Scaffolding Training with Machine Learning: An Experiment on Participant Learning in an On-line Production Community

Corey Jackson1 · Carsten Østerlund1 · Kevin Crowston1 · Mahboobeh Harandi1 · Sarah Allen2 · Sara Bahaadini4 · Scott Coughlin3 · Vicky Kalogera3 · Aggelos Katsaggelos4 · Shane Larson3 · Neda Rohani4 · Joshua Smith5 · Laura Trouille2 · Michael Zevin3

Abstract We conducted an on-line field experiment where we recruited users (N = 386) in a web-based citizen-science project to evaluate the effectiveness of using machine learning (ML) to select tasks to implement a scaffolded approach to learning. We found that users who experienced the ML-scaffolded training performed significantly better on the task (an average accuracy of 90% vs. 54%), contributed more work (an average of 228 vs. 121 classifications), and were retained in the project for a longer period of time (an average of 2.5 vs. 2 sessions). The results suggests that ML approaches can be effectively used to implement scaffolded training.

1 Introduction

We present the design and experimental evaluation of a system that uses computer-selected tasks to scaffold the learning of new participants in an on-line system. Newcomers joining an organization typically go through some kind of orientation and training, formal or informal, to learn how to contribute. In some settings, they may be able learn through observation of the contributions of more experienced volunteers, i.e., through participation as a peripheral member of the group [30]. However, this form of transparency is not possible for all types of work and even when possible, it can take a long time for newcomers to learn through observation.

As a result, explicit training is required for many tasks. However, training has costs. First, developing a training program requires additional work to create appropriate materials. Second, newcomers are not productive during their training period, as their time is devoted to training rather than work. To avoid these costs, organizations might choose to
train “on the job”, by having newcomers simultaneously do the work and receive feedback. Knowing that the work is useful and being given challenging tasks may also be motivating for the newcomer. However, if it takes time to learn to do the task correctly, then the initial contributions may not be of high enough quality to be useful. Furthermore, if newcomers find the task too challenging, they may become discouraged and leave.

We consider the possibility of training in on-line production communities, that is, settings in which individuals co-create some output via an on-line system. To be successful, these communities need to sustain a critical mass of skilled and active participants [11, 33], which requires attracting newcomers and helping them learn to become effective participants in the community. However, the characteristics of on-line communities present particular challenges to newcomer training. First, most communities rely on volunteers who contribute in their free time, reducing their willingness to participate in formal training regimes prior to engaging. Second, the problem of having newcomers be unproductive during training is exacerbated by the skewed distribution of contributions to on-line communities: most volunteers contribute only a few times and only a few become sustained contributors. As a result, increasing the barrier to entry and delaying newcomers’ contributions might result in many not contributing at all. Finally, few on-line communities have the resources to monitor and provide feedback to newcomers, making on-the-job training less attractive.

In short, on-line communities face a particular dilemma in how to handle newcomers. Providing training might increase the quality of contributions, but at the cost of the work done by newcomers during the training period, which for many is the only work they will contribute, and at a cost to those creating or delivering the training. On the other hand, not providing training might mean that initial contributions are not useful [5] and the difficulty of the task may discourage participation.

Scaffolding, that is, gradually introducing learners to carefully selected new materials to support learning, has been shown to enhance learning and has positive impacts on performance and motivation in both authentic classroom settings [42] and on-line learning environments [35,43,50]. However, the design of scaffolded training has been done primarily by human experts who carefully select appropriate materials to be introduced. In this paper, we present a novel approach to scaffolding training by using a machine learning (ML) classifier to select genuine tasks (i.e., tasks currently being worked on in the system) to provide newcomers in an on-line citizen science community. Following this approach, newcomers are introduced to new types of tasks gradually rather than all at once. However, whether an ML-scaffolded training regime using genuine tasks can be effective is an open question. It could be that genuine tasks are not appropriate for training, meaning that the ML-chosen tasks are no better than random tasks for helping the volunteers to learn. To answer this question, we evaluated the ML-scaffolded training in an on-line field experiment. We found that compared to non-scaffolded training, this approach increases performance, retention and the volume of contributions.

