with environmental science experts (used for determining accuracy). These image labels are henceforth referred to as **the primary dataset**. This paper similarly analyzes the image labels associated with the 2008 imagery, although experts were only asked to analyze a subset due to limited resources. For 2008 imagery, experts only voted on images where workers and volunteers disagreed on what the label should be; we refer to this subset as "controversial" images. This might also capture a practical workflow where experts are called only to analyze those images for which participants do not come to a consensus. Image labels associated with the 2008 images are henceforth referred to as **the alternate dataset**.

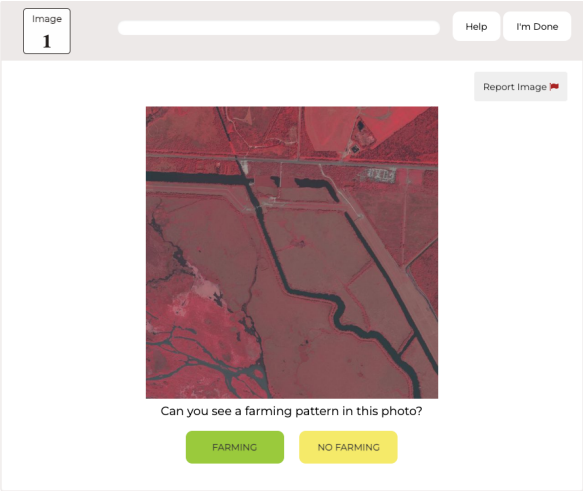

Figure 2: An example of the type of task a participant would encounter, though some features of the interface changed throughout data collection, as discussed in section 3.7.

This paper focuses on understanding differences between Mechanical Turk workers and citizen science volunteers rather than the environmental insights gained from these images. As such, we do not focus on the six wetland patterns as individual categories, instead treating them as one singular image set. For discussion on the environmental understandings from this project, see Spatharioti et al. (2022).

3.3 Recruiting through Community Outreach

Recruitment for volunteers was done primarily through Healthy Gulf, SciStarter (SciStarter 2022), Public Lab (Public Lab 2022), and the River Rally conference (River Rally 2022) via emails, webinars, and other social media. A banner about the project was also included on the website for Foldit (Cooper et al. 2010), a citizen science game about protein folding.

For the 2016 imagery (primary dataset), 835 volunteers participated, primarily from January 2021 to July 2021 (the full recruitment period was July 2020 to November 2021). For the 2008 imagery (alternate dataset), 379 volunteers participated, primarily from July 2021 to September 2021 (the full recruitment period was July 2021 to January 2022).

3.4 Recruiting through Crowdsourcing Marketplaces

For Mechanical Turk, we chose a "recruitment cost" fixed payment scheme where participants were paid USD 1.30 for labeling as many images as they wished. No bonuses were awarded per label contributed. This payment scheme was chosen for a variety of reasons:

1. We feel that this payment scheme is more ethical, allowing workers to determine the worth of their own labor.
2. We avoid implicitly or explicitly enforcing a certain amount of image labeling, a measure we felt would prevent spamming (low effort image labeling).
3. This payment scheme allowed us to understand the natural rate at which Mechanical Turk workers would consider themselves finished with the project at this level of payment.
4. Spatharioti et al. (2017) show that this payment scheme actually *increases* the number of images labeled by Mechanical Turk workers.

This payment scheme led to an average of USD 0.01226 per label, which is commensurate with other payment schemes, e.g. Krause and Kizilcec (2015)

The workflow for workers was the same as volunteers, with the exception that workers did not have the option of revisiting the project. For the 2016 imagery (primary dataset), we recruited 156 workers in February 2022. For the 2008 imagery (alternate dataset), we recruited 433 workers in July 2021.

3.5 Expert Reviews

Subject matter experts from Healthy Gulf reviewed every image in the primary dataset, leading to 6.3 expert votes per image on average.

Due to limited resources, experts were only asked to review a subset of images in the alternate dataset (specifically, the "controversial" images where Mechanical Turk workers and volunteers disagreed on the appropriate label), leading to 1.1 expert votes per image on average.

3.6 Voting and Image Labeling

A "vote" is defined as an input by a single volunteer, worker, or expert about a single image. The image's label is the majority vote by the volunteers or workers: i.e. the "volunteer image label" for an image could be "Yes", composed of 50 individual "No" votes and 100 "Yes" votes. In the case of a tie, the image was considered to be labeled "Yes" — that evidence of a morphology was found.

