This project is co-financed by the European Union

Grant Agreement No.: 824603
Call: H2020-SwafS-2018-1
Type of action: RIA
Starting date: 1/02/2019

ACTION

D 5.2 – Final Guidelines for Task Design

Coordinator: Neal Reeves, KCL
Quality reviewer: Antonella Passani, T6ECO

Deliverable nature	R - Report
Dissemination level	Public
Work package and Task	WP5, T1
Contractual delivery date	M24
Actual delivery date	

Authors

Author name	**Organization**	**E-Mail**
Neal Reeves	**King's College London**	**Neal.t.reeves@kcl.ac.uk**

Abstract	This deliverable presents final guidelines on task design, building on earlier deliverable 5.1. We present high level recommendations from citizen science and scientific crowdsourcing administrators. We also present research conducted by the ACTION consortium on the topic of task design.
Keywords	Citizen science, task design, guidelines, interview, statistical analysis

Disclaimer
The information, documentation and figures available in this deliverable, is written by the ACTION project consortium under EC grant agreement 824603 and does not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

How to quote this document
Reeves, N. (2021) Final Guidelines for Task Design

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document builds on D 5.1 (Initial Guidelines for Task Design) and sets out final guidelines and recommendations for the design of citizen science tasks. In line with the focus of the second round of the ACTION accelerator and the needs of the second cohort of ACTION projects, we focus particularly on projects with a strong online component or *virtual citizen science projects*.

Firstly, we present findings from an interview study with administrators of citizen science and crowdsourced research projects, presenting a set of high-level recommendations synthesised from the comments and experiences of 12 researchers.

We then present an analysis of the impact of designing for short-term interventions and projects such as the Bioblitz and other event-driven citizen science initiatives as identified in D 5.1, analysing engagement in two short-term projects held by the Zooniverse platform over 72 hours.

Thirdly, we analyse the role that data preparation and particularly clustering has on volunteer participation, conducting a small experiment with crowd workers to address whether clustering tasks according to difficulty has a significant impact on worker engagement.

We conclude by introducing and analysing the Virtual City Explorer developed by the KCL team, which allows for data gathering in a virtual environment without the need to visit a location in person. We evaluate the completeness of this approach using a crowdsourcing experiment and comparison with gold standard datasets, before presenting a quality assurance method designed to function with the more complex workflows that the VCE facilitates.

1 INTRODUCTION

If a citizen science project is to be successful, the tasks, tools and resources must be designed and implemented in such a way that they account for the needs, expectations and capabilities of volunteer citizen scientists, who may have little or no knowledge or experience of scientific research. At the same time, projects and particularly tasks must gather or produce data which is of sufficient quality to be used for research purposes.

This deliverable focuses on task design and serves as a follow-up to the initial findings and recommendations around task design presented in the earlier deliverable 5.1. We present final recommendations and guidelines to support citizen science projects. To complement and further the findings presented in D5.1, we consider additional stages of the scientific workflow in this deliverable, beyond and in addition to the collection of data. We also focus particularly on *web-based* online projects, in keeping with the focus of the second ACTION open call and the pilots and initiatives which will form the second ACTION accelerator, to take place in 2021.

We first present findings from an interview process engaging 12 stakeholders using paid and volunteer microtask crowdsourcing such as is common in web-based citizen science. We follow this by presenting research conducted by the ACTION consortium around task design, including an analysis of short-term projects and their impact on participation, an analysis of the impact that clustering of input data has on participant engagement and an introduction of the Virtual City Explorer, a web-based interface for exploring physical environments.

2 Administrator Interviews

To understand the design principles and lessons that citizen science project administrators and researchers acquire throughout the course of their projects, we conducted 12 semi-structured interviews. Because of the similarities between virtual citizen science and microtask crowdsourcing[1], we also included crowdsourcing researchers and administrators within our sample. We present here findings from these stakeholders, grouped as high level design recommendations for scoping and launching crowdsourced initiatives for science.

2.1 Methodology

To facilitate the interview process, we elected to restrict our participant base to only include those with at least one scientific publication with a paid, crowdsourced methodology. Participants were recruited through a snowball sampling method, where participants recommended other participants from among their co-authors and network of contacts. This method was supplemented with targeted recruitment of participants to expand the demographics, platforms and levels of experience present within our sample of participants. We recruited a total of 12 participants. In terms of disciplines, the majority of our participants had explicit or implicit qualifications in the field of computer science or associated fields. However, the published research topics were interdisciplinary and defied easy categorisation, spanning business, social science, linguistics, psychology and behavioural sciences, among others.

Interviews were conducted using the Microsoft Teams video conferencing service and recorded, before later being transcribed and analysed using a thematic analysis approach. We then grouped these themes into recommendations which we present below.

2.2 Choose a suitable approach

Although not a significant phase of the task design problem, participants noted the need to decide whether citizen science or crowdsourcing more broadly were suitable for their problem. Generally, before formulating a task, it is essential that the researchers and designers consider which factors are most important for them and what level of inaccuracy – if any – they are willing to tolerate and can be overcome or accounted for within their later research. As one participant suggested, there are two essential questions which must be answered at a very early stage, prior to proceeding with a crowdsourced or citizen science methodology: what quantity and quality of data does the researcher require?

"For me the process starts from two questions. One is about the quantity of the data. How much data do I need? And the other is about the quality of the data." (Participant 6)

[1] Citizen science and crowdsourcing hold many similarities, with the most common point of divergence the use of payments and rewards, which are rare in citizen science but common in other forms of crowdsourcing (Shanley et al., 2019, Simperl et al., 2018).

Tasks that require a specific level of expertise above that of the average member of the general public, require long periods of participation to make a single contribution, where inaccurate results cannot be tolerated or where large quantities of data are not essential may be better suited for other, more traditional laboratory-based or field study experiments with specially selected or expert participants.

2.3 Formulate a suitable problem

The nature of scientific research questions does not always easily align with the questions and tasks that are to be presented to citizen scientists. While scientific research questions may be broad and multi-faceted, such a question is not always suitable for a volunteer to address. Instead, it is important to think how these questions could be mapped to a task that a volunteer can complete:

“Ultimately [we] find some way to map [a question] to a crowdsourcing task where we at least know what are the inputs and what are the outputs of it... See like can we map the inputs to the outputs in a way that's not too overwhelming for one person in like one microtask session” (Participant 2)

Allowing for too great a diversity in responses – or too much autonomy for workers – can make tasks difficult to troubleshoot or pose barriers to quality assurance at a later stage. This is likely to prove problematic, as citizen scientists are not experts and training opportunities are often limited – there is no guarantee that a volunteer, no matter how well-meaning, will be able to perform a task with complete accuracy:

``The problem is more on the task itself. How do you evaluate whether they did well? So these are not the things that you can easily check with some competency questions where you really know the answer or aren't easy checks on the paragraph of text you can do to verify the quality." (Participant 9)

It is also important that administrators define their input and output data – that is, what is going to be provided to the citizen scientists and what data is expected at the end? Providing too much data to volunteers can overcomplicate a task or make it too difficult or time-consuming. Participants should be provided – and asked to provide – only the data that will be used for subsequent research, to limit the time needed for a single contribution, maximise the number of contributions that can be gathered and prevent the exploitation of contributors by asking them for work which will not be used:

“What you want to know? Because it's not like we want to have people transcribe the whole card. There's probably 30 to 40 fields on on a particular type of card… If we're asking for something we want to make sure that we use it and it's -- if we don't need it, we don't ask for it. We don't want to waste anyone's time.” (Participant 11)

2.4 Account for trade-offs

The use of large-scale crowdsourcing – either paid or voluntary – to conduct research has some significant advantages in terms of the scale and speed at which data can be gathered. However, this is generally conducted by recruiting non-expert participants who may have little to no knowledge of the field or of the scientific method and asking them to participate in research that focuses on concepts with which they may have little or no familiarity. This risks an inevitable trade-off between quantity, speed and accuracy:

“It is an essential trade off that needs to be dealt with because either you do something fast or you do it well... So you have to think that people want to do this. You have to help them do this fast and you also have to understand that you might not get perfect results.' (Participant 9)”

2.5 Account for technology

While not necessarily specific to citizen science, it is important to account for the different technologies that participants might use to access citizen science tasks and how these might fit with the types of task that participants are being asked to complete. Tasks should account for mobile devices and different browsers, particularly where a participant is asked to complete tasks away from their home or computer, such as gathering photos or data from around a city or specific location. Participation in citizen science can be very brief and it is important to remove any unnecessary barriers to participation to harness as much participation as possible.

