CITIZEN SCIENCE:
THEORY AND PRACTICE

Phillips, C., et al. 2018 Assessing Citizen Science Participation Skill for Altruism or University Course Credit: A Case Study Analysis Using Cyclone Center. *Citizen Science: Theory and Practice*, 3(1): 6, pp. 1–13, DOI: https://doi.org/10.5334/cstp.111

RESEARCH PAPER

# Assessing Citizen Science Participation Skill for Altruism or University Course Credit: A Case Study Analysis Using Cyclone Center

Christopher Phillips*, Dylan Walshe*,§, Karen O'Regan*, Ken Strong*, Christopher Hennon†, Ken Knapp‡, Conor Murphy* and Peter Thorne*

A common challenge in citizen science projects is gaining and retaining participants. At the same time, the tertiary education sector is constantly being challenged to provide more meaningful and practical work for students. Can participation in citizen science projects be used as coursework with real practical experiential-learning benefits, without affecting the citizen science project outcomes? We seek to begin to answer this question via a case study analysis with Cyclone Center (CC), which asks participants to classify tropical cyclone characteristics through analysis of infrared satellite imagery. Skill of individual users has previously been shown to be obtainable once classifiers have looked at approximately 200 images using an expectation-maximisation likelihood approach. We use skill scores to determine if participation for course credit or altruism influenced skill for volunteers and students from two universities under three increasingly complex categories of classifications (eye or no eye; stronger, weaker, or the same; and which of six fundamental storm types). A bootstrap resampling approach was used to account for discrepancies between sample sizes. Overall, there is limited evidence for substantive differences in classification performance between credit awarded and altruistic participants, with only one finding of significance at $<p = 0.05$ (Maynooth University showing lower mean agreement with the volunteer consensus on eye vs. no-eye). There is evidence that integrating participation into a larger assessment that requires the students to show understanding of the project may reduce a low-skill student tail. Furthermore, students' perceptions of the coursework compared to more traditional assignments were overall favourable. These findings, if replicated for other citizen science projects, open up possible avenues to more generally increasing participation in, and exploitation of, citizen science projects in the academic sector.

**Keywords:** participation; academia; credit; altruism; climate; volunteers; tertiary education

## Introduction

### *Citizen science and the climate research sphere*

Various definitions of the term citizen science exist. For example, citizen science is defined by the Oxford English Dictionary (2017) as "the collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists." Roy et al. (2012) summarise that "Citizen science is increasingly used as an overarching term for the many varied approaches utilising volunteers in science, from active participation in hypothesis-led science through passive movement of sensors; from addressing highly focussed questions to educational exercises generating data of little scientific value; from using people as data collectors to participants forming the projects, assessing the data, and using the information themselves." Regardless of the precise definition, citizen science has become an increasingly integral part of many aspects of data collection, creation, and analysis, with numerous successful projects operating across a vast array of science and humanities disciplines that have proliferated with technological advances (e.g., Zooniverse.org). Numerous areas of scientific investigation either require, or substantively benefit from, human input such that the proverb "many hands make light work" applies.

In weather and climate research, citizen science approaches are still relatively underdeveloped compared to other research fields, e.g., astronomy or ecology (Muller et al. 2015). Several projects do exist, however. These include substantial networks such as the Cooperative Observer Program (NOAA 2017), the Community Collaborative Rain, Hail and Snow Network (Reges et al. 2016), and Weather

* Maynooth University, IE
† University of North Carolina at Asheville Asheville, NC, US
‡ NOAA's National Centers for Environmental Information, US
§ Department of Biological Sciences, School of Natural Sciences, University of Limerick, IE
Corresponding author: Peter Thorne (peter@peter-thorne.net)

Observations Website (WOW, Met Office 2017), all of which rely upon citizen observations. Climate model simulations are also run by citizen volunteers using spare CPU time (Allen 1999), resulting in numerous published large ensemble analyses (climateprediction.net 2017). Several data rescue projects also have been undertaken, which have rescued millions of old marine (oldweather.org and offshoot projects, 2017) and land (weatherrescue.org, 2017) meteorological observations. Another project is cyclonecenter.org, which is the focus of the present analysis.

### *Historical and ongoing use of citizen science in formal educational settings*

Universities constantly strive to provide novel and informative experiential learning experiences that have meaningful outcomes for students. In this context, citizen science projects potentially can provide valuable win-win propositions, affording structured participation for students with long-term scientific benefits. Participation to date has largely been via programs set up or run by the universities themselves (Mitchell et al. 2017; Oberhauser and LeBuhn 2012; Karlin and De La Paz 2015; Ryan et al. submitted). These programs have provided useful data that have led to published analyses. Both Mitchell et al. and Ryan et al. point to positive student views and valuable learning outcomes that would have been hard, if not impossible, to otherwise achieve in a traditional classroom setting. In all cases, students have built upon the results of a previous cohort, leading to a real sense of ownership by the students.

An alternative engagement route is participation in an existing citizen science project that is open to broad participation by interested members of the public. For the past several years, both the University of North Carolina Asheville and Maynooth University have set, and continue to set, course assignments that involve active participation in Cyclone Center. To our knowledge this is a less common approach to citizen science engagement by university students. However, such engagement raises legitimate concerns about whether student participation, if significant, could inadvertently impact the host project. Various scenarios can be conjectured which may lead either to low-skill or high-skill biases in the data, which could need to be controlled for in some manner. This study aims to assess this issue for the specific case of Cyclone Center.

### *Cyclone Center background*

Cyclone Center (henceforth CC) was developed by a team of scientists at the University of North Carolina Asheville (UNCA), NOAA's National Centers for Environmental Information (NCEI), and the Cooperative Institute for Climate and Satellites – North Carolina (CICS-NC). The project was implemented by Zooniverse, an online citizen science project resource that started out looking at astronomical problems but has since broadened its project portfolio to include environmental and humanities-based projects. The aim of CC is to provide a long-term reanalysis data set over the satellite era (since about 1980) of tropical cyclone characteristics (primarily intensity) which is globally homogeneous over both space and time (Hennon et al. 2015).

Observing tropical cyclones is difficult. Long-term direct observations are limited to the North Atlantic basin near North America, where regular aircraft reconnaissance is flown. Elsewhere, cyclone intensity–usually defined as maximum sustained wind–is almost exclusively inferred from cloud structures apparent in satellite imagery. A set of rules has been developed (Dvorak 1984) which relate cloud top temperature and structure to storm intensity and type. That approach would seemingly suffice to create a long-term homogeneous record. However, because many different agencies analyse storms in subtly different ways across the various ocean basins, and both the guidance for analysis techniques and the experts undertaking analyses have varied through time, using the resulting analyses as unvarnished "truth" is difficult (Knapp and Kruk 2010; Ren et al. 2011).