The majority vote amongst experts for any given image was considered the ground truth, with ties being considered evidence of a pattern.

3.7 Notes on the Retrospectivity of the Analysis

The datasets in question were gathered for purposes other than the analyses in this paper. As such, we consider these analyses to be retrospective, rather than strict A/B tests. In this section we discuss some potential confounding factors:

1. A progress indicator was removed during the data collection period. The potential effects of this change are discussed in section 4.2.
2. Aesthetic UI updates were conducted throughout the data gathering period. These updates were not functional, and we do not expect they affected any aspect of our analyses.
3. Smaller A/B tests were conducted between control populations and populations who were presented minor feature changes. The vast majority of these A/B tests found no statistical significance and our overall population size is notably larger than each individual test, thus we are not overly concerned about the confounding effects of these tests.

Primary Dataset	Accuracy
Mechanical Turk	70.4%
Volunteers	89.6%
Volunteers (random subset)	86.5%
Baseline (always vote no)	67.1%
Alternate Dataset (image subset, see sec. 4.1)	Accuracy on controversial
Mechanical Turk	25.4%
Volunteers	74.6%
Baseline (always vote no)	75.9%

Table 1: Accuracy for various participant types. In the primary dataset, Mechanical Turk workers were highly inaccurate, nearly as much as a baseline naive model that always gave an image the label "no pattern found". Volunteers were nearly 90% accurate however, an accuracy that barely declined even if a small random subset was taken (See section 4.1). In the secondary dataset, experts only voted on images where workers and volunteers disagreed; they agreed with volunteers 74.6% of the time.

While we do not consider the listed effects as invalidating to our analyses, we remain cautious and limit our contributions to descriptive statistics and discussion of effects seen in data exploration, rather than inferential statistical tests.

4 Results

4.1 Workers and Volunteers Accuracy Comparison

We began by simply looking at the various populations in question and their respective accuracies for the primary dataset. For the ground truth accuracy we used the majority expert vote for each image. The results are described in Table 1. As we can see, the Mechanical Turk workers were notably less accurate than their volunteer counterparts, with a 19.2 percentage point difference in their accuracies. As a check, we also considered a hypothetical baseline population that always voted "No pattern" for each image, the most common vote. Seeing as there was only a 3.3 percentage point difference between Mechanical Turk workers and this hypothetical population, it appears that little information was gained from Mechanical Turk workers.

One aside is that our dataset includes far more votes from volunteers than Mechanical Turk workers, with 47.9 votes per image (average) from volunteers, and only 5.4 coming from Mechanical Turk. Is it possible that volunteers were more accurate simply by the wisdom of the (larger) crowd?

We rule out this possibility for two reasons: Firstly, we conducted Monte Carlo simulations of what it would look like if there were similar numbers of volunteer votes as there were Mechanical Turk votes. In other words, we took a random subsample of volunteer votes of the same size as the number of Mechanical Turk worker votes. We conducted this simulation 500 times. Despite limiting the number of votes in these simulations, volunteer accuracy dropped by only 3.1 percentage points. Further discussion of the effects of limiting volunteer votes can be found in section 4.4.

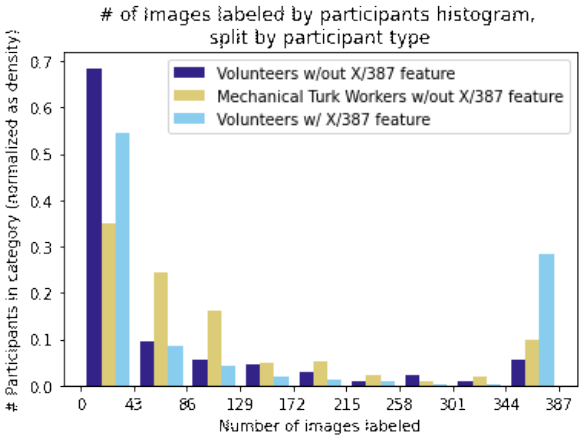

Figure 3: Dropoff curve (a display of when participants left the project) for group across both datasets. Some volunteers were presented a progress indicator feature showing them that they had labeled "X/387" images. Volunteers (with and without the progress feature) often left early, though volunteers who saw the feature had a sizeable cohort that stayed and finished rating the entire dataset. Mechanical Turk workers also left early, though many stayed to rate 50-100 images, and a few rated the entire dataset.