1.1 Gravity Spy

Our experiment was carried out in an on-line citizen-science project, Gravity Spy [52], which is hosted on the Zooniverse platform [47]. Citizen science is a broad term describing scientific research projects that rely on contributions from members of the general public (i.e., citizens in the broadest sense of the term). There are several kinds of citizen-science projects: some have volunteers collect data, while others, including the one we examine in this paper, have volunteers analyze already-collected data. The interactions between volun-
Scaffolding Learning with Machine Learning

The scientific goal of the Gravity Spy project is to classify “glitches” produced by the Laser Interferometer Gravitational-wave Observatory (LIGO). LIGO’s goal is to detect gravitational waves, extremely faint distortions in the fabric of space created by astronomical events such as merging black holes or neutron stars. A challenge for LIGO scientists is that the detectors need to be extremely sensitive to be able to detect gravitational waves, but as a result, they also record large numbers of noise events (referred to as glitches) caused by terrestrial interference or by internal faults or interactions in the detectors. Glitches are a problem because they can obscure or even masquerade as gravitational wave signals, reducing the efficacy of the search. The LIGO scientists would like to have a collection of glitches with a similar appearance to focus their search for the source of that type of glitch in order to improve the detector. However, glitches occur hundreds or thousands of times a day (depending on the threshold used), meaning that there are many more than the LIGO scientists can analyze in detail themselves.

To support the LIGO project, the Gravity Spy project presents volunteers with the spectrogram of a glitch and asks them to classify it into one of the known classes or “none of the above”. The Gravity Spy classification interface is shown in Figure 1. On the left of the classification interface is a spectrogram of a glitch, with time on the x-axis, frequency on the y-axis and intensity represented by the color from blue to yellow. To the right of the spectrogram are the classes from which a volunteer can select to categorize the glitch shown. Currently there are more than 20 known classes of glitch with different causes, with the possibility of more classes being identified (or created) as the detector is altered and improved. When the volunteer clicks on one of the classes, the system shows a larger image.
of a prototypical instance of the class with some more information about the glitch morphology. Once the selection is finalized, the volunteer can add a comment on the image (the “Talk” interface) or classify another glitch. Volunteers end their interaction with the system by logging out or more commonly by simply stopping.

2 Theory

Gravity Spy was designed drawing on theory to increase both volunteer learning and motivation. By learning, we mean specifically learning to do the classification task accurately. Volunteers may also learn about gravitational waves or science more generally, but these kinds of learning are beyond the scope of this paper.

2.1 Learning to classify glitches

To promote accurate glitch classification, the Gravity Spy system integrates four approaches to learning. Of these four, the first three (explicit training, feedback, and presentation of prototypes and exemplars) can be found in many crowd-sourcing and citizen-science projects. The fourth approach (scaffolding supported by ML) adds a new approach to training.

2.1.1 Explicit training

On-line production community sites typically provide a brief introduction to the project that explains its goals and tasks. Citizen-science projects in particular provide training on the scientific tasks, how to interpret images, and how to use the classification interface. In the Zooniverse, the training is provided as a pop-up when a volunteer first visits the site that can be revisited at any time via the “Show the project tutorial” link below the classification options.

2.1.2 Feedback on classification

Feedback on performance is critical in the learning process [7, 12, 15, 31, 36]. The Gravity Spy system therefore has beginning volunteers classify some glitches from a gold-standard dataset (i.e., glitches previously classified by members of the science team). Knowing the correct classification makes it possible to give the volunteers feedback on the correctness of their classifications and to assess their accuracy in classifying those classes of glitch. As an added benefit, [7] found that when feedback was provided for participants on their performance, they were more motivated to contribute than when feedback was not provided.

2.1.3 Prototypical and exemplary glitches

Cognitive theories suggest that people learn to classify images through exposure to prototypes and exemplars of known categories. Prototypes serve as a heuristic: an average representation of an entire category [26]. Exemplars function as examples for the category [27]. When individuals are asked to generalize a category, they evaluate several characteristics and weight each of these characteristics [23, 38, 46, 48]. That is, individuals decide whether an image belongs to a category depending on how much the image is similar to or different from the prototypes and exemplars in certain characteristics and the importance of the
characteristics (i.e., weights). As individuals see more images, they update the weights for the stimuli characteristics. Therefore, to support learning of image classification, volunteers should be continuously provided with good prototypes and exemplars of the classes.

Gravity Spy presents prototypical and exemplary images of glitches to volunteers in two ways. First, the classification interface shows volunteers prototypical instances of the various classes to guide their selection. When a class is selected, a larger image of a prototypical example and a brief description are displayed to reinforce the exemplar. Second, Gravity Spy, like many Zooniverse projects, provides a “field guide”, with prototypical glitches, several exemplars and discussion of the kinds of data to be classified. The field guide is accessed via a link to the far right side of the interface.