CC uses the voluntary contribution of citizen scientists to classify tropical cyclones from satellite images, derived from hurricane satellite (HURSAT) imagery (Knapp and Kossin 2007). The underlying rationale is that a crowdsourced set of analyses, using volunteers who perform many independent assessments of each image under a consistent analysis approach, can be used to create a homogeneous and objective set of records across ocean basins and over time. To this end, the interface of CC is based on a modified version of the Dvorak technique, which attempts to estimate tropical cyclone intensity and development from satellite infrared and visible imagery (Dvorak 1984; Velden et al. 2006). CC uses the Infrared version of the technique exclusively for simplicity and to retain homogeneity. Participants are shown a series of images from a given storm and asked to provide answers about their personal interpretations of the storm type, strength, and progression (cyclonecentre.org 2017). Questions such as "Pick the storm image that appears stronger" and "Pick the cyclone type, then choose the closest match" are put to the classifiers, who are then asked to choose from a number of options. Classifiers base their choices on a number of static images provided. These images act as a guide to help the classifier select the correct storm type: eye (EYE), embedded center (EMB), curved band (CBD), shear (SHR), post tropical (PT) and no storm (NS) scenarios (**Figure 1**).

Cox et al. (2015) undertook a comparative analysis to recognise better and lesser performing projects involved in Zooniverse in terms of participation and volunteer retention rates. CC was found to be an outlier, due to the relative complexity of the tasks involved compared to other projects. Even though CC employs a simplified interface with the Dvorak technique, each image still requires more in-depth analysis than is the norm for Zooniverse-hosted citizen science projects. One of the main challenges for projects such as CC is maintaining participant interest. Infrequent and scarce volunteer contributions pose an issue for the collection and cohesion of large citizen science data sets, because the skill of individual low-volume users is hard to assess. Across a broad range

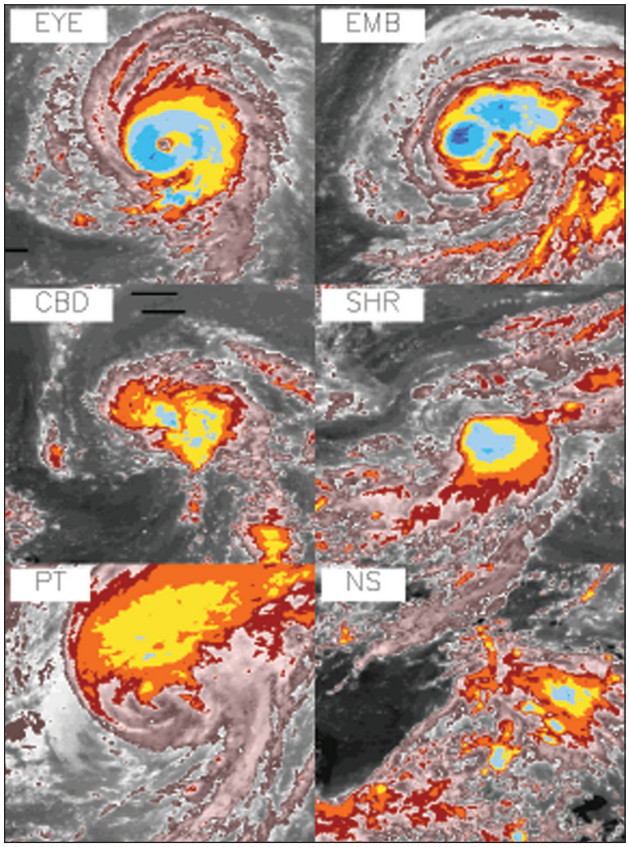


**Figure 1:** Sample images for (top left to bottom right): eye (EYE), embedded center (EMB), curved band (CBD), shear (SHR), post tropical (PT) and no storm (NS) scenarios taken from the typological guidance available to users as they classify on cyclonecenter.org.

of citizen science projects, relatively few participants contribute the bulk of the input. Franzoni and Sauerman (2014) explored some mechanisms that could increase contribution to a citizen science project, including "supporting a broader set of motivations." One such motivation could be the receipt of course credit for participation. To our knowledge, however, little if any assessment has been undertaken to uncover whether project outcomes are affected by participant status—that is, whether participation is undertaken on a voluntary basis or for tertiary (or earlier stage) education course credit. Ostensibly, participants motivated by receiving course credit would constitute a pool of medium-to-high engagement users possessing academic credentials and experience sufficient to provide valuable contributions.

Assessing the impact of different participation pathways first requires a robust quantitative measure of participant skill. Knapp et al. (2016) have developed and documented an Expectation Maximisation (EM) algorithm that can appropriately assess the skill relative to consensus of individual classifiers in CC. The ultimate aim of the EM algorithm is to identify the "true" classification using the range of available answers and the tendencies of each classifier who contributed to its analysis. Determining a skill score requires approximately 180 classifications to be conducted by each classifier (Knapp et al. 2016). Therefore, if more participants contribute larger numbers of classifications, skill can be determined by the EM algorithm which, in turn, would benefit all future classifications.

### *Historical assessments of participant motivation impacts*

Motivations of both altruistic and credit-awarded participants are difficult to gauge. Raddick et al. (2010) concluded that altruists do not have only one motivation for project participation. The top two reasons for the participants they studied were "interest" in the project's subject matter (46%) and "contribution" to science (22%), while "learning" was a much lower motivation (7%). The potential of participation in research to benefit the education of students has been discussed by Padilla-Walker et al. (2005), whose study showed that students who participated in research during their education received higher academic performance scores than students who did not participate.

This study addresses whether course credit versus altruistic involvement has any impact on citizen science project results, using the case study of CC. The closest analogy to our study is a suite of analyses of the distinctions in performance of work done by both paid and unpaid crowdsourcing participants. We stress, however, that substantive differences exist between participating for pay and participating for academic credit, which probably limits what can be inferred by directly comparing the two types of cases. Thus, while the findings of studies focused on paid versus unpaid participants provide potentially useful comparisons, direct comparisons should be treated with caution.

Mao et al. (2013) concluded that both paid and unpaid workers were found to have similar performances when working on the same task, especially in terms of their accuracy scores. However, they found that the payment scheme benefitted the work to be done, because payment guaranteed higher rates of participation without hindering accuracy scores. Mason and Watts (2009) found increases in the quantity of completed tasks due to the presence of payment schemes, and that these incentives had no influence on accuracy. Rogstadius et al. (2011) also found that payment schemes did not influence performance and output accuracy. Thus, in general, no negative impacts have been found in comparisons of paid versus altruistic participation.

Our study was developed to determine whether a similar pattern holds for altruistic participation versus participating for academic credit. Code and data used in our analysis are available in an appendix (see supplemental data) and also electronically via https://www.maynoothuniversity.ie/icarus/icarus-data (Maynooth University 2017).

## Data and Methods

This study was originally undertaken as a piece of course-graded assessment by the first four authors as part of their Masters in Climate Science at Maynooth University. Co-author Knapp provided skill score information arising from the method described in Knapp et al. (2016) to co-author Thorne, who anonymised the data so that no

**Table 1:** Summary of the participant groups. MU is Maynooth University; UNCA is University of North Carolina Asheville.

| | MU | UNCA | Volunteers |
|---|---|---|---|
| Background | Geography Students (BSc, MSc and PhD) | Meteorology Major/Minor Students | Unknown |
| Details of training | 1 hr lecture<br>1 hr tutorial<br>Information available on the website | 1 hr lecture<br>Information available on the website plus 20 minutes of guided practical | Information available on the website |
| Motivation | Core Credit (20% of assessment marks for completing 500 images), required to complete essay for 80% marks balance (assessment worth 50% of overall course marks) | Extra Credit – 2% max (classifying 1500 images equated to 2%) | No known credit basis for involvement (Altruistic) |

possible way of ascertaining individuals remained. Two of the masters students had previously undertaken the undergraduate course outlined in the next sub-section.