Secondly, the alternate dataset had far more Mechanical Turk worker votes than volunteers (21.7 worker votes and 7.7 volunteer votes for each image on average), yet we saw similar results. In the alternate dataset, experts were asked only to look at images where Mechanical Turk workers and volunteers disagreed on the correct label (limited to this subset to reduce expert workload). In this subset, experts agreed with volunteers over the workers 74.6% of the time.

We attempted to ameliorate the differences in accuracy via the application of the Dawid-Skene Expectation Maximization algorithm (Dawid and Skene 1979). This algorithm attempts to deprioritize voters who are found to be unreliable. While this algorithm was somewhat effective for Mechanical Turk workers in the primary dataset, leading to a 12.5 percentage point increase in accuracy, the application of the algorithm actually decreased accuracy of every other group[1]. For this reason, we suspended further analysis with the Dawid-Skene algorithm, though future work could attempt to identify why the algorithm was ineffective or look at other algorithms for crowdsourced labeling.

4.2 Number of Images Labeled by Participant Type

When looking at the number of images labeled by different groups, we took into consideration a progress indicator feature which showed participants that they had labeled

[1] Accuracy was reduced by 6.7, 2.7, and 10.4 percentage points for volunteers in the primary dataset, workers in the alternate dataset, and volunteers in the alternate dataset, respectively.

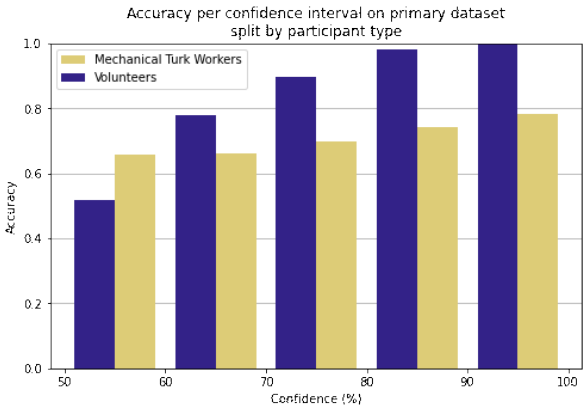

Figure 4: Accuracy as a function of confidence (binned), split by participant type in the primary dataset. As participant confidence increased, their accuracy did as well. This change was far more extreme with the volunteers however.

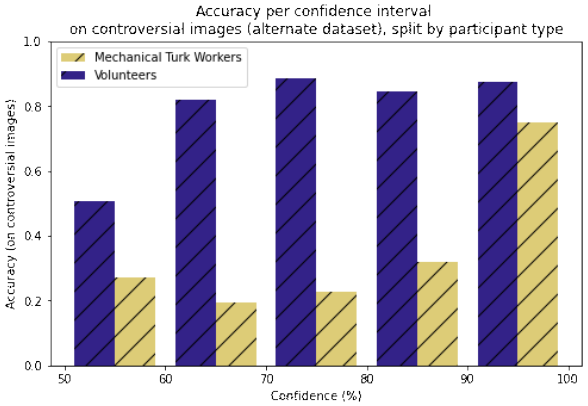

Figure 5: Accuracy as a function of confidence (binned), split by participant type in the alternate dataset. Since the alternate dataset only had expert labels for controversial images, this figure **cannot** be directly compared to Figure 4. This figure shows that, for an image where Mechanical Turk workers and volunteers disagree, the Mechanical Turk workers are only likely to be correct if they are 90-100% confident (which, as Figure 7 shows, only occurs in exceedingly few images). On the other hand, for an image where the two groups disagreed, the volunteers were more likely to be correct at any confidence level — though there was a dip with images at confidence 50-60%.