2.2 Scaffolding Training

Finally, the zone of proximal development concept emphasizes the need to adjust learning opportunities to the learner’s current abilities [14]. Studies of learning through legitimate peripheral participation similarly suggest that learners gradually expand their access to central activities [30]. The emerging literature on learning in on-line settings specifically and eLearning more broadly [10,16,17,22,32,39] suggests that learning emerges as participants gradually expand their engagement with a task. Bringing these concerns together, the concept of scaffolding suggests the importance of carefully sequencing participants’ learning opportunities [21].

Volunteers’ progress in on-line production communities has been analyzed and supported from these perspectives. In a study of Wikipedia, for instance, Bryant et al. [4] showed how novices often start out by simply reading other’s articles before they start making their initial contributions and gradually access more involved tasks. Preece and Shneiderman [40] similarly suggested that participants in peer-production sites move from “readers to leaders”. The Fold It citizen science system [6] provides a series of increasingly challenging tasks to help newcomers learn the task (protein folding), though these are already-solved tasks. In other learning environments, we see increased calls for scaffolded mechanisms to deliver formative feedback [44] to guide users in the learning process.

The innovation in the Gravity Spy system is that we implemented scaffolding using real tasks selected by a machine-learning (ML) system that classifies all glitches added to the system into one of the known classes, with an accompanying confidence in the classification. Details of the ML algorithm and classification are reported in [1,52]. The system introduces new kinds of tasks through a series of training levels. The different kinds of tasks are the classes of glitch to be classified.

New volunteers start in Level 1, in which they are presented with glitches to classify that are expected to be of one of only two distinctive classes—“blips” vs. “whistles” in the current system—and given those two choices or “none of the above” in a simplified version of the classification interface. Once the volunteer can successfully classify glitches of the initial two classes (currently assessed by accuracy in classifying gold-standard data), the volunteer can advance to the next training level, in which they see glitches of additional classes. In other words, to scaffold volunteer learning, the system gradually expands the number of classes presented to the volunteers. Table 1 shows the Gravity Spy training levels at the time of this study, with the number of glitch class options and the names of the classes presented in each level.

Once volunteers have completed all rounds of training, introducing the classes of glitches, they are considered fully qualified and are given glitches to classify at varying levels of ML
Table 1 Gravity Spy glitch classes by training level.

<table>
<thead>
<tr>
<th>Level</th>
<th>Glitch Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (3)</td>
<td>blip, whistle, none of the above</td>
</tr>
<tr>
<td>2 (6)</td>
<td>blip, whistle, koi fish, power line, violin mode, none of the above</td>
</tr>
<tr>
<td>3 (10)</td>
<td>blip, whistle, koi fish, power line, violin mode, chirp, low frequency burst, no glitch, scattered light, none of the above</td>
</tr>
<tr>
<td>4 (20)</td>
<td>blip, whistle, koi fish, power line, violin mode, chirp, low frequency burst, no glitch, scattered light, helix, 45Mhz light modulation, low frequency noise fluctuations, paired doves, 50hz, repeating blips, scratchy, tomte, wandering line, extremely loud, none of the above</td>
</tr>
</tbody>
</table>

We expect this ML-scaffolded process to support better learning (i.e., to help volunteers become more accurate at classifying) for two reasons. First, because the ML has a high confidence in the classification of the glitches, it is most likely that they are of the identified class and so will be exemplary glitches that will help the volunteer to learn how to identify that class. Second, focusing attention initially on just a few classes enables volunteers to master those classes before adding complexity (i.e., staying in the volunteer's zone of proximal development).

And as the ML can be wrong, it is still useful to the project to have a human judgment even for glitches for which the ML reports high confidence. For example, when a new class of glitch appears in the data, the ML will attempt to classify them as one of the known classes. It has happened that the new glitches are confused with an existing class, resulting in incorrect ML classifications with high confidence. These classification errors can be corrected even by new volunteers.

The first three types of learning support (tutorials, feedback and inclusion of prototypical and exemplary images in the interface) are standard in on-line citizen-science projects, but the fourth approach (scaffolded introduction to classes) is novel for citizen-science projects. Furthermore, while scaffolding is a well-accepted approach, implementing scaffolding by using ML to select among real tasks is, as far as we know, novel. Assessing the success of the ML-scaffolded approach is the focus of the experiment reported in this paper. For conciseness, we refer to it as Machine-Learning Guided Training (MLGT). We hypothesize:

**H1**: Volunteers who go through the MLGT will be more accurate in their classifications than volunteers who do not go through the MLGT.