### *The three groups*

Three groups of participant classifications were analysed for this study as follows:

*MU – Maynooth University*. A group of 25 undergraduate, masters, and PhD students, all with some prior academic knowledge of climatology and tropical cyclone behaviour, participated in slightly different ways. The undergraduate students undertook classification of tropical cyclone imagery as part of the module "Climate Science at the Public Interface" for 50% of their marks for the module, sectioned into 20% for classifying 500 images and 80% for writing an essay. The aim of the essay was to track the evolution of a particular storm from the individual set of images the student classified. Students were required to understand and discuss their chosen storm, which was intended to guard against the risk of students randomly clicking on CC boxes to get to 500 classifications for credit. Masters students were graded similarly, but the grades on the module were worth 66% of the final course marks. For the PhD students, classifying of cyclones constituted the sole means of gaining course credit.

*UNCA – University of North Carolina Asheville*. A group of 28 undergraduate students participated in classifications to earn extra credit to supplement their Major/Minor meteorology degree. The extra credit was awarded for a particular course (e.g., Intro to Meteorology, Tropical Meteorology). Students received one bonus course point for each 75 images classified up to cap of 20 points (1500 classifications), with a total course credit equal to 1000 points. These students had no associated essay or other course assessment.

*Volunteers* – A group of 244 members of the public participated in the project, undertaking classifications for apparent altruism. These participants have undisclosed backgrounds in the field of meteorology and cyclone behaviour, although three of the co-authors are amongst this group, and all three hold relevant doctorates and publications.

Summary information for each of these groups is provided in **Table 1**.

### *Skill scores data*

The three CC classification aspects that were quantified in Knapp et al. (2016) were considered for this study: eye or no eye; intensity trend; and storm type. For each, the trace of scores returned by the EM algorithm of Knapp et al. (2016) were used to assess skill for each user. The trace constitutes the sum of the diagonal elements of an n *x* n array, where n is the number of choices available, and all rows and columns sum to 1. The number in each cell corresponds to the frequency whereby the user chooses this option relative to the bias-adjusted voting of all users. Thus, a user who agreed 100% of the time with the bias-adjusted voting majority would have a score of unity in each cell in the diagonal (user and majority of bias-adjusted voters agree) and zero in all other entries (user and majority of bias-adjusted voters disagree). More skilful users will therefore have scores approaching n, and scores cannot be lower than zero.

Three aspects of the CC user choices were considered in the present analysis. These become increasingly complex problems for users to address.

- The first category was to classify if an eye was present in the image (eye or no eye) (n = 2). Eyes are visually striking features in tropical cyclone imagery, and whether an eye is present is a purely binary choice that, at least for well-developed storms, is arguably the easiest choice required of a given user. An eye is present in the top left image of **Figure 1**, where a prominent occurrence of a circular warmer scene is evident in the middle of cold clouds, whereas no eyes are present in all other occurrences in **Figure 1**.
- The second category was to assess the intensity of a storm compared to an image of a storm from 24 hours prior (stronger, weaker, or the same) (n = 3). This classification requires a user to interpret the physical presentation of the storm (cloud structure, temperature). Because it permits three choices, it is a bit more challenging than category one. An example is provided in **Figure 2**.
- The third classification category was to determine a tropical cyclone type, which involves comparing the given image with example images and classifying according to a best fit to one of the six storm types

(n = 6; **Figure 1**). Cyclones can be chosen by users to be curved band, embedded center, eye, shear, post tropical, or no storm. Considerable ambiguity exists, even amongst professionals, about how to classify images by storm type, so this category is the most challenging in terms of both complexity and the number of options from which to choose.

### *Quantitative analysis approach*

To assess differences among groups, we carried out statistical testing of classification skill. Firstly, the populations were tested for normality. The Shapiro-Wilk test was deemed most powerful based on the relatively small dataset and the different sample sizes for the three populations (Razali and Wah 2011). As discussed further in the results, none of the populations satisfied normality assumptions. Therefore, the significance of any differences among scores from MU students, UNCA students, and volunteers were assessed using non-parametric tests.

Given the non-normally distributed data and small sample sizes, the Wilcoxon sign rank test was employed to assess differences in the mean skill of the three groups using the programming language R (R Core Team 2013). Details of the code used for all tests are supplied (Appendix 1, supplemental material). The Wilcoxon test was chosen over the conventional Student t-test or Welch's t-test, because it performs better when investigating data that are not distributed normally and which have small sample sizes (Imam et al. 2014).

To assess the significance of any difference in variance among groups, a series of tests were considered including Bartlett's test, Levene's test, one way analysis of variance (ANOVA) test, Fisher's test, and Kruskal-Wallis H Test. Bartlett's test and Levene's test were deemed not applicable because they assume that data are normally distributed. A one-way ANOVA test, Fisher's test, and the Kruskal-Wallis test have previously been tested against each other, with the Kruskal–Wallis test being regarded as more powerful without any modifications (Yusof et al. 2013). Therefore, we used the Kruskal-Wallis test to assess the significance of any differences in variance among the three populations.

Following initial analyses, we decided to split the volunteers into two groups–500 or more classifications and fewer than 500 classifications–to check for a significant difference between medium (denoted < 500) and high (denoted ≥ 500) intensity users. The majority of both UNCA and MU students had completed at least 500 classifications. Therefore, a comparison to the group of high-intensity contributors from the volunteer pool may be more appropriate if there are distinctions between the medium- and high-intensity volunteer sub-groups.

Given the discrepancy in population sizes amongst the much larger volunteer pool and the two academic participant groups, a bootstrap approach was employed to create multiple random draw sub-populations of the volunteer group of equivalent size to the two academic groups. Using the programming language R, a random subsample of 25 users was created from the volunteer group, and Wilcoxon and Kruskal – Wallis tests were carried out. This was repeated 1000 times, and the number of tests showing significance was tallied. The results were then converted to a percentage frequency of occurrence.