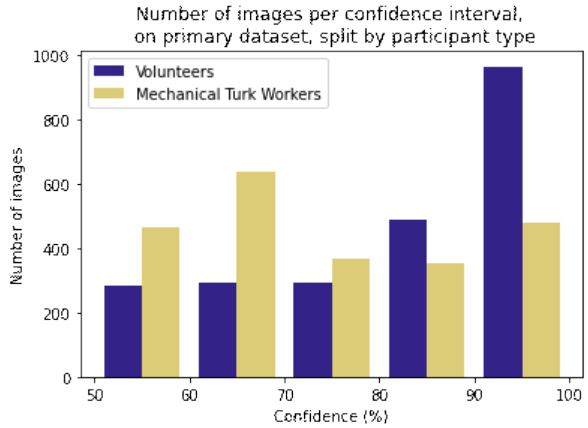

Figure 6: The number of images at each confidence level for the participant types in the primary dataset. Volunteers had many images at high confidence, indicating that they often agreed with one another. Mechanical Turk workers however often disagreed, which is illustrated by the fact that their modal confidence is the 60% bin.

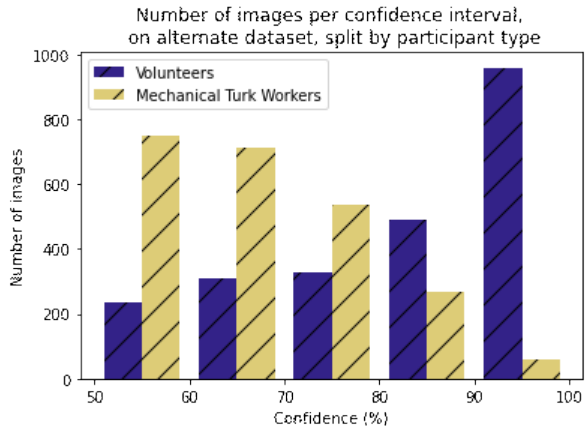

Figure 7: The number of images at each confidence level for the participant types in the alternate dataset. This figure provides further evidence of the discussion in Figure 6, though the results here are even more striking: volunteers usually agreed with each other, whereas with most images, workers disagreed with one another.

"X/387" images. Some volunteers saw this progress indicator and some did not, but no Mechanical Turk workers did.

Among Mechanical Turk workers and volunteers who did not see their progress, Mechanical Turk workers labeled many more images. Across both datasets workers labeled an average of 106 images per worker, whereas volunteers (without progress indicator) labeled 65 images per volunteer. The volunteer mean was skewed by a large subgroup of volunteers who labeled fewer than 10 images before leaving, presumably those who had a passing interest in the project, took a look at how the project functioned, and then moved on.

However, adding the progress indicator flipped this volunteer-worker discrepancy: These volunteers (who comprised of 58.7% of our total volunteer population) had a similar subgroup of members who labeled fewer than 10 images, but also had a sizeable subgroup of those who labeled the entire data set. Volunteers who were shown the size of the dataset labeled 135 images on average, roughly double volunteers who were not given a progress indicator, who labeled 65 each on average. Coincidentally, the overall average of all volunteers (with and withough progress indicator) was 106

images.

There are a variety of potential causes for this increase in volunteer labeling. The "387" could have caused an anchoring effect, volunteers may have felt compelled to complete the entire dataset, and / or volunteers may have felt comforted that the dataset was not practically endless. This is in line with prior work showing that Mechanical Turk workers completed more images when shown a progress indicator (Spatharioti et al. 2017).

The feature showing the size of the dataset was not activated for any Mechanical Turk worker. We wanted workers to label as many images as they felt was appropriate for the amount we were paying, however preliminary Mechanical Turk projects showed that workers might be mistaking the "X / 387" as the amount they were required to complete (and were thus feeling undue pressure to complete more images than they would have otherwise). To prevent this pressure, we disabled the feature. This discussion can be visualized in Figure 3, which depicts the number of participants leaving vs number of images completed. (Number of participants leaving is by density, not absolute value to account for differences in population sizes).

4.3 Accuracy as a Function of Participant Confidence

Participant confidence is a metric describing to what extent participants agreed with each others' votes. Specifically, confidence was defined as the percent of participants who voted for the majority answer. For example, if there were 100 votes of "Yes" and 50 votes for "No", inter-participants agreement would be $\frac{100}{100+50} \approx 67\%$. Confidence could also be thought of as inter-participant agreement. By definition, confidence could not go below 50%.