“If they can't load the survey because what I built doesn't work on their browser, then they can't do the task.” (Participant 5)

2.6 Provide Context

One of the factors that most drives participation in research through citizen science and other crowdsourcing methods is the desire to contribute to and participate in genuine scientific research (Jun, Morelle and Reinicke, 2018; Simperl et al., 2018). However, the need to render tasks in a way that is simple for participants with any level of knowledge and experience can risk trivialising the research process and harming volunteer engagement, as volunteers fail to understand why an otherwise simple task that could be performed by anyone can still be valuable for scientific research:

“Because yeah, doing a job on a platform can be boring. So many of these jobs ask questions that for a human are trivial. Sometimes they seem like total nonsense to lay people who do not do research. So I think that knowing that they took part in an academic effort and they are helping science may encourage them." (Participant 6)

Our participants suggest that this can be overcome through a combination of providing feedback on the progress of research and contextualising the task to account for and explain these concerns:

“You know, I think it just goes back to sort of the narrative of why they are participating and why they would like to -- what they would like to contribute to? And then being able to demonstrate that” (Participant 11)

2.7 Provide Feedback

Volunteers in citizen science and crowdsourcing desire reassurance that what they’re doing is correct and valuable. Where possible, administrators proposed to integrate feedback into the task workflow, allowing volunteers to understand how well they were performing without the need for additional effort:

“I think they also if they can get it, they are interested in getting feedback on their work quality, during the task at least.” (Participant 2)

This can be a challenge, however, given the nature of citizen science, where often the correct response – or accuracy of a given response – is not known prior to gathering data. One proposed method for this was to use gold standard questions for which the answer is known, which additionally functions as a quality assurance method, in addition to serving as an opportunity to provide feedback to contributors.

2.8 Solicit – or pay attention to – feedback from contributors

When initially designing and scoping citizen science projects, we noted a number of assumptions administrators are required to make, including the capabilities of workers, their motivations and expectations, the difficulty of a task and how immediately accessible it may serve to be. These did not always accurately align with reality when volunteers commenced their participation and comments, queries and complaints for volunteers were relatively common.

It can be tempting to ignore these comments, but participants suggest that they in fact include vital information about how volunteers perceive and understand tasks and should be used to adjust, update and simplify tasks to match the expectations of volunteers, while of course balancing the needs of research.

“Incorporate feedback from pilots... And really, take seriously what goes wrong and then try to fix that. Whether it is like a technical failure or just something that is confusing. It's definitely worth fixing that because you'll have happier, happier workers and better results in the end.” (Participant 2)

Moreover, even where such communications do not provide explicit feedback – most notably, when they are a request for clarification – there can still be value in responding to and communicating with volunteers to provide more information and clarity, to reassure them and allow them to feel that their voice is heard:

"Even if it's just for me to say like I appreciate the email, you know, this is how what happened, happened. Oftentimes I'll get replied that just says thanks. Appreciate it and that's the end of it, so sometimes it seems like they just kind of want clarity." (Participant 5)

2.9 Avoid ambiguity

Participants noted that citizen science – and crowdsourcing – tasks function most effectively when they are unambiguous. Multiple choice tasks with discrete, finite options are simpler and gather more valuable results than more open ended questions. This recommendation extends beyond the tasks themselves to the instructions and onboarding process that citizen scientists can follow. The requirements needed to begin and to complete a task should be clearly laid out in a manner which is easy to follow and does not permit deviation:

"Yeah, part of that is the task design where you're looking for very clear, unambiguous instructions. So like fill out this field, fill out this field you know. Where there's not a lot of -- there isn't a lot of ambiguity in what you should do. You shouldn't have to sort of make judgment calls" (Participant 11)

3 Short-term Citizen Science

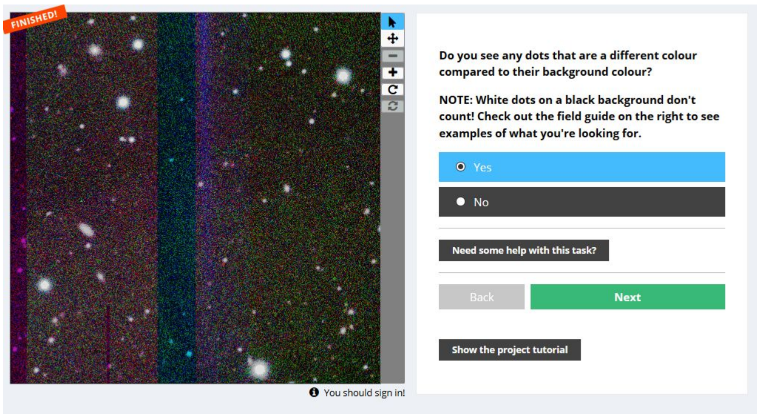
While citizen science projects may run for many months or even years, increasingly initiatives are being launched which are intended to last for only a brief time. These include annual events such as Bio Blitzes, competitions within existing projects and even projects designed to run solely for a matter of days. As identified in D 5.1, these short-term activities are particularly common in pollution, where many projects are 'restricted' to occur at set times and places. One of the outstanding areas of uncertainty from D 5.1 was the extent to which these limitations influence participation in citizen science projects, if indeed at all.

To address this question, in this chapter we explore engagement in two projects conducted over the course of 72 hours within the Zooniverse platform. We analyse contributions over time in terms of the submission of classifications and data from volunteers and socially driven discussion participation within both projects, as well as the timing of initial and final participation. We also compare these participation patterns with contemporary projects launched within the Zooniverse platform at the same time.

3.1 Background

The Zooniverse platform has frequently partnered with BBC Stargazing Live to launch new projects, including extensions to existing projects and short-term projects designed to last 48 hours, covering the three broadcasts of Stargazing Live which take place 24 hours apart[2]. In 2017, two such projects were launched: Planet Nine in association with BBC Stargazing Live and Exoplanet Explorers, launched in conjunction with the new ABC Stargazing Live in Australia. Projects are launched during an initial one-hour live show, with updates 24 hours later live on air and 'final results' (as available) are revealed in a final broadcast 48 hours after the first launch.

Both Planet Nine and Exoplanet Explorers were relatively simple tasks – participants were presented with either an image (P9) or a graph (EE) and asked to answer a binary, yes-or-no answer. The interface for both projects can be seen in figures 1 and 2 below.



Do you see any dots that are a different colour compared to their background colour?

NOTE: White dots on a black background don't count! Check out the field guide on the right to see examples of what you're looking for.

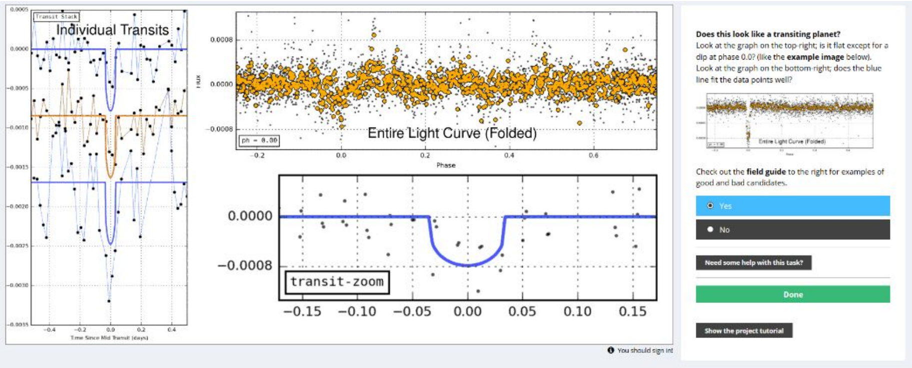
● Yes

● No

You should sign in!

[2] In truth, these projects last slightly longer, generally covering at least a subsequent 24 hours before they are retired and closed to new contributions.

Figure 1 – Planet Nine classification interface showing asset (left) and task process (right). Note that ‘next’ proceeds to a subsequent image and not a continuation of the classification task.



Does this look like a transiting planet?
Look at the graph on the top-right; is it flat except for a dip at phase 0.0? (like the **example image** below).
Look at the graph on the bottom-right; does the blue line fit the data points well?

Check out the **field guide** to the right for examples of good and bad candidates.

● Yes

○ No

Need some help with this task?

Done

Show the project tutorial

Figure 2 – Exoplanet Explorers classification interface showing asset (left) and task process (right).

3.2 Methodology

To conduct this analysis, we used records listing the time of each classification or discussion message, the associated username and the rough geographic location. We first produced graphical indicators to explore how participation grew over time, before removing all anonymous classifications prior to proceeding further. We then used Python to calculate each volunteer’s first and last classification, first and last comment and their period of participation.