### *Qualitative data collection on students' perceptions*

To get a better impression of the students' thoughts and ideas about the assignment, a questionnaire was designed and issued to all 25 members of the MU through their university email portal. As many of the students had graduated, it is unclear how many were still checking their university email account when the questionnaire was distributed. Participants in the UNCA and altruistic groups were kept anonymous

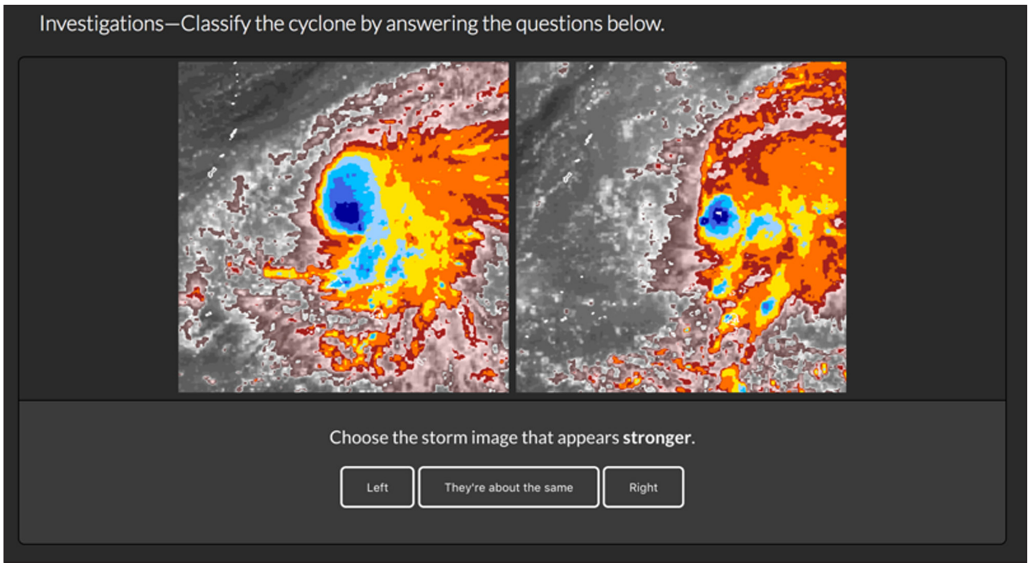


**Figure 2:** Example of the storm weaker/stronger/same step. Users are presented with two images from the storm 24 hours apart, with the earlier occurrence on the left, and required to decide which is stronger based upon their satellite presentation (cloud top temperatures and organization). In this case, the left hand image would be correctly identified as the stronger of the two, because it shows a greater mass of cold clouds and a greater degree of structure. Users always go on to classify the right-hand image regardless of their selection.

from the authors of this study (as was the skill of individual participants from MU), so contact with these participants was not possible. Of the MU participants, 14 members answered anonymously (Appendix 2, supplemental material). We wanted to understand whether the students felt that participation in research is of educational value to them as they learned about the research process while being assessed in a non-conventional manner. We also wanted to understand whether the students felt that their participation in the research added variety to their course (Landrum and Chastain 1995). Given the good return rate of questionnaire respondents (56%, assuming that all students received and read the email, higher if they did not), it can be inferred that the opinions expressed by the MU group are likely to be largely representative of the (still small) set of participants who took part for course credit at MU.

## Results

### *Testing for normality*

Results for a Shapiro-Wilk test for each group and for each of the three classification tests are shown in **Table 2**. The assumption of normality could be rejected in all cases except eye or no eye for the particular case of the MU student classifiers. Given the propensity for non-normal distributions in the skill scores for each population and each classification challenge considered, we make use of nonparametric tests in subsequent analyses, as outlined above.

**Table 2:** Shapiro – Wilk test for normality for the three test categories. Values below 0.05 mean that normality of the distribution can be rejected at the 95% confidence level and are shown in bold.

| Classification Category | MU | UNCA | Volunteers |
|---|---|---|---|
| Eye or no eye | 0.125 | **0.001** | **<0.001** |
| Stronger, weaker, or the same strength | **0.028** | **0.002** | **<0.001** |
| Type 6 | **0.031** | **0.009** | **<0.001** |

### *Eye or no eye*

Box plot results for eye or no eye classifications (n = 2) are shown in **Figure 3**. The Wilcoxon test showed no significant differences between the means of the different groups except for MU students vs all Volunteers (p-value < 0.05) and MU students vs volunteers ≥ 500 (p value < 0.05) (**Table 3**). The MU students have the lowest mean skill score (~1.72) for eye or no eye, with the volunteers ≥ 500 showing the highest skill score (~1.85). The Kruskal-Wallis test results for eye or no eye showed that the differences in variance of all the groups was not statistically significant (**Table 3**). UNCA show the largest variance (skill scores between 1 and 2) with volunteers ≥ 500 showing the least amount of variance (skill scores ranging between 1.6 and 2). Although the MU students have somewhat lower overall mean skill, they do not have as long low-skill propensity tails as the distribution for UNCA or medium-intensity (<500 classifications) volunteer groups (**Figure 3**).

### *Stronger, weaker, or the same strength*

The second category considered was when users were asked to classify if a cyclone is stronger, weaker, or the same strength as a previous image of the storm from 24 hours prior (**Figure 2**). This question has three potential answers, and users are expected to consider both cloud temperature and storm structure to make a determination. No statistically significant differences were present

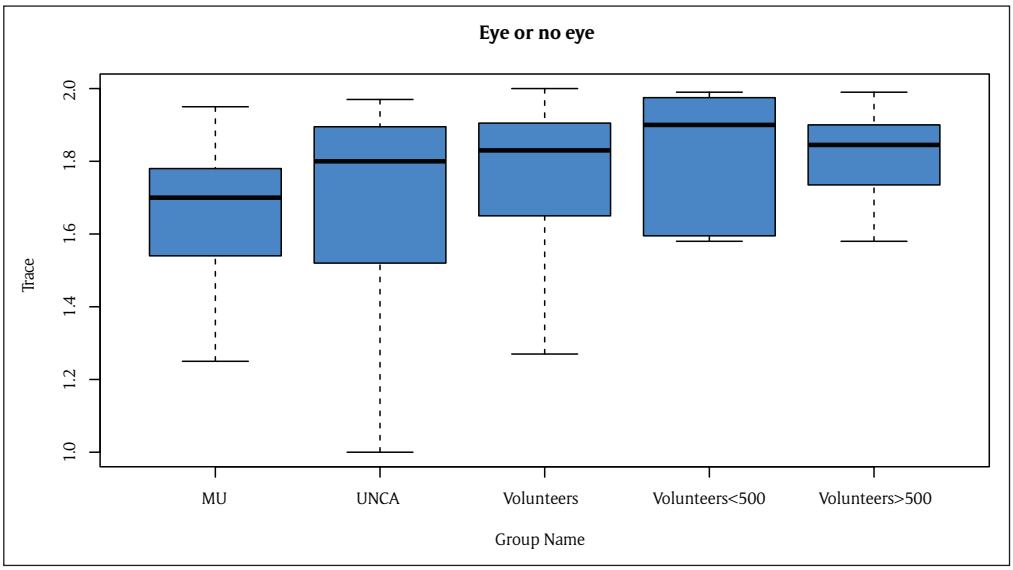


**Figure 3:** Boxplot of the results for eye or no eye. There are two choices, so skill is bound between 0 and 2, where 2 would indicate complete consistency with the bias-adjusted voting majority classification on all occasions. A trace value of 1 would represent completely random selections by the users and thus a complete lack of skill for this task. Boxes indicate interquartile ranges and whiskers denote 5–95% ranges. The median trace value in each population is indicated by the horizontal bar in the interquartile box. Both the full volunteer sample and the subsets of medium and high intensity users are shown.

between any of the groups for either the Wilcoxon or Kruskal-Wallis tests for this classification step (**Table 4**). A visual representation (**Figure 4**) shows that MU have the smallest variance of skill score of 0.8, with UNCA and volunteers < 500 having the largest variance of 1.2, with UNCA varying between 1.2 and just under 2.4 and the volunteers < 500 varying between just under 1.4 and 2.6. The MU and UNCA student groups do not have the high-skill tail that is present for both volunteers ≥ 500 and volunteers < 500. MU also lack a low-skill tail present in all other populations for this classification type (**Figure 4**).