2.3 Motivation to contribute

In addition to learning, we expect the MLGT to improve volunteer motivation. As citizen-science projects depend entirely on the contributions of volunteers, volunteer motivation has been a consistent topic of research and researchers have identified a range of motivations for
citizen science. For example, surveys and interviews show that citizen science volunteers are motivated to participate in projects by the opportunity to make a contribution to science [3, 28, 29, 41]. Accordingly, presenting volunteers with authentic tasks should be more motivating than providing training tasks. Following [9], we consider motivation for an initial contribution and for sustained contribution separately.

For the first, citizen-science projects, like most on-line production community systems, exhibit a highly-skewed distribution of contribution: a few volunteers contribute a lot while many contribute only a little. Indeed, many new visitors to a project do not contribute at all but rather leave before making a classification. We hypothesized that new volunteers seeing a complex task feel overloaded and discouraged from contributing. For example, the full Gravity Spy interface offers 20 options (since increased to 22), some with fairly subtle distinctions. A new user could feel unable to perform the task accurately or at all. This problem is not unique to Gravity Spy: the Zooniverse Snapshot Serengeti project for example, asks volunteers to identify animals shown in images into one of 54 species, many of which would be unfamiliar to a novice. We expected that the MLGT, with its scaffolded design introducing volunteers to the classes a few at a time, will be less challenging and thereby more inviting. We therefore hypothesize:

**H2**: Volunteers who go through the MLGT are more likely to classify more subjects than volunteers who do not go through the MLGT.

We expect a scaffolded approach to presenting new tasks to also motivate continued participation. The system is designed to appeal to volunteers’ sense of accomplishment. The initial Gravity Spy page (Figure 2) shows all of the training levels, but volunteers can only access the ones they have “unlocked” by successfully completing the lower levels (Note that volunteers are free to choose to work on any of the unlocked levels, not just the highest one.). The system also provides an encouraging messaging when mastery at the current level is achieved and the next level is unlocked. In a sense, the MLGT slightly gamifies the project by posing levels of possible accomplishment [37], which we expected to motivate volunteers to continue to contribute. [2, 19] propose that gamification can be an effective motivator for citizen scientists. We therefore hypothesize:

**H3**: Volunteers who go through the MLGT will contribute more classifications than volunteers who do not go through the MLGT.

**H4**: Volunteers who go through the MLGT will contribute for a longer period of time than volunteers who do not go through the MLGT.

![Fig. 2](image-url) Levels for the Gravity Spy project. When the buttons are shaded (e.g., Level 4: Black Hole Merger and Level 5: Universe Cosmic Background) a volunteer cannot access the level.
3 Experiment Design

To test the hypotheses developed above, we conducted a randomized controlled on-line experiment in the Gravity Spy project.

3.1 Procedure

The experiment tested the impact of the MLGT on volunteer accuracy and contribution. During the experimental period, half of new participants were randomly assigned to the treatment condition and the other half to the control condition. The exemplar glitches in the field guide and the use of feedback on gold-standard data were the same in the treatment and control conditions.

The treatment condition for the experiment was the MLGT described above. Volunteers in the treatment condition start in Level 1 and were promoted to higher levels depending on their performance. As they were promoted, classification options were introduced a few at a time, expanding the number of options and exemplary glitches shown in the classification interface and the classes of gold-standard data and glitches presented. Additionally, each level includes its own tutorial to gradually introduce features of the system that aid in the classification task and information about the new glitch classes.

As the control condition, we assigned volunteers to a slightly modified version of the Gravity Spy Level 4 (labeled M.A.), in which volunteers see all glitch classes and their exemplary glitches as options for classification and are given gold-standard data and glitches to classify that may be of any class. The modification needed to the level was to expand the tutorial to include content included in the tutorials from lower levels, to ensure that the tutorial content was the same in both conditions; presented in four parts in the treatment and all at once in the control.

The modified Level 4 was used as the control condition as it matches the approach taken in other on-line citizen-science projects in the Zooniverse that make all options available to all volunteers from the beginning of their participation. However, the design of the Zooniverse system is such volunteers could choose to classify in a lower level even if they were initially assigned to classify at Level 4 (see Figure 2).

Subjects for the experiment were volunteers who joined the Gravity Spy project during the experimental period from 30 October 2016 to 19 December 2016. When Zooniverse volunteers created an account or logged into their account for the first time after the experiment launched, they were randomly assigned to either the control or to the treatment group. Volunteers retain this assignment when they visit the project on future sessions. To assess the impact of the treatment on a volunteer, it is necessary that they receive the treatment from their initial interaction with the Gravity Spy system. Therefore, we did not include data from volunteers who had already contributed to the Gravity Spy system before the start of the experiment.