Following this, we gathered daily classification statistics for the initial 72 hours of 14 Zooniverse projects launched within the three months prior or three months following Planet Nine and Exoplanet Explorers. We used statistical analysis, through the Mann-Whitney U test to compare participation between the projects as a means to identify whether participation was comparatively higher in these short-term projects or whether participation is simply consistently high during the initial launch of a project.

3.3 - Classification participation over time

Classifications in Planet Nine can be seen within figure 1. In Planet Nine, while participation starts extremely high, there is a rapid drop in engagement after approximately 6 to 8 hours. Engagement then begins to gradually rise until the second broadcast, at which participation rises once again, before rapidly falling a second time. A second, smaller boost in participation is seen after the third and final broadcast, before a final drop and engagement remains low after this point.

Classifications in Exoplanet Explorers can be seen within figure 2. As with Planet Nine, participation dwindles over time and is mostly revitalised by the broadcast events. However, the drop in participation is much slower and less pronounced than in Planet Nine and participation takes longer to fall off following the last broadcast event. Classification levels in Exoplanet Explorers were consistently lower than within Planet Nine.

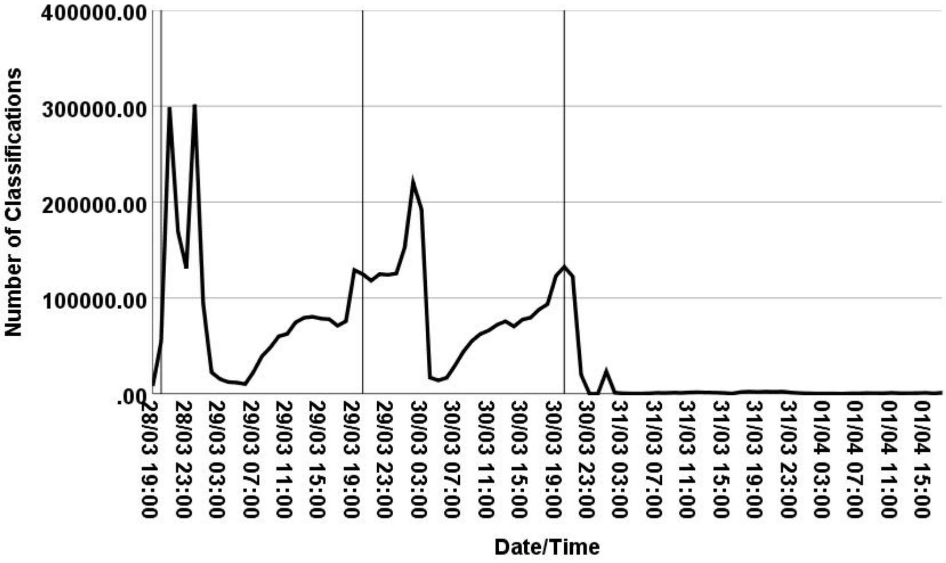


Figure 3 – Line graph showing hourly classification count within Planet Nine

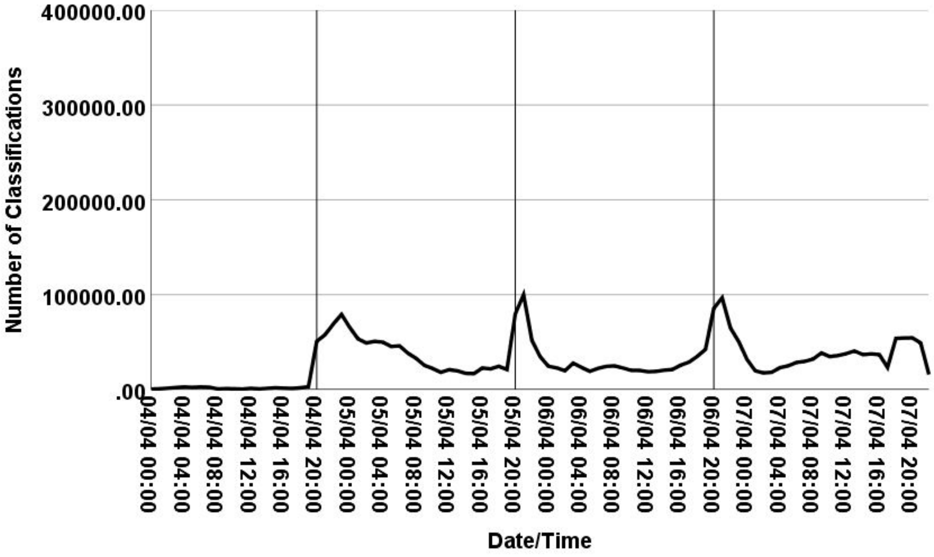


Figure 4 – Line graph showing hourly classification count within Exoplanet Explorers.

3.4 - First Classifications

We also analysed the number of users making their first classification within each project[3]. The vast majority of Planet Nine participants joined the project within the initial eight hours of heavy participation, with a smaller peak corresponding to the second broadcast event and a third, significantly smaller peak for the third and final broadcast event. Nevertheless, over the course of the first 24 hours, a number of participants made their first classifications outside of the periods corresponding to broadcast events, likely because they had not heard of the project until that point.

[3] Because each project launched ‘from scratch’, every user within Planet Nine and Exoplanet Explorers was considered a ‘new user’

In addition to the live broadcast, dissemination also occurred within the Zooniverse blog, project page and social media accounts, particularly within the first 24 hours of the project and it is likely that this was responsible for the prolonged period of initial activity.

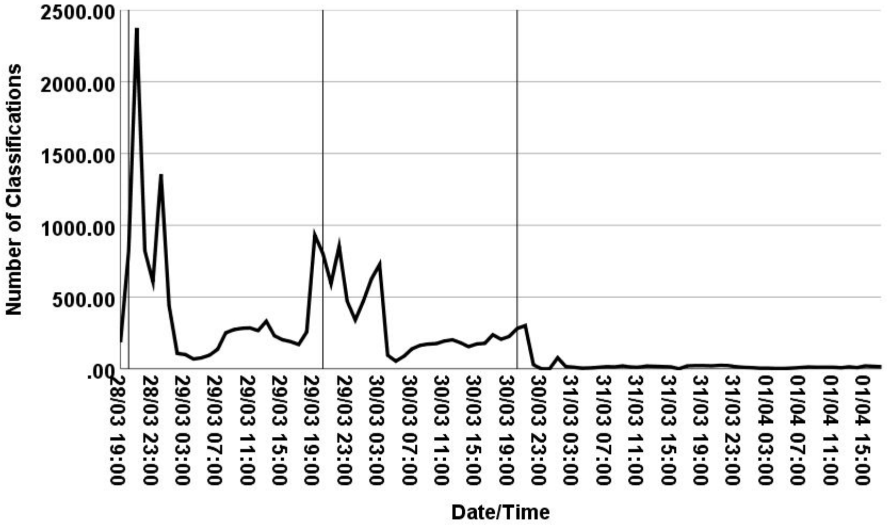


Figure 5 – Line graph showing hourly number of classifications made by a new user within Planet Nine

Within Exoplanet Explorers, however, participation was closely correlated to broadcast events, with limited participants joining the project outside of these points. We note a greater peak following the first broadcast event and a slow decline rather than a sudden drop off – again, we suggest this may have been due to dissemination efforts.

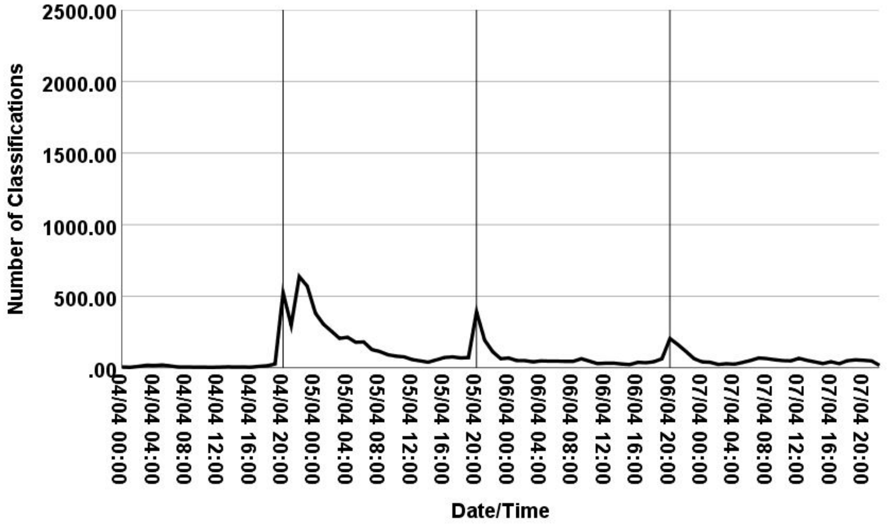


Figure 6 – Line graph showing hourly number of classifications made by a new user within Exoplanet Explorers

3.5 - Talk participation over time

Unlike task participation, discussion participation within Planet Nine shows limited correlation with broadcast events. While each broadcast did see an associated peak in participation, this peak is offset and occurs unpredictably, occurring 4 hours after the first broadcast, 8 hours after the second broadcast and roughly in time with the third. Participation is much more random and rather than diminishing over time, actually appears to increase, up until the final broadcast.