Analyzing the volunteers in the primary dataset, we found that confidence appears correlated with accuracy (Figure 4). For example, if we take the subset of images where volunteers were 50–60% confident, volunteers agreed with experts for only 61% of these images. On the other hand, if we take the images when volunteers were 90%+ confident, they agreed with the experts on every image in this subset. This apparent correlation holds with the alternate dataset (Figure 5), though not as strongly. This is likely due to the set up of the alternate dataset (with experts voting only on images of disagreement between workers and volunteers) rather than any inherent differences between the two volunteer populations.

These same results did not hold for Mechanical Turk workers; it seems that the likely prevalence of workers who were "spamming" (repeatedly voting without appropriate consideration for what the correct label may be) caused large number of images to be placed in a confidence category arbitrarily. However, in the alternate dataset, the highest-confidence worker images did lead to high accuracy, though there were very few images that fell into this category.

To further illustrate these differences in confidence, we can look at a histogram generally describing the number of images that were at various confidences across the two groups (Figures 6, 7).

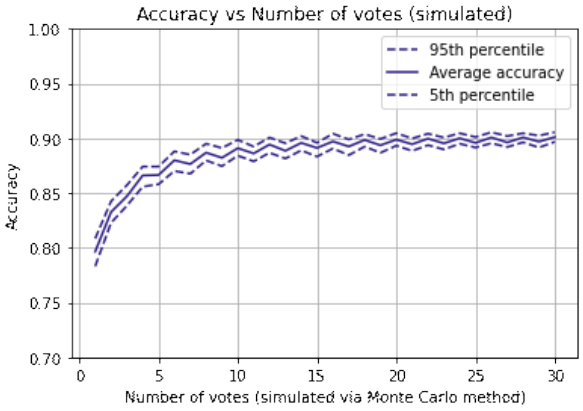

Figure 8: Number of votes vs accuracy for images rated by volunteers in the primary dataset. As number of volunteer votes increases, accuracy increases, however there are diminishing returns at $N = 5$ and increasingly diminishing returns at $N = 10$. The dashed lines indicate the 95th / 5th percentile runs in all 200 runs, indicating that the accuracy level in these simulations had low variation.

We can see that volunteers tended to agree with each other, with the modal confidence being 90%+. On the other hand, Mechanical Turk workers generally tended to disagree with one another, a potential indication of voting with less care. The modal confidence is between 60–70% for the primary dataset, and 50–60% for the alternate dataset, with few images reaching anything close to unanimity.

In an ideal case, confidence can be used as a proxy for image difficulty. Experts can be asked to spend limited time on a subset of images with low confidence, or perhaps these images could be scrutinized to improve tutorials. These results indicate however that these applications of participant confidence may require further thought when using data from Mechanical Turk workers.

4.4 Accuracy as a Function of Number of Votes

In section 4.1 we discussed how limiting the volunteer votes to 5 votes per image led to an only 3.1 percentage point drop in accuracy; this leads to a natural question: were the number of votes received from volunteers necessary for the level of accuracy seen? In order to understand this question, we created subsets of the data where each image only had N volunteer votes and varied N from 1 to 30. In order to prevent random effects from skewing the results, we repeated the simulation for each N 200 times and averaged the accuracy values across all the simulations. We found diminishing returns after $N = 5$ and increasingly diminishing returns after $N = 10$, which is illustrated in Figure 8.

This idea of increasing votes leading to increasing accuracy was not consistent across the dataset however. If one assumes that volunteer confidence is an appropriate proxy for difficulty (i.e. assuming that a high confidence image is one that is easy, and a low confidence image is difficult), then accuracy gains differ based on the difficulty of each image. We conducted further analysis into how accuracy changed

with increasing votes and found that high difficulty images (50-66% confidence) do not gain from an increased number of votes (presumably since volunteers will be inaccurate on them through and through). Low difficulty images (83.3-100% confidence) also do not gain from increased votes (presumably since they are so easy that many votes are not required). Medium difficulty images (66-83.3% confidence) are where the benefits of large number of voters were most prominent. These statements are subjective, based on the rough criteria that going from 3 votes to 10 votes had little effect for high and low confidence images, but caused an approximately 10 percentage point increase in accuracy for medium confidence images.

Thus, we recommend that organizations focus volunteers on medium difficulty images rather than having them waste votes on high or low difficulty images.