3.2 Data Collection and Analysis

Three outcome variables were used to test the hypotheses and to test the effectiveness of the MLGT: accuracy, classifications and sessions.

- **Accuracy** is the fraction of the volunteer’s classifications of gold-standard data that agree with the science team’s classification.
– **Classifications** is the total number of classifications a volunteer contributed to the project during the experiment period.
– **Sessions** is the number of occasions on which a volunteer contributed.

Accuracy is used to assess learning while classifications and sessions show motivation.

Contributing over multiple sessions is of interest as it shows sustained contribution to the system, rather than just trying it out. The intuition behind the definition of a session is that volunteers often log in to the system, contribute for some time and then take a break, e.g., until the next day. As volunteers do not always log out of the system when they are done classifying, we defined sessions by looking at the gap between classifications. A session is defined as a set of classifications where the gap between the classifications is not greater than thirty minutes (the threshold used in previous studies of similar citizen-science projects [34]). A gap of more than thirty minutes indicates the start of a new classification session.

All statistical analyses were performed using R Studio. For the comparisons between the control and treatment groups, we first assessed variables for normality using the Shapiro-Wilk test. Data that were normally distributed were compared using independent samples t-tests. The independent samples t-test is a standard group test to determine whether there is statistical evidence that the associated population means are significantly different. When the data were non-normal, we used the nonparametric Mann-Whitney-Wilcoxon test. The Mann-Whitney-Wilcoxon test determines whether the population distributions are unlikely to be identical.

### 3.3 Ethics Review

The experiment protocol was reviewed by our university’s human subjects institutional review board (IRB). The experimental procedure posed minimal or no risk to the participants, as the control process was the process used in nearly all other citizen-science projects on the Zooniverse and the treatment was the same as used by default in Gravity Spy. We did not collect any data about the subjects; only the count and timing of the classifications they did and the agreement of those classifications with gold-standard data. Indeed, the site does not collect demographic information of any kind and volunteers are identified only by a self-selected volunteer ID. A section of the initial volunteer agreement provided when volunteers sign up for a Zooniverse account is a disclosure that site administrators run experiments to improve the system and volunteer experience. As collecting informed consent would require volunteers to provide identifying information that was not otherwise collected, we were permitted to run the experiment without requesting specific informed consent for this experiment.

### 4 Results

The chart in Figure 3 shows the flow of new volunteers through the experiment. After the experiment had been run, we discovered an omission in the data collection. It appears that the system assigned volunteers to control or treatment when they first visited the site, but did not record the assignment until they reached the classification page. As a result, volunteers who dropped out while viewing the tutorial seem not to have been recorded in the system. As assignment was random, we believe that roughly equal numbers of volunteers were assigned to control and to treatment, but we ended up with unequal numbers of new volunteers in the control and treatment groups. Further, we do not have an exact count of volunteers
who started the experiment. The final population of new volunteers and the population we
analyzed was 386: 246 volunteers in the treatment and 140 volunteers in the control.

![Flow chart showing how many volunteers visited Gravity Spy, created an account, and were assigned
to either the treatment or control.](image)

**Fig. 3** Flow chart showing how many volunteers visited Gravity Spy, created an account, and were assigned
to either the treatment or control.

4.1 Descriptive statistics

222 of the treatment and 99 of the control group volunteers contributed classifications. Figure 3 shows how many of the volunteers made contributions at each workflow level. Note
that in the treatment condition, contributing at one level is a prerequisite for being able to
contribute at a higher level, so the counts are cumulative. That is, 222 volunteers in the treat-
ment condition contributed at Level 1; 104 of these did enough classifications to graduate
to Level 2, while 118 did not; and so on up to 30 volunteers who contributed at Level 4. In
contrast, volunteers in the control condition started in a modified Level 4 (M.A.) but had the
option to try a lower level at any time, meaning that these counts are not cumulative.

4.2 Hypothesis 1: Learning

We first report on the effect of the GTS on volunteer classification accuracy, shown in Ta-
ble 2. We assessed each volunteer’s accuracy by examining whether their answers agreed
with the science team answers for the gold-standard data. However, of the 321 volunteers
who did classifications, 160 did not see any gold-standard data, decreasing the sample size
in both the control (N = 46) and the treatment (N = 115). As hypothesized, the average ac-
curacy was significantly higher in the treatment group. The average level of agreement with
gold-standard data was 60% (SD = 35.9) for volunteers in the control and 95% (SD = 9.1)
for volunteers in the treatment. A Mann-Whitney-Wilcoxon test indicates the difference is
significant, W = 2395.5, p < 0.001. As a result, H1 is supported.
### Table 2
Volunteer accuracy on gold data in control and treatment groups, overall and for just level 4.