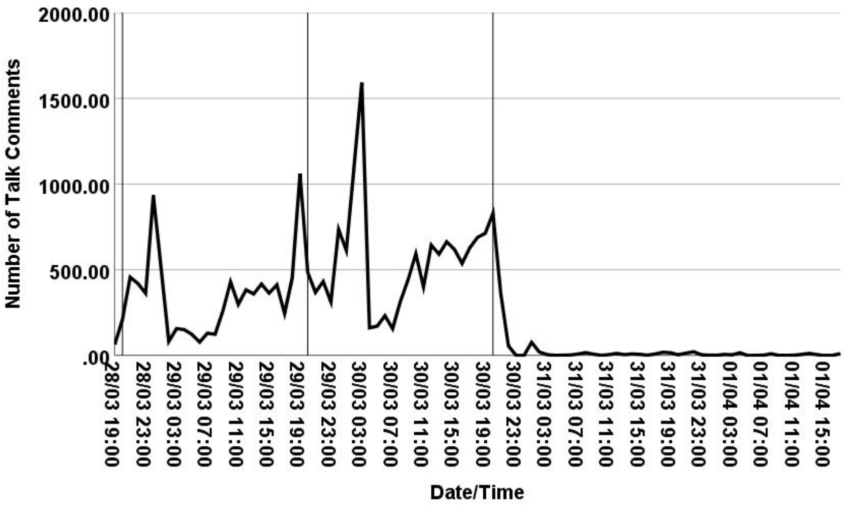


Figure 7 – Line graph showing hourly number of discussion comments posted to Planet Nine's Talk page

Discussion participation within Exoplanet Explorers followed a similar pattern, with peaks in participation following broadcast events, but offset by between one and four hours in each case. There was, however, no clear pattern in terms of increased or decreased participation over time and participation rose and fell seemingly at random. Discussion participation also persisted for some time after the project was officially retired, although the data did not cover a sufficient period to determine whether this phenomenon persisted or whether participation was dropping for a final time at the end of the sampled period.

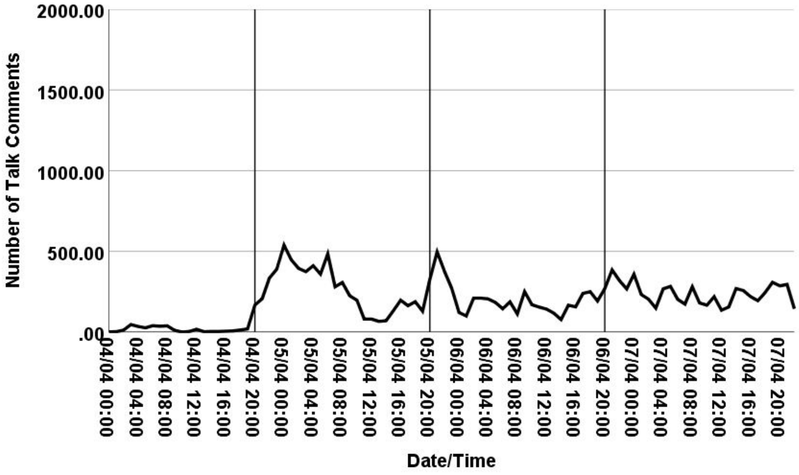


Figure 8 – Line graph showing hourly number of discussion comments posted to Exoplanet Explorer’s Talk page

3.6 – First discussions

Perhaps surprisingly given hourly participation patterns within Planet Nine discussions, participants first discussion comments followed the pattern of first classifications, peaking roughly in line with broadcast events and diminishing over time. There was, however, an element of randomness, with a significant number of participants making their first comments in the 12 hours following the second broadcast event. We believe this is likely due to an eagerness to discuss or contribute to the preliminary results which were presented at this time.

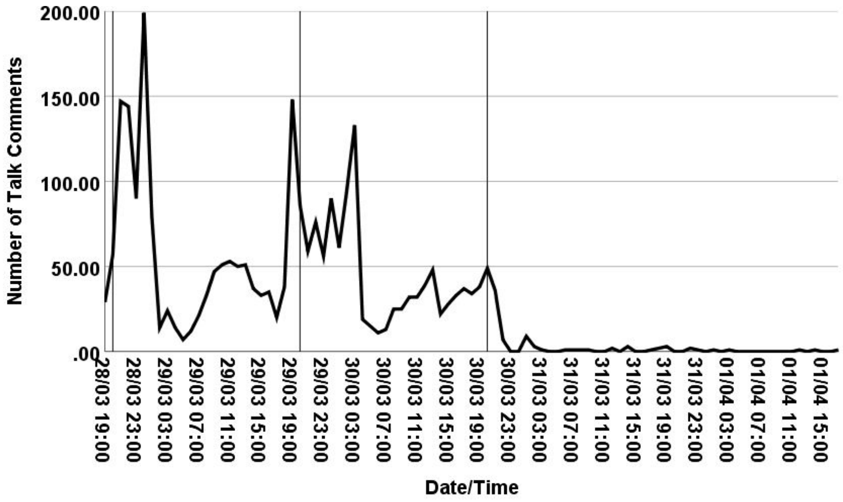


Figure 9 – Line graph showing hourly number of talk comments made by a new user within Planet Nine

Within Exoplanet Explorers, the pattern of first discussion comments also largely followed the pattern of first task contributions. However – and perhaps surprisingly – the highest peak coincided with the retirement of the project, with a high number of comments made at this time and little drop off in participation following this point.

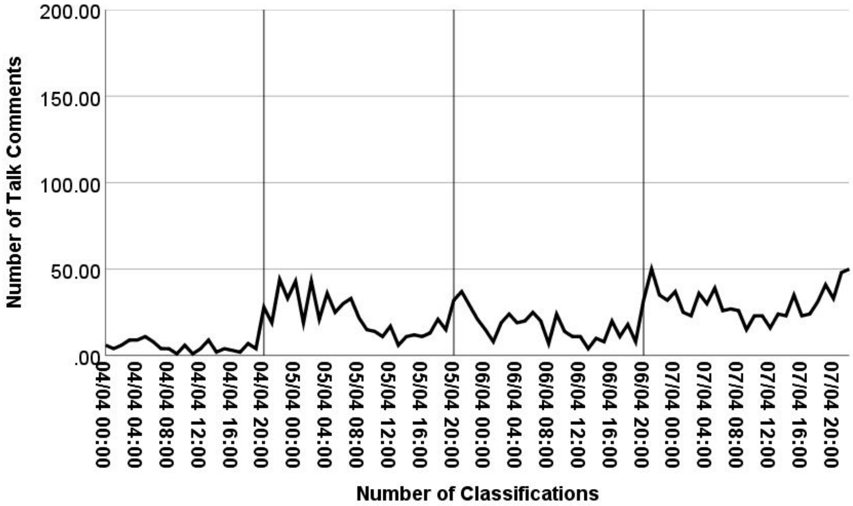


Figure 10 – Line graph showing hourly number of talk comments made by a new user within Exoplanet Explorers

3.7 – Comparison with contemporary projects

We also compared participation within the two projects with participation over the first 72 hours of other projects launched in the three months before and after Planet Nine and Exoplanet Explorers were launched. Notably, contributions to Planet Nine and Exoplanet Explorers are an order of magnitude greater than in other projects launched during the same time period. Exoplanet Explorers and Planet Nine each garnered millions of classifications -- 1,663,459 for Exoplanet Explorers and 4,074,437 for Planet Nine -- while the most popular long-term project launched during this time (Backyard Worlds: Planet Nine) gathered just 475,724 classifications in its first 48 hours. This is equally true of the number of Talk comments made within these projects, with 11,071 comments made in Exoplanet Explorers and 20,967.00 comments made in Planet Nine, compared to just 2,432 comments made in the next highest project -- Backyard Worlds: Planet Nine[4]. Ultimately these figures suggest that participation levels in short-term projects exceed the temporary intense contribution levels of VCS project launches. While this is likely influenced by the use of a live broadcast, it should be noted that a large proportion of the classifications came from outside the broadcast area in both cases (i.e., outside Australia in the case of EE and the UK in the case of Planet Nine) with significant classifications from the US, Canada and Europe. Moreover, in the case of Exoplanet Explorers, participation was seen either side of the broadcasts, suggesting a degree of interest was present prior to – and persisted beyond – the use of the life television broadcast events.