**Table 3:** Wilcoxon (mean) and Kruskal-Wallis (variance) test results for eye or no eye. Significant results (at 95% level) are shown in bold.

| Eye or no eye | Wilcoxon (p-value) | Kruskal-Wallis (p-value) |
|---|---|---|
| MU vs UNCA | 0.322 | 0.304 |
| MU vs Volunteers | **0.009** | 0.802 |
| UNCA vs Volunteers | 0.283 | 0.584 |
| MU vs Volunteers < 500 | 0.071 | 0.317 |
| MU vs Volunteers ≥ 500 | **<0.001** | 0.646 |
| UNCA vs Volunteers < 500 | 0.490 | 0.401 |
| UNCA vs Volunteers ≥ 500 | 0.124 | 0.485 |

**Table 4:** Wilcoxon and Kruskal Wallis test results for Stronger, Weaker, or the same strength. No cases are significant at the 95% level, although UNCA versus volunteers > 500 comes very close.

| Stronger, Weaker, or the same strength | Wilcoxon (p-value) | Kruskal-Wallis (p-value) |
|---|---|---|
| MU vs UNCA | 0.465 | 0.364 |
| MU vs Volunteers | 0.623 | 0.874 |
| UNCA vs Volunteers | 0.150 | 0.595 |
| MU vs Volunteers < 500 | 0.926 | 0.372 |
| MU vs Volunteers ≥ 500 | 0.289 | 0.474 |
| UNCA vs Volunteers < 500 | 0.308 | 0.375 |
| UNCA vs Volunteers ≥ 500 | 0.055 | 0.334 |

### *Storm type classification*

The final classification choice was assigning the storm as one of six types (**Figure 1**). No statistically significant difference was found between any of the groups for the Wilcoxon or Kruskal-Wallis tests (**Table 5**). MU scored the highest mean skill with the least amount of variance. Volunteers < 500 scored the lowest mean skill with the largest variance (**Figure 5**). The variance for all groups is larger for this case than for the previous cases, which had fewer choices available to classifiers. The MU group lack the high, but in particular the low-skill tails present in remaining groups of classifiers, and show the highest mean skill score for this step (**Figure 5**).

**Table 5:** Wilcoxon and Kruskal Wallis test results for Type 6 (cyclone type assignment). Values significant at the 95% level would be indicated in bold (no occurrences).

| Type 6 | Wilcoxon (p-value) | Kruskal-Wallis (p-value) |
|---|---|---|
| MU vs UNCA | 0.262 | 0.645 |
| MU vs Volunteers | 0.413 | 0.407 |
| UNCA vs Volunteers | 0.665 | 0.466 |
| MU vs Volunteers < 500 | 0.181 | 0.525 |
| MU vs Volunteers ≥ 500 | 0.928 | 0.695 |
| UNCA vs Volunteers < 500 | 0.744 | 0.429 |
| UNCA vs Volunteers ≥ 500 | 0.222 | 0.609 |

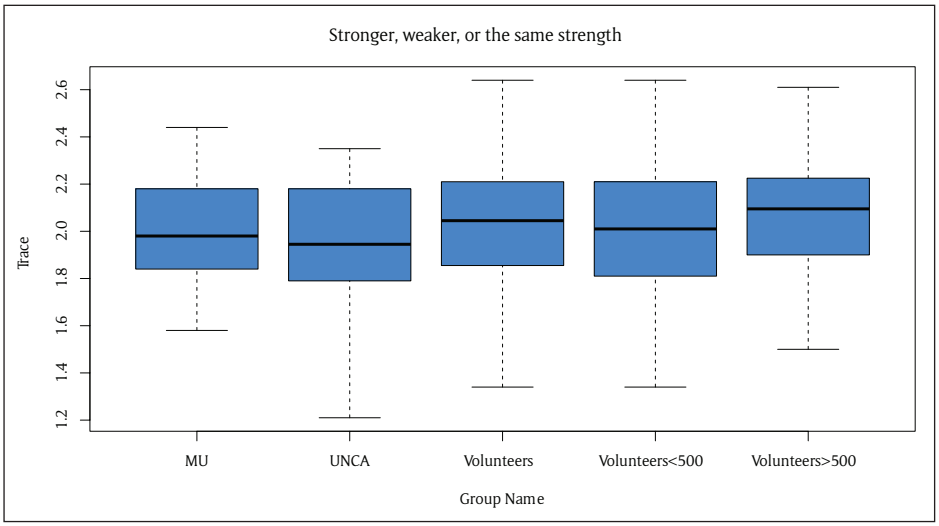


**Figure 4:** Boxplot of the results for Stronger, Weaker, or the same strength question (n = 3, so trace value bounded between 0 and 3). A trace value of 1 would represent completely random selections by the users and thus a complete lack of skill for this task.

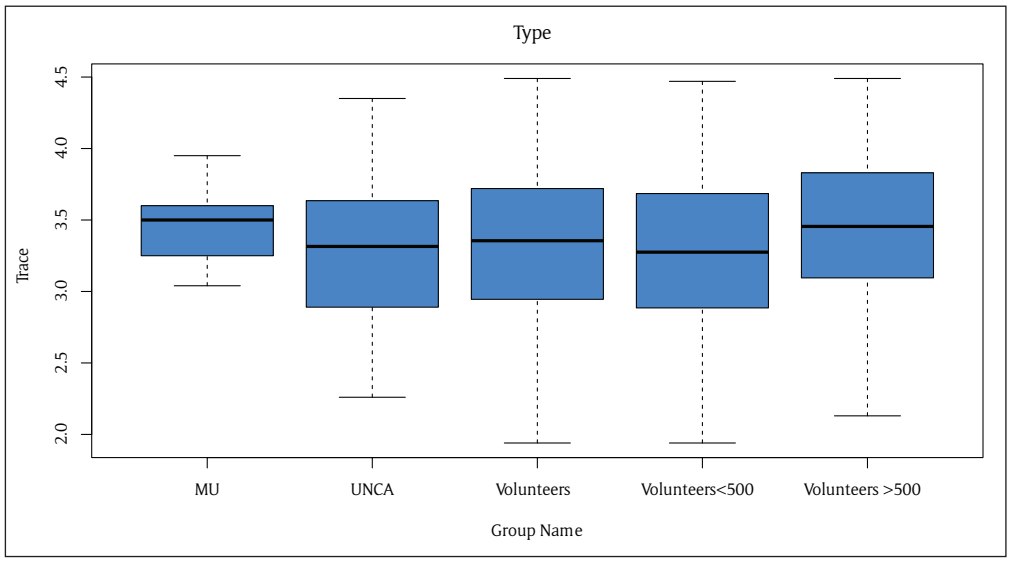


**Figure 5:** Boxplot of the results for the storm typing question (n = 6, so trace value bounded between 0 and 6). A trace value of 1 would represent completely random selections by the users and thus a complete lack of skill for this task.