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>Mann-Whitney</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy on Gold Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Levels</td>
<td>90% (SD = 1%)</td>
<td>54% (SD = 23%)</td>
<td>103.5***</td>
</tr>
<tr>
<td>N = 115</td>
<td>N = 46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levels 4 &amp; M.A.</td>
<td>56% (SD = 9%)</td>
<td>46% (SD = 25%)</td>
<td>t(56.82) = -2.09*</td>
</tr>
<tr>
<td>N = 15</td>
<td>N = 45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** = p < 0.001, ** = p < 0.01, * = p < 0.05.

#### 4.2.1 Gold Data for Level 4 and Modified Apprentice Workflow.

The analysis above includes data for all workflow levels. However, note that in the initial training levels in the MLTS, volunteers select from only a subset of the classes, which could explain the higher accuracy. To address this bias, we compared accuracy for classifications done in the two groups at Level 4, in which there are the same number of options. In the control group, which starts at Level 4, 45 volunteers saw gold-standard data. Of the 30 volunteers in the treatment who reached Level 4, 15 saw gold data. We found 46% (SD = 25%) agreement in the control and 56% (SD = 9%) in the treatment group. While the distribution of accuracy scores for all project classifications did not follow a normal distribution, the distribution of accuracy scores for Level 4 and MA classifications did, so we analyzed these accuracy scores using the parametric independent samples t-test. We found the difference in accuracy to be statistically significant at t(56.82) = -2.09, p = 0.04.

#### 4.3 Hypotheses 2: Initial contribution

Our second hypothesis was that volunteers who went through the MLGT are more likely to contribute classifications. In the experiment, 41 (30%) of volunteers in the control did not classify versus 24 (10%) volunteers in the treatment. We conducted a test of proportions to determine whether the number of volunteers classifying in each group was significantly different. The results of the chi-squared (χ²) revealed volunteers in the treatment were more likely to make an initial classification χ²(1) = 37.84, p < 0.001. Accordingly, H2 is supported. In addition, we believe that the dropout rate during the initial tutorial was much higher for the control group compared to the training group, as reflected in the different final sample sizes. Unfortunately, we do not have the data on the size of the dropout to test the hypothesis at this point in the volunteers' interaction.

#### 4.4 Hypotheses 3: Contributions

Our third hypothesis was that volunteers who went through the MLGT would contribute more than those who did not. We found that volunteers in the treatment group contributed many more classifications than volunteers in the control: the average number of classifications for volunteers in the control group was 121.1 (SD = 722.7) compared with 228.2 (SD = 677.8) classifications in the treatment group (Table 3). The results of the Wilcoxon rank sum test revealed a significant effect of the MLGT on the total number of classifications volunteers contributed (W = 7609.5.5, p < 0.001). Accordingly, H3 is supported.
### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>Mann-Whitney</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classifications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Levels</td>
<td>228.2 (SD = 677.8)</td>
<td>121.1 (SD = 722.7)</td>
<td>6770.5***</td>
</tr>
<tr>
<td>N = 222</td>
<td>N = 99</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sessions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Levels</td>
<td>2.5 (SD = 5.8)</td>
<td>2 (SD = 5.9)</td>
<td>8783**</td>
</tr>
<tr>
<td>N = 222</td>
<td>N = 99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *** = \( p < 0.001 \), ** = \( p < 0.01 \), * = \( p < 0.05 \)

4.5 Hypothesis 4: Sessions

Our final hypothesis concerned the duration of engagement with the Gravity Spy project. The data on the number of sessions suggests that the MLGT increased interest in the project. Volunteers in the control contributed on average 2 (SD = 5.9) sessions while volunteers in the treatment contributed in an average of 2.5 (SD = 5.8) sessions (Table 3). A Mann-Whitney-Wilcoxon test indicates that the distribution of sessions in the training and control groups are significantly different, \( W = 8783, p = 0.005 \). Accordingly, H4 is supported.

5 Discussion

The research results add to our knowledge about how training design can contribute to learning and motivation in on-line production communities. Building on prior research on learning and motivation, we proposed that a scaffolded introduction to the work of the Gravity Spy project would be more beneficial and motivating for volunteers (and the project) than simply having volunteers “thrown in the deep end.” Further, we implemented an approach to scaffolding in which the materials provided to newcomers were selected by an ML classifier from the actual tasks of the project, rather than being curated by an instructional designer. Overall, the results show that this approach had the hypothesized impacts of improving volunteer accuracy (H1), increasing conversion to a contributor (H2), increasing the number of classifications (H3) and improving retention (H4).