Project	Launch	Total	Median
Astronomy Rewind	22/3/17	45,827	471

[4] In spite of the similar name, this was a distinct project from Planet Nine in terms of focus, task question and design

Backyard Worlds: Planet Nine	15/2/17	475,724	9,226.5
Bash the Bug	7/4/17	14,361	262
Colorado Corridors	2/6/17	13,525	272
Count Flowers for Bees	10/5/17	16,643	234.5
Elephant Expedition	10/5/17	66,764	234.5
Etch a Cell	6/4/17	2,126	39
Galaxy Nurseries	31/5/17	89,769	1,428
Michigan ZoomIn	3/5/17	81,633	1,466
Muon Hunters	28/2/17	407,488	8,489.33
Planet Four Ridges	17/1/17	203,484	3030.50
Plastic Tide	26/4/17	139,336	1488.50
Steller Watch	15/3/17	245,868	3,653.50
Western Montana Wildlife	12/4/17	77,253	3,653.50
Exoplanet Explorers	4/6/17	1,663,459	24,625
Planet Nine	28/3/17	4,074,437	75,095.50

Table 1 – Comparison of project launch date, total classifications received over 72 hours and median hourly classification rate for contemporary projects and EE/P9

Similar results can be seen in the discussion activity for Exoplanet Explorers and Planet Nine when compared to other projects launched at a similar time. Once again, participation was an order of magnitude greater than in the other projects. We do note, however, that astrophysics projects show a generally elevated level of discussion participation relative to other projects within the Zooniverse and so at least some of this increased level of participation is likely related to the domain. There is also a high level of cross-project effects within the Zooniverse[5], where participants from one project contribute to another, which may have further influenced this.

Project	Launch	Total	Median
Astronomy Rewind	22/3/17	631	7
Backyard Worlds: Planet Nine	15/2/17	2,432	47
Bash the Bug	7/4/17	80	0.5
Colorado Corridors	2/6/17	68	0
Count Flowers for Bees	10/5/17	56	0
Elephant Expedition	10/5/17	111	1
Etch a Cell	6/4/17	20	0
Galaxy Nurseries	31/5/17	973	16
Michigan ZoomIN	3/5/17	716	10.5
Muon Hunters	28/2/17	855	15
Planet Four: Ridges	17/1/17	843	13
Plastic Tide	26/4/17	319	5
Western Montana Wildlife	12/4/17	454	6
Exoplanet Explorers	4/6/17	11,071	197
Planet Nine	28/3/17	20,967	406

Table 2 – Comparison of project launch date, total talk comments received over 72 hours and median hourly comment rate for contemporary projects and EE/P9

3.8 Conclusions

From this analysis, we draw three major conclusions which have implications for the use of short-term projects or campaigns in citizen science:

1. Short-term citizen science projects can gather significantly more data than would otherwise be gathered through more conventional project designs/timelines

[5] See for example Luczak-Roesch et al., 2014

2. However, participation will naturally diminish over time and this may occur very rapidly in short-term projects
3. Dissemination plays a key role in managing this diminishing engagement and also attracting participation in a timely manner to facilitate the short-term engagement.

4 Task Difficulty and Clustering

Citizen Science projects and initiatives often include concepts with which citizen scientists may be unfamiliar. As a result, it is essential that volunteers receive adequate training and preparation to complete the tasks prior to carrying them out. Yet training can be unengaging, driving participants to abandon tasks, while also using up valuable time that participants spend on projects, which can otherwise be extremely brief.

In this chapter, we present the findings of a study exploring the impact of clustering tasks according to difficulty on the quantity and quality of submissions made by volunteers. We first group images according to the perceived difficulty identified by volunteers, including an easy, difficult and randomised cluster. We then compare the number of submissions made, the accuracy of submissions and the average time taken to complete classifications between each quantity.

4.1 Methodology

In this experiment, we presented participants with a series of images taken from tweets around hurricanes Irma and Maria and asked them whether the tweet contained any: people, hazards, buildings, important information, resources or evidence of repairs in progress. Despite seeming relatively simple, the task itself was actually rather difficult as the images taken had not been gathered with the intention of facilitating this kind of activity – instead, they were simply images taken around the recovery process by various individuals and shared on Twitter. This meant that in many cases, there were a large number of details to identify or one or more features may have been somewhat obscured.

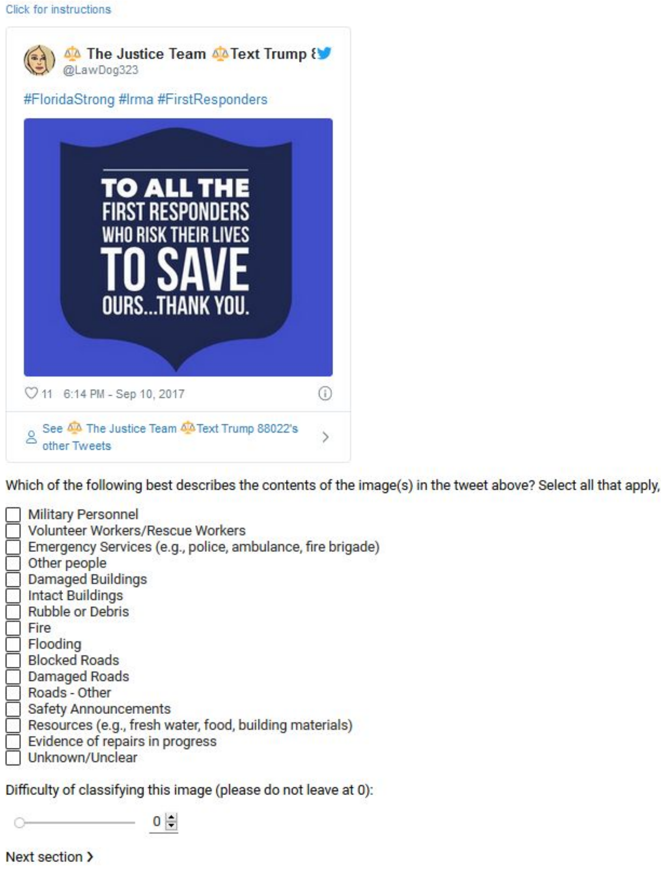
We first ran an experiment to understand how difficult each image was. We presented each image to 5 Mechanical Turk[6] workers and asked them to identify the content of the image, as well as to provide a difficulty score out of 5, where 5 is very difficult and 1 is not at all difficult. The classification interface can be seen in figure 11 below. We then produced a series of three clusters based on a combination of the difficulty scores provided by workers and the percentage agreement of workers regarding content – these clusters included an easy cluster, where scores were low and agreement was high, a hard cluster where scores were high and agreement was low and a third random cluster, where images were selected at random. We then presented these to a new series of workers[7] and asked them to identify the content of the image, with no requirement to indicate a difficulty score.

We then calculated the accuracy of each classification for each image relative to an expert provided gold standard produced through careful manual analysis of the image and conducted a Kruskal-Wallis H test to identify whether there was a statistically significant difference in the accuracy scores for each of the three conditions. We similarly monitored the number of

[6] To minimise differences between the three conditions, we used Mechanical Turk to offer the same flat payment to all workers, so as to avoid the possibility that volunteers from one condition may have different motivations than those in the other two conditions

[7] Workers who helped to classify the images initially were not eligible to participate in this second experiment

classifications made by each worker and conducted a Kruskal-Wallis H test to similarly identify whether a statistical significance in engagement was observed in response to image difficulty.



Click for instructions

The Justice Team Text Trump
@LawDog323

#FloridaStrong #Irma #FirstResponders

TO ALL THE FIRST RESPONDERS WHO RISK THEIR LIVES TO SAVE OURS...THANK YOU.