### *Bootstrap resampling*

The 1000 subsamples taken from the volunteer group were subjected to a mean and variance test. The results of the mean test (**Table 6**) show that in all cases except one, the overall mean skill across subsamples did not change significantly. From the Volunteers 1000 subsamples for eye or no eye, 470 samples (47% of cases) were shown to be significantly different from the MU students mean score. This is consistent with the whole population comparison results, which also showed a significant difference. No bootstrapped volunteers subset resulted in p values less than 0.05 for the Kruskal-Wallis test for any of the groups across any of the classification categories.

### *Overall assessment of results*

The majority of comparisons between students and altruistic participants yielded no statistically significant differences in aspects of skill as measured by the EM algorithm of Knapp et al. (2016). The one significant finding is for a difference in the mean skill for MU students for the eye or no eye classification, with the MU students agreeing somewhat less frequently with the consensus estimate. In no cases is the variance statistically significantly different. However, from a consideration of boxplot results, the MU students appear to avoid a low-classification skill tail (whereby individual users more consistently disagree with the consensus classification), which is more readily apparent in the UNCA and volunteer users' distributions. This finding hints at a potential avenue for further investigation and a possible impact of embedding the classifications in a broader piece of coursework. However, given the small sample size involved, this result must be treated with caution. A larger sample size would be required to robustly assess aspects of skew between the populations.

### *Qualitative student responses*

There were 14 respondents to the questionnaire from MU, which explored the outlook that students had on participating in citizen science during the CC aspect of the module. The responses have been summarised in **Table 7**. Despite the good response rate, with such a small population we limit the analysis here to a qualitative commentary. The majority of responses stated that the classification of tropical cyclones on CC as part of their credit was of moderate difficulty. This suggests that students did not find the task too easy for their university level education, or so difficult as to discourage them from completing the task. Responses indicated that most participants found the task of classifying cyclone satellite imagery enjoyable, which is very important for maintaining engagement with tasks of this nature. Most respondents specified that they felt their knowledge of cyclones, as well as their skill of classifying images, increased as they completed more classifications. This is beneficial information for determining whether it is useful to award course credit for completion of such citizen science tasks. If students feel as if they are benefitting their knowledge and skills by participating in research of this type, an argument can be made for awarding credit for academic engagement, which also benefits CC through additional classifications accrued.

**Table 6:** Number of Volunteer subsamples whose mean had a p value of < 0.05 in comparison with either MU or UNCA for the three classification categories, given as a percentage.

| Group (Classification category) | Percentage of tests < 0.05 |
|---|---|
| MU vs Volunteers (Eye or no eye) | 47 |
| UNCA vs Volunteers (Eye or no eye) | 5.4 |
| MU vs Volunteers (Stronger, Weaker, Same) | 1.0 |
| UNCA vs Volunteers (Stronger, Weaker, Same) | 7.3 |
| MU vs Volunteers (Type 6) | 2.9 |
| UNCA vs Volunteers (Type 6) | 0.4 |

**Table 7:** Questionnaire results from MU students. For each question, the response options available to students are provided together with the distribution of tallies.

| | | | | | | |
|---|---|---|---|---|---|---|
| Question 1 | How would you rate classifying cyclones? | Extremely Difficult | Difficult | Moderate | Easy | Extremely Easy |
| | Answers | 0 | 4 | 9 | 1 | 0 |
| Question 2 | Did you enjoy classifying cyclones? | Yes | No | | | |
| | Answers | 11 | 3 | | | |
| Question 3 | Did you feel you improved your skills of classifying cyclones as you completed more classifications? | Yes | No | | | |
| | Answers | 14 | 0 | | | |
| Question 4 | How would you rate your knowledge of cyclones before classifying? | Very poor | Poor | Moderate | Good | Very Good |
| | Answers | 1 | 5 | 6 | 2 | 0 |
| Question 5 | How would you rate your knowledge of cyclones after classifying? | Very poor | Poor | Moderate | Good | Very Good |
| | Answers | 0 | 1 | 3 | 8 | 2 |
| Question 6 | What was your motivation to stop classifying cyclones? | I had completed the required amount to receive course credit | I lost interest | Other | | |
| | Answers | 12 | 0 | 2 | | |
| Question 7 | How did you feel this assessment compared to other modules in terms of the workload? | It was heavier | It was about the same | It was lighter | | |
| | Answers | 1 | 6 | 7 | | |
| Question 8 | Would you have preferred an alternative mode of assessment | Yes | No | | | |
| | Answers | 0 | 14 | | | |

When asked if they would prefer a different method of assessment for this module, all students responded "no." This result suggests that students were satisfied with the assessment requirements of the module, possibly due to the added variety compared to more traditional exam or essay-based assessments. The high level of agreement in student responses in the questionnaire is encouraging, but a larger sample would be required to make definitive conclusions.

## Discussion

Several caveats pertain to our analyses. First and foremost, the small available sample sizes of participant groups limit the robustness of quantitative statistical conclusions. In particular, larger groups of student user populations would have enabled more robust inferences to be drawn and formal quantitative investigation to be made of aspects such as skew in the distributions. In that context, we believe that our results should be interpreted as advisory rather than final. Repetition, either with larger student samples and/or by other citizen science projects, is required to either confirm or refute our findings. To that end, MU has repeated its assignment exercise over the past two years with a first-year class having in excess of 250 students participating, although requiring that only 250 images be classified by each student. This expanded sample should provide a more substantial population from which more robust statistics can be inferred. However, this group consists of students who are far earlier in their university courses and commensurately earlier in their training. This may confound a clean comparison to our present results. Similarly, UNCA continues to engage students. Hence, the pool of for-credit students has grown and continues to grow and, in the case of MU, diversify. In the future, we hope to revisit the present analysis using these much larger populations of for-credit students, which may permit more advanced statistical enquiry and hence more robust conclusions.

Results might also reasonably be expected to be specific to different citizen science projects. To that end, CC is likely a good case-study project, because it is one of the more technically challenging citizen science projects (at least within the Zooniverse family). Intuitively, were there to be any substantive performance differences between volunteer and for-credit student users, they would be most acute in such technically challenging projects. Nevertheless, there would be considerable value in repeating this type of study across a range of citizen

science projects encompassing a broad range of technical difficulty.

In cases where significant differences occur between volunteer and for-credit user performance in the classification steps, the analysis tells us solely that there is a difference. It cannot tell us which group shows higher absolute skill, as the EM skill of Knapp et al. (2016) is relative to consensus, which is dominated by the volunteer population. If all of the volunteer users consistently misinterpret certain imagery, they all can be consistently and systematically wrong. In practical terms, this means that our finding that the MU and the volunteer group diverge in mean EM-based skill scores for eye scenes (eye or no-eye) cannot directly be taken to imply that the volunteer group has the higher absolute skill in identifying eye scene types. It may be, for example, that the MU group, following instructor guidance, was better at spotting emerging eye features than the average user. If this were the case, divergence from the consensus would lead to lower average apparent skill scores (agreement with consensus) under the EM algorithm of Knapp et al. (2016) being obtained by the MU group, despite actually exhibiting better absolute skill. Ascertaining absolute skill would require a distinct metric, for example, of comparison to a trained expert's (or preferably small group of experts') classification of the same images.