First, we discuss the role of scaffolding in support of learning and motivation of newcomers. Second, we discuss the implication of our finding for the design of informal on-line learning environments in particular.

5.1 The Benefits of Scaffolding Work for Learning and Motivation

The literature on learning systems has increased in calls for scaffolding access to materials suggesting learners in on-line settings might find materials confusing and thus increase attrition [44]. In computer-supported collaborative learning environments, scaffolding has been shown to enhance participant learning [?] and support community level benefits such as increasing the volume of contribution and retention among learners [?]. The study we presented above showed similar findings, however, our main contribution is that scaffolding has similar effects when materials are served to learners using ML-supported training.
From a learning theoretical perspective, one can conceptualize this process as 1) legitimate peripheral participation where newcomers gradually expand their access to central activities [30]; and 2) an ML-supported approach that introduces work fitting each participant’s zone of proximal development [13]. While these two perspectives do not exclude one another, we can argue that the first highlights the phased introduction of the tutorials, prototypical and exemplary glitches and feedback on gold-standard data. The second emphasizes the ML-supported introduction of work fitting the participants’ level of skill. The research findings do not indicate if leveling or gradual introduction of more difficult glitches matters the most. Future research may help untangle the benefits of these different design principles.

When it comes to motivation, the work design literature suggests that motivating work exists at the intersection of familiarity and challenge [18]. We suspect that the ML-supported selection of glitches helps approximating the right level of familiarity and challenge. Again, we are not certain whether it is the phased introduction of work, organized into levels or the ML-supported selection of image difficulty that play the most important motivational role. Future research can parse out the effects of these two design choices.

Interestingly, we found that a number of volunteers in the control group who started in the modified Level 4 chose to contribute at lower levels after the experiment finished, as shown on the right side of Figure 4. For instance, 28 volunteers contributed in Level 1 after having classified in the modified Level 4. We do not know for sure whether these volunteers were motivated by the “leveling” or the introduction of easier glitches in earlier phases, but we suspect the latter explanation. Volunteers who have worked at Level 4 have already been presented with a comprehensive tutorial and collection of prototypical glitches, so there should be no need to go back to experience the same material being introduced in smaller chunks. Likewise, there should be little appeal to those motivated by levels to move backwards. As a result, the most convincing explanation is that volunteers shifting to the lower levels are seeking work matching their zone of development, neither too hard nor too easy.

Beyond the positive findings associated with learning and motivation among volunteers, the Gravity Spy design offers benefits to the science team behind the project. First, scaffolding allowed volunteers to work on real data during their training and not solely on pre-classified glitches. As a result, participants start contributing to the science work from the very beginning. They do not have to complete the training before they become productive members of the project.

5.2 Training Contributors in on-line Production Communities

Transforming non-experts into high performing contributors remains a challenge for many on-line systems. While our results are promising, they also point to future research opportunities.

First, volunteers come to on-line communities with varied competencies. Some newcomers might arrive with more background knowledge on the task or be more proficient learners. For example, amongst citizen scientists, many people contribute because of a prior interest in science [20,45]. Therefore, participants would likely benefit from a personalized tutoring system that starts at their current level rather than from scratch [24].

To properly target training requires an estimate of a volunteer’s current level of knowledge. Few citizen-science projects evaluate volunteers’ knowledge level at all. Those that do generally rely on proxies, such as the number of classifications contributed. These are quite crude measurements of volunteers’ skill level. In the present project, we rely on responses
Bayesian methods offer a promising approach to modelling user knowledge as they can incorporate prior knowledge about a volunteer and update it from experience. Such models are widely used to improve the performance of ML systems and human learning [25, 49, 51]. For instance, Corbett and Anderson’s [8] Bayesian knowledge tracing (BKT) model has been applied to model learning in tutoring system as students practice different skills. In our setting, such models could use responses to both gold and non-gold data. However, the citizen science context has to account for the possibility that the ML classification might be incorrect, rather than the volunteer’s classification.

Knowing the volunteers’ level also opens up new possibilities for interpreting their contributions, specifically for making decisions about the class of a glitch. It should be possible to achieve confidence in the collective assessment with judgments from fewer experienced participants than novices. Similarly, images that are confidently rated by the ML might be retired with fewer classifications or with classifications from less experienced volunteers. We also suspect that asking users to classify a set of pre-defined glitches to assess their performance (see [51]) might also be useful for personalized learning.