4.5 Association with Healthy Gulf and Accuracy

Could differences between individual volunteer subgroups cause notable differences in the ways in which they interacted with the citizen science task? Volunteers had varying reasons for participating and varying ways in which they found the project. Some were members of Healthy Gulf, some followed citizen science advertising platforms, some were recommended the project by their company or school, and others stumbled on the project organically.

Here we focus on an important divider between volunteers: whether or not they were members of Healthy Gulf. To enable anonymity of participants, we did not track where participants had come from, however an optional survey presented upon completion of the task allowed 38.0% of participants to self-identify regarding their potential association with Healthy Gulf. All analyses in this section were completed with volunteers from the primary dataset.

Volunteers who indicated in the survey that they were not members of Healthy Gulf ($n = 249$) labeled an average of 195 images, 84.0% more than the general average volunteer (an unsurprising result, seeing as filling out the post-task survey would reasonably be associated with higher interest in the task). Volunteers who did claim to be associated ($n = 91$) had an even higher average: 307, which is 250.9% more images labeled than the average volunteer.

In terms of accuracy, Healthy Gulf members and non-members were both 79.0%[2] accurate at the task. This 10.6 percentage point decrease from the general average could possibly be caused by fatigue, though data exploration found no correlation between number of images labeled and accuracy (see Figures 9, 10). While this decrease is disheartening, it contains a silver lining: it seems that organizations like Healthy Gulf may not need to worry about volunteers outside their organization creating lower quality labels.

5 Discussion

The accuracy of the citizen science volunteers seen in this project provides confidence that volunteer votes can be used with minimal concern in future projects. Furthermore, the straightforward correlation between volunteer confidence and accuracy sets up a potential workflow for projects without copious resources: simply limit the experts to confirming images where volunteer confidence is low, solicit more votes when confidence is medium, and trust the volunteers when they are confident.

Unfortunately, the same could not be said for Mechanical Turk workers. The lack of accuracy, issues with internal voting disagreement, and lack of straightforward correlation between confidence and accuracy caused difficulties in using image labels generated primarily through Mechanical Turk (such as some preliminary attempts to train machine learn-

[2]No difference with three significant figures.

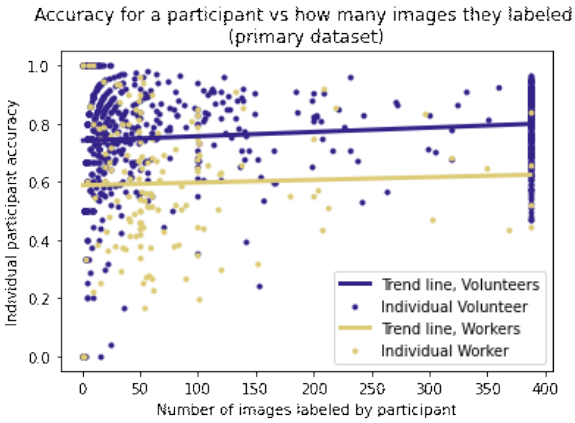

Figure 9: Accuracy for participants depending on how many images they labeled (in the primary dataset). There appears to be no correlation, indicating that fatigue from image labeling likely did not cause a notable decrease in accuracy.

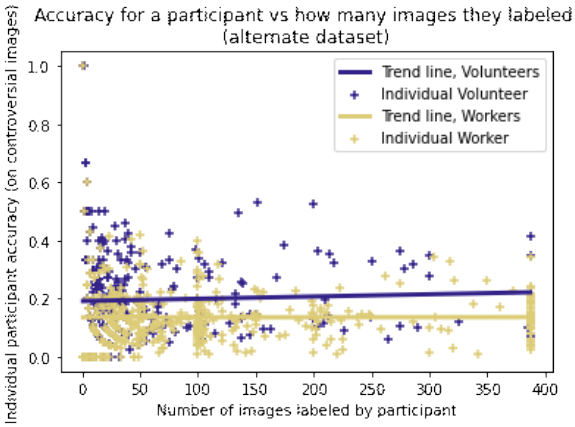

Figure 10: Accuracy for participants depending on how many images they labeled (in the alternate dataset). There appears to be no correlation, indicating that fatigue from image labeling likely did not cause a notable decrease in accuracy. Interestingly, few volunteers were over 60% accuracy, despite general volunteer accuracy being 74.6% on this dataset. This implies that volunteers by themselves struggled, but volunteer votes coming together is what led to the higher general accuracy number.

ing models).