♡ 11 6:14 PM - Sep 10, 2017

See The Justice Team Text Trump 88022's other Tweets

Which of the following best describes the contents of the image(s) in the tweet above? Select all that apply,

☐ Military Personnel
☐ Volunteer Workers/Rescue Workers
☐ Emergency Services (e.g., police, ambulance, fire brigade)
☐ Other people
☐ Damaged Buildings
☐ Intact Buildings
☐ Rubble or Debris
☐ Fire
☐ Flooding
☐ Blocked Roads
☐ Damaged Roads
☐ Roads - Other
☐ Safety Announcements
☐ Resources (e.g., fresh water, food, building materials)
☐ Evidence of repairs in progress
☐ Unknown/Unclear

Difficulty of classifying this image (please do not leave at 0):

0

Next section ›

Figure 11 – Image classification interface showing image extracted from tweet, question with answer options and difficulty selection.

4.2 Results

For tweet classification, we found a highly statistically significant difference between the random, easy and difficult conditions (p=0.008). However, when conducting a Dunn's test to compare accuracy between pairings, we observed no statistically significant difference between the random and easy and easy and difficult conditions. We did, however, observe a statistically significant difference between the *random* and *difficult* conditions (p=0.008), suggesting that the most 'difficult' images as identified by workers were indeed harder than easy and random images.

However, when comparing the *number* of contributions completed by workers, we found no statistically significant difference between the three conditions (p=0.14). This suggests that easy tasks are no more or less engaging than difficult tasks or those that are randomly assigned. We also note that approximately one third of participants completed just one submission and this was observed between all three conditions, suggesting that participation followed the long-tail phenomenon common to citizen science and online participation and that this phenomenon was not significantly impacted by filtering tasks based on ease.

4.3 Conclusions

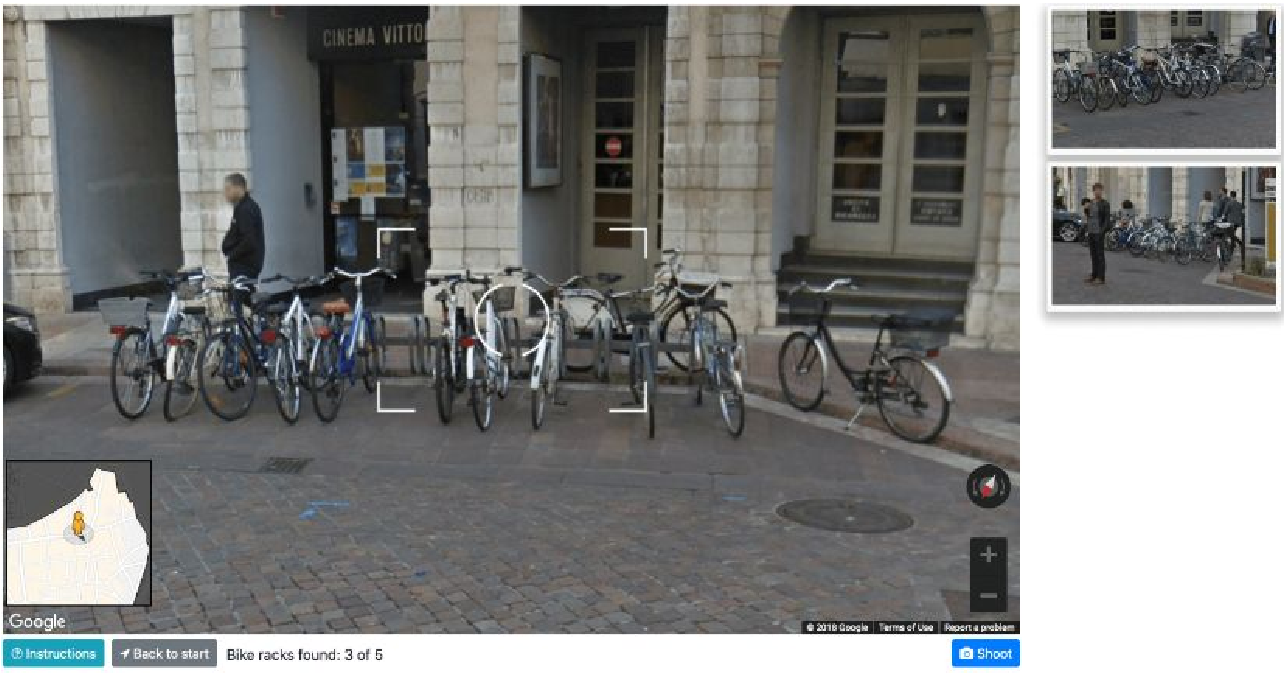
From this experiment, we note that clustering images according to perceived difficulty had no effect on the engagement of participants. We would recommend that presenting images at random is the most effective way to gather high quality data, particularly when accounting for the time and effort required to judge and cluster images in this manner.

5 Virtual City Explorer

With the disruption and restrictions posed by the novel coronavirus, the ability and willingness of volunteers to travel to gather data has been significantly diminished. In this section, we present the Virtual City Explorer, which allows participants to explore a virtual environment based on particular locations, to search for and identify items within street view images. We use the example of bike racks as easy to identify items which are common to many cities, but such capabilities could be extended to other contexts including sources of pollution, within the context of ACTION.

5.1 Presenting the Virtual City Explorer

From a participant’s perspective, the Virtual City Explorer (VCE) is an exploration interface in which they can drag and click to move themselves about through a virtual environment made up of images gathered from Google Street View (see figure [12] below). The participant is able to move around any area featured on Google Street View or – if so desired – an external requester (for example, a citizen science project administrator) may define an explicit area in which the participant can be permitted to explore. The participant is then randomly placed at a geographic point within this area or may be placed at a specific area of the requester’s choosing or based on gaps in existing data. Requesters may ask participants to explore a certain proportion or region of the larger defined area, identify a certain number of potential bike racks or otherwise limit tasks to simplify completion for participants.



Bike racks found: 3 of 5

Figure 12 – VCE exploration interface, showing the ‘shoot’ capture function.

As the participant moves around the images they can look for specific points of interest. Here we use the example of a bike rack. As the participant moves, they view and analyse each image to see if a bike rack is present. If one is not present, they can move to a new point in the predefined area to continue searching. Conversely if they *do* find a bike rack, then they are able to record the geolocation of that bike rack following a simple process that we describe as similar to taking a picture. Using a specific option within the VCE interface, the worker is able to zoom, move their cursor and display and save a 'picture' of the bike rack that they have found. We ask them to take three images of the bike rack in question and use triangulation of these results to determine an approximate GPS coordinate location that the bike rack is likely to occupy. The external requester is then able to use the images taken within the VCE and the approximate location for further analysis as required.

5.2 The Taboo System

One potential risk of the methodology followed by the VCE is a large quantity of erroneous or otherwise low value data. Many participants may identify the same bike rack or bike racks and in the absence of restrictions within the task itself, may assume their task is complete and leave, leading to a large number of classifications for a small area of the map. We therefore developed a taboo system, whereby requesters may mark certain known objects – for example, bike racks they are already aware of – within the VCE as pre-classified, preventing participants from submitting classifications of that particular object. This can also be conducted dynamically, such that when a certain threshold of submissions has been reached, a certain object or small area (e.g., coordinate) within the VCE becomes taboo and no further submissions can be made for that particular object.

For example, consider a bike rack that is positioned close to the point at which a requester asks a participant to start searching. This rack is likely to be classified by most – if not all – of the participants contributing to the experiment. Under the taboo system, the requester can either mark this particular bike rack as taboo prior to starting or have the system mark the bike rack as taboo as soon as a pre-defined number of participants mark the bike rack.

5.3 Completeness (accuracy)

To judge the completeness and therefore accuracy of the VCE approach, we compared the work of recruited crowdworkers to 4 quasi-gold standard datasets, using two geographical areas – one representing Trento in Italy and another representing Washington in the US. These two areas were chosen in part because of their unique differences, with the Trento area featuring many small, branching streets, while the Washington area is largely grid based and rigidly laid out. In the case of the Trento area, we used a dataset provided by the Trento municipality and the Open Street Map citizen science/volunteer geographic initiative. In the case of Washington, we used the Rackspotter service launched by the local government and the Open Street Map volunteer geographic initiative.

Completeness statistics can be seen in table 3 below. In the case of Trento and when compared to the municipality-provided dataset, our approach found 27 of 39 racks, missing a total of 12 bike racks. However, our approach identified 5 bike racks not recorded by the municipality and indeed, analysis of images and data submitted by the workers suggests that these are indeed racks which had been missed by the municipality. Similarly, when compared with the Open Street Map service, our approach found 27 of a total of 52 bike racks, missing 35 bike racks but identifying 8 new bike racks that had been unidentified. We propose that the reason for this low completeness in the Washington area was due to the sheer density of bike racks within the city in conjunction with the reasonably low number of workers recruited for the study.