Second, volunteers learn from resources beyond engaging in the task and receiving feedback from their tagging of gold-standard data. It is well-known in learning systems that formative feedback is important for learning. However, creating meaningful opportunities for feedback remains a challenge [15]. As indicated by the learning literature, a number of resources can help participants master a task. In the context of Gravity Spy volunteers may engage with FAQs, tutorials and comment forums to learn the practices and norms for contribution. However, these resources can be voluminous and disorganized, making it difficult for volunteers to know which are relevant to them at the given time. Also, they are relatively static and less personalized. As a result, volunteers could benefit from a scaffolded access to project resources in addition to the introduction to the tasks. To implement such an approach requires a better understanding of the types of resources participants find helpful at various stages of their participation. For instance, it seems intuitive that tutorials would be most effective with newcomers while FAQ and comment forums would predominantly benefit more advanced participants, but these intuitions should be tested with data. Additionally, given the rich conversations that occur via discussion for, one might explore the extent to which these materials can act as feedback opportunities. Several studies have experimented with this form of feedback as an intelligent tutoring system e.g., [12, 44]. For instance, Easterday et. al., [12] suggests several features for a crowd-based design critique system where designers learn through formative feedback from peers that might be applicable in this context.

Finally, our focus in this paper has been entirely on how volunteers learn to do the image classification task, but as we noted, volunteers may learn and be motivated by learning about the topic of the project or about science more generally. This kind of learning can be supported by the additional project resources noted above, but we as yet know little about what and how volunteers learn from these materials and how best to support them.

5.3 Limitations

Field experiments in on-line production systems pose challenges that may limit researchers’ inferences about volunteers and the community of the study. The research presented here is no different. While the true experimental design does control many threats to internal validity, there are two caveats.
The most apparent limitation here is the nature of the assignment of volunteers to the conditions of the study. First, the system did not correctly record assignment of volunteers to condition, meaning that our analysis starts part way into the study. We believe, but do not have data to show that more volunteers dropped out in control condition than in the treatment condition. However, this threat, known as experimental mortality, would be expected to leave the control condition with more motivated volunteers than the treatment, and so does not explain our findings that the treatment group seemed more motivated. In other words, while it is unfortunate that the experimental data were not completely captured, this lacuna does not compromise the main contribution of the study.

Further, the system did not prevent volunteers in the control group from contributing at other levels and in fact, 18 control group members at some point did visit other levels. This threat is known as design contamination, meaning that some of the control group received the treatment. This contamination would make the control and treatment groups more similar than otherwise, so again, this threat does not explain our findings. Indeed, as both threats tend to reduce the difference between the control and treatment groups, the actual effect of the treatment may be greater than we observed.

The second limitation concern the measurement of volunteer accuracy. We could not control the frequency with which volunteers see gold data or which class of glitch is shown. As a result, some volunteers did not see any gold data and so could not be included in our analysis of accuracy. However, this omission should affect the control and treatment groups equally.

Finally, the trade-off for a design with strong internal validity is weaker external validity. We have shown that our training approach works in the Gravity Spy setting, but can not say for sure how the approach will work elsewhere. However, the underlying theoretical rationale for the approach suggests that it could be useful in citizen science projects more generally and perhaps for other kinds of on-line communities. Furthermore, the training has a number of parameters, e.g., how many classes to introduce and after what level of performance. The experiment has tested only point in this design space, and so does not provide insight into the optimal settings.

6 Conclusions

In summary, we have presented an approach to newcomer training that offers learners ML-selected tasks in an attempt to fit their zone of proximal development, work that is not too easy or too difficult. Our experiment shows that this approach is successful in increasing the accuracy of the volunteers while also increasing motivation and contribution. Even these initial classifications are useful to the project, as ML assessment are not perfect and so need to be checked.

This approach to training addresses the dilemma faced by on-line communities in particular, as making good use of newcomers’ contributions is important in setting where many volunteers only contribute a few times. Equally important, this approach to training scales to large numbers of participants who can engage in on-the-job training without requiring more experienced workers to evaluate the work quality.

Although our focus has been on learning in an on-line production community, it should be possible to apply the approach to other settings in which many newcomers need to learn to perform a variety of tasks. The main limitation is the need to train an ML-model to do the task. However, ML technology is rapidly improving and being applied to more kinds of work, suggesting that there will be many future applications. Often, the hope is to use
the ML to complete automate the task, but in many cases, this hope may be too optimistic. The approach presented here offers a path to creating a collaboration between human and machine learning that takes advantages of the strengths of each while enabling both to learn, improve and contribute.

Acknowledgements

References