While Mechanical Turk was not the appropriate choice for this project, that's not necessarily the case for all projects. The accessibility and monetary benefits of Mechanical Turk cannot be understated, especially seeing as the gathering of a reasonable number of volunteers took multiple months to complete. Mechanical Turk is a valuable tool to gather pilot data, for tasks amenable to the Mechanical Turk framework, or in projects where spamming is less of a concern.

There are other values within these projects beyond getting accurate image labels however. The goals of Healthy Gulf include outreach, advocacy, and education; building up a strong network of volunteers allows Healthy Gulf to interact with these volunteers in community oriented projects. In fact, one of Healthy Gulf's more recent projects, identifying oil spills and environmental damage in Nigeria[3], was spearheaded and developed in direct collaboration with people who were originally crowdsourcing volunteers. As an interesting point of fact, the diversity of Mechanical Turk workers likely had the unintended benefit of bringing diverse voices to the project.

Even if an organization doesn't hold these exact education and advocacy goals, building up a network of trust and engagement with volunteers allows that organization to call upon a helpful population for future projects.

Future Work and Limitations

1. This analysis was exploratory, and while steps were taken to limit confounding effects, we nonetheless recommend that future work replicate our analyses with measures to ensure population homogeneity and appropriate controls.
2. This was an image labeling task looking at satellite imagery, which is difficult to analyze and does not contain much diversity. Future work could look at generalizing these results beyond the task presented in this paper in order to determine which tasks are appropriate for Mechanical Turk workers and which require volunteers or even experts.
3. Differences in motivation and expertise likely contributed to the differing outcomes of the two populations. A study specifically designed around isolating these differences and understanding their effects on outcomes would be beneficial.
4. There are platforms other than Mechanical Turk, and even within Mechanical Turk, there are filters such as the Mechanical Turk Masters Qualification. How do these populations compare to the ones discussed in this paper?
5. This study did some preliminary work looking at the Dawid-Skene Expectation Maximization Algorithm and found that the algorithm actually decreased accuracy in some cases. Understanding the reasoning behind this quirk or finding algorithms that are more generally effective would be useful for organizations looking to increase the effective accuracy of their human computation populations.

[3]https://scistarter.org/land-pollution-lookout

6 Conclusion

This work looked at two similar sets of image label data provided by our partner organization, Healthy Gulf. A retrospective analysis of their attempts to utilize Mechanical Turk workers and their volunteer network uncovered systemic differences in the two populations. While Mechanical Turk workers labeled more images than volunteers, their labels were of a lower quality and their confidence for individual images did not correlate well with their actual accuracy on those images.

Further analysis indicating that only a few volunteer votes were necessary for subjectively good accuracy and that having more votes was most effective for medium difficulty images sets up a nicely simple workflow: for images where volunteers are highly confident, take their labels as truth. For images of medium confidence, solicit further volunteer votes. Finally, for images of low confidence, spend limited expert resources on those images specifically. Further work can examine effective ways to identify these images while they are being voted on.

Of course, Mechanical Turk workers are still a useful population for human computation, especially in simpler tasks, for rapid preliminary gathering of data, or for use in projects where the concerns of lower accuracy can be ameliorated.

Finally, we cannot focus overly much on the raw numbers. Organizations should keep in mind the subtler benefits of maintaining an engaged volunteer network, both to the organization itself, but also to the world at large.

Acknowledgements

The authors would like to thank all participants, along with Caroline Nickerson (outreach), Asha Padmashetti (web development), Drishti Sabhaya (web development), Akash Sethumurugan (web development), Scistarter, Healthy Gulf, Public Lab, and River Rally. This material is based upon work supported by the National Science Foundation under grant no. 1816426.

References

Allf, B. C.; Cooper, C. B.; Larson, L. R.; Dunn, R. R.; Futch, S. E.; Sharova, M.; and Cavalier, D. 2022. Citizen Science as an Ecosystem of Engagement: Implications for Learning and Broadening Participation. *BioScience*, 72(7): 651–663.