	Trento		Washington	
Reason	Municipality	OSM	Rackspotter	OSM
Missed	11 (64.7%)	11 (31.4%)	92 (89.3%)	4 (57.1%)
Absent	2 (11.8%)	8 (22.9%)	4 (3.9%)	1 (14.3%)
Unreachable	2 (11.8%)	8 (22.9%)	2 (1.9%)	0
Missed cluster	2 (11.8%)	4 (11.4%)	5 (4.8%)	2 (28.6%)
Triangulation impossible	0	4 (11.4%)	0	0
Total	17	35	103	7

Table 3 – completeness of VCE for Trento and Washington in comparison with gold standard datasets

5.4 Effectiveness of the Taboo Strategy

We conducted an experiment to ascertain the effectiveness of the taboo strategy. For each of the two areas, we conducted two experiments, asking workers to identify bike racks within the VCE using a crowdsourcing platform. In one condition, however, workers were given a basic task, with no taboo. In the other, workers were given the taboo condition and had to find 5 bike racks to complete the task. We conducted this for 60 workers in each case.

An analysis of the number of bike rack detections (i.e., classifications) for Trento can be seen in figure 13 below. Firstly, we note that setting a taboo leads to a significant reduction in the number of classifications made, likely because the available pool of bike racks is also significantly smaller, with this effect growing as the number of workers recruited grows. Nevertheless, we note that the number of *confirmed* bike racks – that is, the number of accurately identified unique bike racks is actually greater in the taboo condition, once a threshold of approximately 25 workers is reached. In the Washington area, however, this threshold is much lower and the taboo condition outperforms the basic condition much more readily (see figure 14).

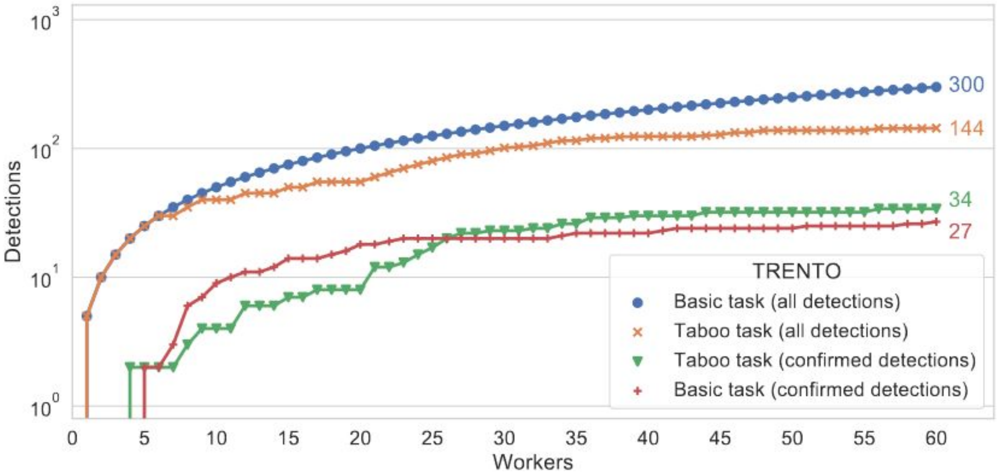


Figure 13 – Number of bike racks detected according to number of workers recruited for basic (no taboo) and taboo task variants for Trento.

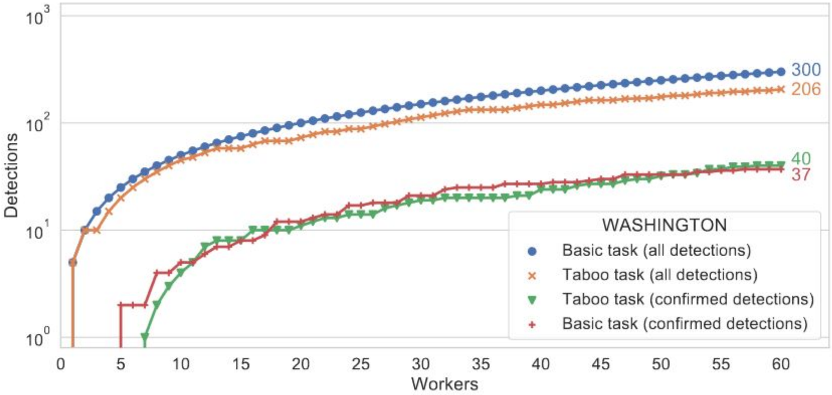


Figure 14 – Number of bike racks detected according to number of workers recruited for basic (no taboo) and taboo task variants for Washington.

We further analysed the number of detections received per bike rack for each area in the two conditions. For Trento (figure 15), two bike racks received 5 detections, 6 received 4 and the remaining 26 received three confirmations, allowing for redundancy and easier, more accurate identification of the coordinates that the bike racks occupy. In the case of the basic condition, however, the number of classifications varied, with the majority of the bike racks receiving far in excess of 5 detections, representing a large volume of wasted effort and resources. A total of 7 bike racks from the taboo condition were not identified at all in the basic condition.

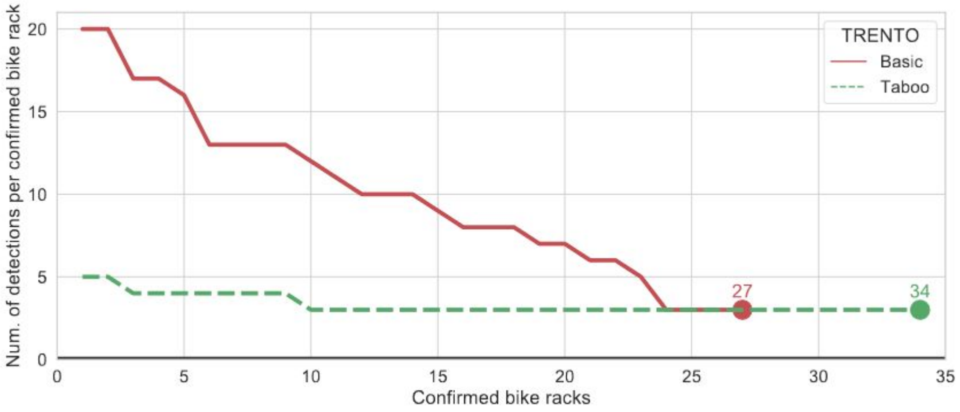


Figure 15 – Number of detections per confirmed bike rack count for Trento

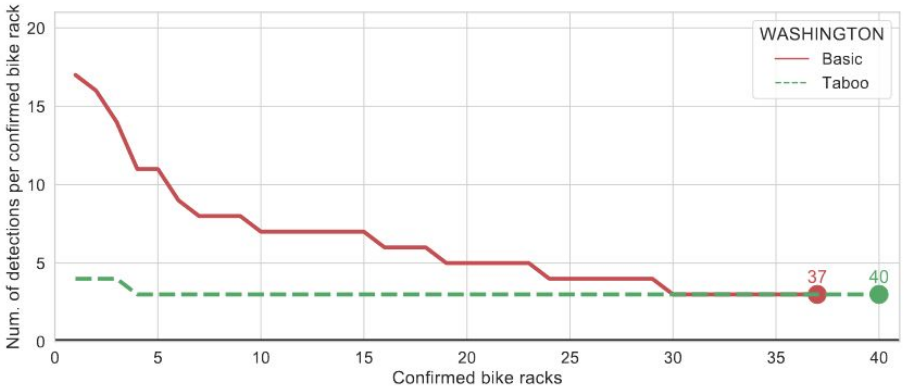


Figure 16 – Number of detections per confirmed bike rack count for Washington

The number of detections within the Washington area (figure 16) was broadly similar for the taboo condition, with most receiving three detections. However, the basic condition was generally less effective than in the Trento condition, with a significant proportion of the workers failing to identify many of the bike racks. Nevertheless, the basic condition identified 37 of the 40 bike racks identified in the taboo condition, although again a large volume of time and resources would be effectively wasted by using this method.

5.5 Conclusions

The VCE is an effective system for gathering data on the geolocation of static objects recorded within Google Street View, offering requesters a great deal of control over the area and quantity of data that they wish to explore. The taboo system offers a relatively simple, yet nonetheless effective and accurate method of quality assurance. We further note that the taboo system could be a useful strategy in encouraging more diverse data collection processes in many of the projects analysed within D 5.1 which use more traditional offline or mobile data gathering strategies.