The study was a "study of opportunity" in that the different student coursework assignments were developed independently and with minimal cooperation, with no view to their being used in the present study. A future study of this nature would benefit greatly from participating universities coordinating in advance. This could ensure consistent delivery of training materials. It also could involve better coordination of assignment of credit structures in advance. By designing appropriately similar or distinct approaches, more formally testing distinct hypotheses would be possible. These could be, for example, around course credit models and the effects of integrating participation into broader assignments. Ideally, a subsequent study would set out to more formally test questions around participation modalities by deliberately designing and delivering the course assignments as part of the study. For robustness, this would be done consistently across two or more distinct citizen science projects.

The student questionnaire had a high response rate, but the group was small, and no commensurate sample was undertaken for UNCA students. Were the analysis to be repeated, we would suggest that student feedback be collected from the outset and again upon completion of the associated assignment. This would provide a more robust collection and analysis of student perceptions. Use of student assignments in citizen science projects requires a value proposition to the students whereby some novel engagement or learning outcomes result that are not easily obtained by other means of teaching and assignment delivery. While we firmly believe that participating in citizen science is more meaningful than a traditional essay assignment, there is a distinction between believing so and knowing so, which requires robust collection of student reflections. However, the motivations of students are difficult to determine. We note the distinctions outlined in the Data and Methods Section. UNCA associated classification with at most 2% of course credit, whereas MU associated 50% of their course credit with the assignment. Furthermore, MU students had to undertake additional analysis to gain all associated credit, whereas for UNCA, credit revolved solely around classification counts. Assumptions about student motivations can be made on the basis of the module selection, as participation in the courses and hence CC is optional for students in both cases. In general, interest plays a major role in the module selection of students. Students from MU were briefed on the module outlines and assessment method, so part of the motivation to take part in CC for these students can be assumed to be interest. Students from UNCA who participated in CC can be assumed to have done so for a mix of extra credit and personal interest.

We can be even less certain of the reasons for participation of the altruistic user group. As CC requires solely a username registration, we have no indication of the level of educational attainment or post-education engagement of project volunteers in tropical cyclone analyses. Three of the volunteers are the two course tutors and one other author of this paper, but they represent less than 1% of the altruistic sample, and are thus unlikely to substantively bias the results. Beyond that we have little knowledge about participants.

Further, while unlikely, we cannot rule out that other student groups were amongst the sample labelled "altruistic" who had been given assignments by course tutors. If use of citizen science platforms becomes an increasingly common feature in participatory-learning focussed university coursework assignments, then a mechanism to notify project science teams and track which are student assignment-based classifications would be highly advisable. This would permit subsequent analysis that may control for the effects, if found to be necessary.

Finally, the Knapp et al. (2016) analysis results used in this analysis relate to three steps common to all image classifications. For most classifications, users are asked subsequent questions around image features. These questions are dependent upon the selected storm type (the third and most complex question we considered), such that each time a type is considered, the same set of subsequent questions are asked. However, these follow-on questions are distinct for each storm type. EM-algorithm skill scores for these subsequent steps were not available to us and, even if they were, the sample sizes would be much smaller. These issues could be considered in any follow-up study.

## Conclusions

The aim of this investigation was to determine if participation in the citizen science project Cyclone Center for course credit or for volunteer purposes influenced the quality of the resulting classifications. Overall, we find no compelling evidence for a significant difference between student and altruistic participants based upon a skill metric that measures closeness to consensus opinion. Although one of our two student groups recorded a significantly lower mean skill score on one of the three tasks in the project, they also displayed a smaller low-skill tail than the volun-

teer group. On this basis, we suggest that the motivations for classifying cyclones, whether for course credit or altruism, do not affect the quality of the classification result. Therefore, we regard the awarding of course credit to students who participate in citizen science projects of this type to be suitable. Similar assignments for other citizen science projects could increase project participation rates and provide valuable learning outcomes for students.

Recommended future work includes testing the robustness of the findings with increased sample sizes for all groups and understanding low-tail (low skill) propensity. This can be achieved by re-running the credit awarded modules in both MU and UNCA a sufficient number of times, then analysing with the increased sample sizes. It also could be achieved by running similar exercises with other citizen science projects. If our findings hold true across other citizen science projects, this opens up avenues to use citizen science projects in participatory-learning based activities more broadly in university courses. This could be valuable to the students, the universities, and the citizen science projects themselves.

## Additional Files

The additional files for this article can be found as follows:

- **Appendix 1.** R code and data description. DOI: https://doi.org/10.5334/cstp.111.s1
- **Appendix 2.** Questionnaire. DOI: https://doi.org/10.5334/cstp.111.s1

## Ethics and Consent

To ensure unbiased analyses and that ethics considerations were met, all user identification beyond the group that each user belonged to was redacted prior to analysis – no personally identifiable information beyond which of the three sub-classes each participant arose from was at any point disclosed. All students and volunteers were fully aware that their contributions may be analysed as a result of their participation.

## Acknowledgements

The authors would like to thank all of the Cyclone Center team for facilitating this research. We appreciate the help of Chris Brunsdon and Martin Charlton of NCG, Maynooth University, who provided training in R. We are grateful to all the contributions made by numerous citizen scientists for participating in Cyclone Center. Special thanks must be extended to the students of both MU and UNCA who participated in the project for attainment of credit during their university education. These participants have contributed an indispensable 78,739 analyses of observations of satellite imagery to the Cyclone Center project. Two anonymous reviewers are thanked for their reviews of the original submission.

## Competing Interests

The first four authors completed this analysis as part of their masters course in climate science at Maynooth University in Autumn semester 2016. Following completion of the assignment, they agreed to lead its drafting as a paper. Two of these authors also participated as undergraduate students in the original Maynooth University course assignment analysed herein. These students had no way of identifying their participation scores. No other potential competing interests exist.

## References

Allen, M. 1999. Do it yourself climate prediction. *Nature*, 401: 642. DOI: https://doi.org/10.1038/44266

Climateprediction.net. 2017. http://www.climateprediction.net/publications/?type=29&letter=&theme [Last accessed 19/5/17].

Cox, J., Oh, E.Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, C. and Holmes, K. 2015. Defining and measuring success in online citizen science: A case study of Zooniverse projects. *Comput. Sci. Eng.*, 17: 28–41. DOI: https://doi.org/10.1109/MCSE.2015.65

Cyclone Center. 2017. https://www.cyclonecenter.org/ [Last accessed 16/01/18].

Dvorak, V.F. 1984. Tropical cyclone intensity analysis using satellite data. *NOAA/NESDIS Tech. Rep.*, 11: 47.