Buytaert, W.; Zulkafli, Z.; Grainger, S.; Acosta, L.; Alemie, T. C.; Bastiaensen, J.; De Bièvre, B.; Bhusal, J.; Clark, J.; Dewulf, A.; Foggin, M.; Hannah, D. M.; Hergarten, C.; Isaeva, A.; Karpouzoglou, T.; Pandeya, B.; Paudel, D.; Sharma, K.; Steenhuis, T.; Tilahun, S.; Van Hecken, G.; and Zhumanova, M. 2014. Citizen Science in Hydrology and Water Resources: Opportunities for Knowledge Generation, Ecosystem Service Management, and Sustainable Development. *Frontiers in Earth Science*, 2.

Chandler, D.; and Kapelner, A. 2013. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets. *Journal of Economic Behavior & Organization*, 90: 123–133.

Cooper, E. A.; and Farid, H. 2016. Does the Sun revolve around the Earth? A comparison between the general public

and online survey respondents in basic scientific knowledge. *Public Understanding of Science*, 25(2): 146–153.

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; et al. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307): 756–760.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Dickel, S.; Schneider, C.; Thiem, C.; and Wenten, K.-A. 2019. Engineering Publics: The Different Modes of Civic Technoscience. *Science & Technology Studies*, 8–23.

Eickhoff, C.; and de Vries, A. P. 2013. Increasing Cheat Robustness of Crowdsourcing Tasks. *Information Retrieval*, 16(2): 121–137.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640.

Haywood, B. K. 2016. Beyond Data Points and Research Contributions: The Personal Meaning and Value Associated with Public Participation in Scientific Research. *International Journal of Science Education, Part B*, 6(3): 239–262.

Healthy Gulf. 2022. Healthy Gulf. https://www.healthygulf.org/. Accessed: 2022-06-09.

Hendricks, M. D.; Meyer, M. A.; and Wilson, S. M. 2022. Moving Up the Ladder in Rising Waters: Community Science in Infrastructure and Hazard Mitigation Planning as a Pathway to Community Control and Flood Disaster Resilience. *Citizen Science: Theory and Practice*, 7(1).

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. Worker motivation in crowdsourcing – a study on Mechanical Turk. In *Proceedings of the Americas Conference on Information Systems*.

Krause, M.; and Kizilcec, R. 2015. To play or not to play: Interactions between response quality and task complexity in games and paid crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Lewandowski, E. J.; and Oberhauser, K. S. 2017. Butterfly Citizen Scientists in the United States Increase Their Engagement in Conservation. *Biological Conservation*, 208: 106–112.

Mao, A.; Kamar, E.; Chen, Y.; Horvitz, E.; Schwamb, M. E.; Lintott, C. J.; and Smith, A. M. 2013. Volunteering versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing. In *In Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*.

Mason, W.; and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1): 1–23.

Munro, R.; Schnoebelen, T.; and Erle, S. 2013. Quality Analysis after Action Report for the Crowdsourced Aerial Imagery Assessment Following Hurricane Sandy. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*.

Paolacci, G.; and Chandler, J. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science*, 23(3): 184–188.

Public Lab. 2022. Public Lab. https://publiclab.org/. Accessed: 2022-06-09.

River Rally. 2022. River Rally. https://www.rivernetwork.org/connect-learn/river-rally/. Accessed: 2022-06-09.

Sarkar, A.; and Cooper, S. 2018. Comparing paid and volunteer recruitment in human computation games. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*, 1–9.

SciStarter. 2022. SciStarter. https://scistarter.org/. Accessed: 2022-06-09.

Spatharioti, S.; Boetsch, E.; Eustis, S.; Gandhi, K.; Rota, M.; Apte, A.; Cooper, S.; and Wylie, S. 2022. An Effective Online Platform for Crowd Classification of Coastal Wetland Loss. *Conservation Science and Practice*. Forthcoming.

Spatharioti, S. E.; Govoni, R.; Carrera, J. S.; Wylie, S.; and Cooper, S. 2017. A Required Work Payment Scheme for Crowdsourced Disaster Response: Worker Performance and Motivations. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*, 475–488.

Tinati, R.; Luczak-Roesch, M.; Simperl, E.; and Hall, W. 2017. An Investigation of Player Motivations in Eyewire, a Gamified Citizen Science Project. *Computers in Human Behavior*, 73: 527–540.

Tinati, R.; Van Kleek, M.; Simperl, E.; Luczak-Rösch, M.; Simpson, R.; and Shadbolt, N. 2015. Designing for Citizen Data Analysis: A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4069–4078.