6 Guidelines and Recommendations

In this section we present our final set of recommendations and guidelines for task design in citizen science. We first begin by recapping the recommendations made in D 5.1, before formulating a final set of guidelines which incorporate the research findings of D 5.1 and 5.2.

6.1 Initial Recommendations

In the initial guidelines formulated through D 5.1, we identified 5 preliminary recommendations for the design of citizen science tasks. Here we reiterate those recommendations and re-evaluate them in the light of the research carried out within this deliverable.

1. **Consider volunteer participation patterns**
   Our findings within D 5.1 demonstrated that data collection in a pollution context was often best achieved by exploiting ubiquitous devices such as smartphones and sensors, to facilitate ad-hoc data gathering. We further note from our findings in D 5.2 that *short-term* engagement can be an effective way to gather data for event-based or temporary initiatives.
   It is important to carefully consider the level of activity expected of volunteers, the context in which they will gather data and what tools and materials they will have with them at the time. This should be developed in response to the research problems which are defined and the data to be collected.
2. **Consider project aims when defining volunteer engagement**
   Even otherwise very similar projects can have very different aims. A project which aims to raise awareness and promote education is much less likely to be concerned with the validity of gathered results than an investigative, research-driven project. At the same time, educating volunteers may be much more easily achievable in restricted, one-on-one or small group sessions than remotely.
3. **Match task types to expected crowd size**
   Some citizen science tasks can be completed much more quickly, easily and rapidly than others and require much fewer participants. If a project aims to simply gather examples of plastic pollution in a small area, then a simple cataloguing task is fine. On the other hand, if the project requires information which volunteers may be unfamiliar with, then a more supplementary-based methodology may be necessary, in which case the size of crowd may become unmanageable. For every volunteer contributing images or records, an equal number of volunteers describing those records is required. Carefully consider the types of task and activity that will be carried out and how these might align with the size of the crowd involved.
4. **Diversify input devices**
   Truly accessible citizen science should be developed to allow distinct data-input devices and technologies, including mobile and desktop devices and distinct browsers. The importance of this recommendation was reiterated by the project administrators who took part in our interview process.
5. **Open up the scientific process**

Participants can offer potentially vital knowledge and experiences which project administrators have overlooked. Moreover, allowing volunteers to participate in multiple stages of the scientific process promotes learning and can strengthen volunteers' intrinsic motivations to participate in a project, leading to increased contributions and appreciation of both the project and scientific research.

6.2 Final Recommendations

1. **Choose a suitable approach**
   Citizen science is most suitable for gathering or analysing research data where participants can be motivated to engage without requiring payment or rewards -- i.e., where a task is inherently engaging, supports research for the public good or can be designed in such a manner that they are inherently fun and enjoyable (e.g., Games With A Purpose). Prior to selecting citizen science as an approach, it is important to consider whether the task and research aims align with these goals. If not, consider alternative methods such as paid microtask crowdsourcing or more traditional, lab-based or field studies.
2. **Formulate a suitable problem**
   Consider the task that is to be presented to volunteers. While overarching research questions made be broad, citizen science works best when the questions presented to volunteers are specific and discrete, with limited ambiguity. Think about how the task maps to the activities that volunteers will complete and what resources volunteers will need as well as any specificities and restrictions that will define the task -- for example, will the task require being present in a specific location? If so, the resources, input devices and task steps will be different to a task that can be carried out at home.
3. **Account for trade-offs**
   The use of citizen science entails inevitable trade-offs between the quantity of data to be gathered, the speed at which data is to be gathered and the accuracy of the gathered data. Prior to commencing the research process, it is essential to consider and identify which of these factors is to be prioritised and take appropriate steps to safeguard this factor, while taking steps to mitigate threats to the additional trade-off factors. For example, if a project is to emphasise *accuracy and quality* of submissions, the task completion time is likely to increase and this can limit engagement. It is important to then streamline and simplify the task completion process to support faster data gathering or otherwise take steps to encourage engagement to account for these trade-offs.
4. **Account for technology**
   As outlined in the initial guidelines reiterated above, it is important to consider the technology and software that volunteers are likely to use to complete your task. Does the task need to support both mobile and desktop devices or is the task designed to be completed outside of the home? Does the task support multiple browsers? Wherever possible, support diverse technologies to lower any barriers to entry. If participants cannot access your task, then they are unlikely to put in the effort to overcome these barriers and continue contributing. If these barriers are technological, it is also possible that volunteers will not be able to overcome these barriers or will not know how. While it may require a significant commitment of time and resources, it is essential that these issues are resolved

upfront and prior to task publication, as participants who encounter these barriers may be unwilling or unable to return to tasks and/or may otherwise remain unaware that these issues have been resolved.

5. **Provide Context**
   Citizen science tasks can often be designed and implemented in such a way that they are trivial and simple for volunteers to complete. This is essential for encouraging accessibility and gathering high quality data, but can obfuscate or trivialise their research value, with the potential to harm volunteer engagement. Tasks, project resources and educational resources should provide additional context on the value that volunteer contributions pose for the research process.
6. **Provide Feedback**
   While citizen science tasks are generally designed to be easily understood and completed by all participants, not all projects are able to achieve this. Moreover, even where tasks are otherwise easily understood, participants want and need feedback on the accuracy of their responses and the value of their contributions to scientific research. Providing feedback to participants -- either within tasks or through communication features such as forums -- can encourage participant engagement with citizen science.
7. **Solicit Feedback**
   Tasks should not necessarily remain static. The design process involves a number of assumptions and trade-offs which may not align with the expectations of participants. Soliciting feedback from participants is key to ensure that the needs of all stakeholders are met, with the potential for increased task quality and engagement, as well as volunteer engagement.
8. **Avoid Ambiguity**
   While the requirements and processes involved within a task may be clear to task designers, these do not necessarily align with the understanding and motivations of volunteers. To avoid misunderstanding, miscommunication and other issues, avoid ambiguity wherever possible. Support participants through the task process by using discrete, clear questions and limit the need for autonomy and personal judgement. Consider offering multiple choice answers rather than free text responses, for example.
9. **Consider Time-scales**
   While citizen science is an effective way to gather large volumes of data for scientific purposes, volunteer engagement is sporadic, asymmetrical and often brief. It can therefore take a significant amount of time to gather larger datasets. This can be offset by focusing on restrictive, limited-time activities such as BioBlitzes, where volunteers are asked to gather or analyse data over a short period of time. While this approach can be very effective, it is less effective for tasks with more longer-term aims such as public engagement and education. It is essential to consider the implications and long-term aims of the approach to be used and which factors are most important -- is it essential to gather data quickly or in large quantities? Do the research aims warrant longer term engagement and community building or is one off engagement desirable?
10. **Do not pre-prepare input data**
    Scientific research often entails concepts or data with which volunteers may be unfamiliar. While these data must be selected carefully, our findings suggest that there is no significant value to be gained from the pre-formulation of input data, for example grouping data

according to perceived difficulty. Instead, presenting tasks at random requires minimal time and resources, with no negative impacts on participant engagement.

7 Conclusions

In this deliverable, we have presented recommendations by citizen science and crowdsourcing researchers. We further presented research exploring the role that the design and implementation of tasks plays in influencing participant engagement in citizen science activities, with associated conclusions and recommendations for designing, refining and managing citizen science activities.

While much of our findings have focused on designing for accessibility and engagement, we note also that designing to ensure high quality results is an important aspect of citizen science, particularly given that any results must be used for subsequent research, as well as for use in policy and decision making. We would therefore advise that in addition to the recommendations set out in this deliverable and D 5.1, stakeholders consider and account for the need for quality assurance. Such methods will be addressed separately in an additional deliverable – D 5.3, which will present final guidelines and methods for quality assurance which should be adopted and considered in conjunction with the task design strategies described within this deliverable.

8 References

Jun, Eunice, Morelle Arian, and Katharina Reinecke. "The potential for scientific outreach and learning in mechanical turk experiments." *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 2018.

Luczak-Roesch, Markus, et al. "Why won't aliens talk to us? Content and community dynamics in online citizen science." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. No. 1. 2014.

Maddalena, Eddy, Luis-Daniel Ibáñez, and Elena Simperl. "Mapping Points of Interest Through Street View Imagery and Paid Crowdsourcing." *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.5 (2020): 1-28.

Reeves, Neal T., and Elena Simperl. "Efficient, but Effective? Volunteer Engagement in Short-term Virtual Citizen Science Projects." *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019): 1-35.

Simperl, Elena, et al. "Is virtual citizen science a game?." *ACM Transactions on Social Computing* 1.2 (2018): 1-39.