Franzoni, C. and Sauermann, H. 2014. Crowd science: The organization of scientific research in open collaborative projects. *Research Policy.*, 43: 1–20. DOI: https://doi.org/10.1016/j.respol.2013.07.005

Hennon, C.C., Knapp, K.R., Schreck, C.J., III, Stevens, S.E., Kossin, J.P., Thorne, P.W., Hennon, P.A., Kruk, M.C., Rennie, J., Gadéa, J.-M., Striegl, M. and Carley, I. 2015. Cyclone Center: Can citizen scientists improve tropical cyclone intensity records? *Bull. Amer. Meteor. Soc.*, 96: 591–607. DOI: https://doi.org/10.1175/BAMS-D-13-00152.1

Imam, A., Mohammed, U. and Moses Abanyam, C. 2014. On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test. *IOSR Journal of Mathematics*, 10(1): 1–6. DOI: https://doi.org/10.9790/5728-10140106

Karlin, M. and De La Paz, G. 2015. Using camera-trap technology to improve undergraduate education and citizen-science contributions in wildlife research. *Southwest Nat.*, 60: 171–9. DOI: https://doi.org/10.1894/SWNAT-D-14-00005.1

Knapp, K.R. and Kossin, J.P. 2007. New global tropical cyclone data set from ISCCP B1 geostationary satellite observations. *J. Appl. Remote Sens.*, 1: 013505. DOI: https://doi.org/10.1117/1.2712816

Knapp, K.R. and Kruk, M.C. 2010. Quantifying interagency differences in tropical cyclone best-track wind speed estimates. *Mon. Wea. Rev.*, 138: 1459–1473. DOI: https://doi.org/10.1175/2009MWR3123.1

Knapp, K.R., Matthews, J.L., Kossin, J.P. and Hennon, C.C. 2016. Identification of tropical cyclone "storm types" using crowd-sourcing. *Mon. Wea. Rev.*, 144: 3783–3798. DOI: https://doi.org/10.1175/MWR-D-16-0022.1

Landrum, R.E. and Chastain, G. 1995. Experiment spot checks: A method for assessing the educational value of undergraduate participation in research. *IRB: A Review of Human Subjects Research*, 17(4): 4–6. DOI: https://doi.org/10.2307/3564152

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. *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, 94–102. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/viewFile/7497/7408.

Mason, W. and Watts, D.J. 2009. Financial incentives and the "performance of crowds". *SIGKDD Explorations*, 11(2): 100–108. DOI: https://doi.org/10.1145/1809400.1809422

Maynooth University. 2017. https://www.maynoothuniversity.ie/icarus/icarus_data [Last accessed 04/03/2018].

Met Office. 2017. https://wow.metoffice.gov.uk/ [Last accessed 19/05/17].

Mitchell, N., Triska, M., Liberatore, A., Ashcroft, L., Weatherill, R. and Longnecker, N. 2017. Benefits and challenges of incorporating citizen science into university education. *PLoS ONE*., 12(11): e0186285. DOI: https://doi.org/10.1371/journal.pone.0186285

Muller, C.L., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., Overeem, A. and Leigh, R.R. 2015. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *Int. J. Climatol*, 35: 3185–3203. DOI: https://doi.org/10.1002/joc.4210

NOAA. 2017. http://www.nws.noaa.gov/om/coop/what-is-coop.html [Last accessed 19/5/17].

Oberhauser, K. and LeBuhn, G. 2012. Insects and plants: engaging undergraduates in authentic research through citizen science. *Front. Ecol. Environ.*, 10: 318–20. DOI: https://doi.org/10.1890/110274

OED. 2017. *Citizen science – definition of citizen science in English|Oxford Dictionaries*. [online]. Available at: https://en.oxforddictionaries.com/definition/citizen_science [Last accessed 20/01/2018].

OldWeather. 2017. https://www.oldweather.org/ [Last accessed 20/5/17].

Padilla-Walker, L.M., Thompson, R.A., Zamboanga, B.L. and Schmersal, L.A. 2005. Extra credit as incentive for voluntary research participation. *Teaching of Psychology.*, 32(3): 150–153. DOI: https://doi.org/10.1207/s15328023top3203_2

R Core Team. 2013. R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. [online]. Available at: http://www.R-project.org/ [Last Accessed 21/01/17].

Raddick, M.J., Bracey, G., Gay, P.J., Lintott, C.J., Murray, P., Schawinski, K., Szalay, A.S. and Vandenberg, J. 2010. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review.*, 9. 010103-1. DOI: https://doi.org/10.3847/AER2009036

Razali, N. and Wah, Y. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1): 21–33.

Reges, H., Doesken, N., Turner, J., Newman, N., Bergantino, A. and Schwalbe, Z. 2016. COCORAHS: The evolution and accomplishments of a volunteer rain gauge network. *Bull. Amer. Meteor. Soc*. DOI: https://doi.org/10.1175/BAMS-D-14-00213.1

Ren, F., Liang, J., Wu, G., Dong, W. and Yang, X. 2011. Reliability analysis of climate change of tropical cyclone activity over the western North Pacific. *J. Climate*, 24: 5887–5898. DOI: https://doi.org/10.1175/2011JCLI3996.1

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 321–328.

Roy, H.E., Pocock, M.J.O., Preston, C.D., Roy, D.B., Savage, J., Tweddle, J.C. and Robinson, L.D. 2012. Understanding Citizen Science & Environmental Monitoring. *Final Report on behalf of UK-EOF*. NERC Centre for Ecology & Hydrology and Natural History Museum. Available from: http://www.ceh.ac.uk/sites/default/files/citizensciencereview.pdf.

Ryan, C., Duffy, C., Broderick, C., Thorne, P.W., Curley, M., Walsh, S., Daly, C., Treanor, M. and Murphy, C. (submitted). Integrating data rescue into the classroom, Submitted to *Bull. Amer. Met. Soc.*

Velden, C., Harper, B., Wells, F., Beven, J.L., II, Zher, R., Olander, T., Mayfield, M., Guard, C., Lander, M., Edson, R., Avila, L., Burton, A., Turk, M., Kikuchi, A., Christian, A., Caroff, P. and McCrone, P. 2006. The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bull. Amer. Meteor. Soc.*, 87: S6–S9. DOI: https://doi.org/10.1175/BAMS-87-9-Velden

Weather Rescue. 2017. https://www.zooniverse.org/projects/edh/weather-rescue [Last accessed 18/01/18].

Yusof, Z., Abdullah, S. and Yahaya, S. 2013. Comparing the performance of modified $F_t$ statistic with ANOVA and Kruskal Wallis test. *Applied Mathematics & Information Sciences*, 7(2L): 403–408. DOI: https://doi.org/10.12785/amis/072L04

**How to cite this article:** Phillips, C., Walshe, D., O'Regan, K., Strong, K., Hennon, C., Knapp, K., Murphy, C. and Thorne, P. 2018 Assessing Citizen Science Participation Skill for Altruism or University Course Credit: A Case Study Analysis Using Cyclone Center. *Citizen Science: Theory and Practice*, 3(1): 6, pp. 1–13, DOI: https://doi.org/10.5334/cstp.111

**Submitted:** 23 May 2017 **Accepted:** 16 March 2018 **Published:** 04 June 2018



]u[ *Citizen Science: Theory and Practice* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS