CITIZEN SCIENCE:
THEORY AND PRACTICE

Dickinson, JL and Crain, R. 2019. An Experimental Study of Learning in an Online Citizen Science Project: Insights into Study Design and Waitlist Controls. *Citizen Science: Theory and Practice*, 4(1): 26, pp. 1–13. DOI: https://doi.org/10.5334/cstp.218

RESEARCH PAPER

An Experimental Study of Learning in an Online Citizen Science Project: Insights into Study Design and Waitlist Controls

Janis L. Dickinson and Rhiannon Crain

The field of citizen science needs experimental studies to develop a better understanding of the connection between citizen science participation and learning. We conducted an online experiment to test whether social interaction and the primary learning activity, mapping, led to increased content-learning in the citizen-science project, YardMap. Participants were randomly assigned to three treatment/control groups: the full, socially-networked mapping project, the mapping project with its social tools disabled, and a waitlist-control group, whose members took the pre-test and post-tests along with the active participants, but were blocked from participating in the project for a two-month interval between the tests. Based on general linear models, post-minus-pre-test scores (learning gains) did not differ between the two treatments, nor between the treatment and control groups. Individual level of activity in the project did not affect learning gains, but the learning gains were negatively associated with pre-test score for all three groups, indicating that learning occurred, not only in the treatment groups, but also in the control group. Retrospective analysis of participant completion rates, effort, and responses focused on developing possible explanations for this outcome. This analysis uncovered factors that reduced the ability to detect learning, which may be common to many Citizen Science research studies. Of particular interest were factors related to the use of a waitlist control design in online settings. This research concludes with new recommendations for the design of controlled studies of informal science learning, including new controls to explore the mechanisms by which waitlist control participants learned as much as participants in the treatment groups.

Keywords: citizen science; social learning; waitlist control; online learning

Introduction

Controlled studies are needed to determine whether citizen science projects meet the specific learning objectives for which they are designed (Phillips et al. 2012; Phillips et al. 2018; Wells and Lekies 2012). In the context of conservation, such learning objectives can include interest in science and the environment, self-efficacy, understanding of the nature of science, motivation, science inquiry skills, and behavior change/stewardship (Phillips et al. 2014, 2018), but they also include learning relevant content, including facts and concepts that serve as the basis of a particular area of inquiry. Observational studies that provide evidence for conceptual- and content-learning have tended to conclude that learning arises due to participation in the citizen science project, but these conclusions are often based on "snapshot" survey methods or pre-post testing without controls. One reason controls are needed in citizen science contexts is that, unlike students in classrooms, citizen science participants are self-selected learners who have many free-choice learning outlets. This makes it difficult to separate learning that happens as a consequence of participating in a project from learning that happens as a consequence of preexisting motivations that bring participants to citizen science.

For example, Land-Zandstra et al. (2016) pointed out that one of the primary motivations for participation in the iSPEX air monitoring project was a preexisting interest in air quality and its impact on health and the environment. That preexisting interest might have driven learning on its own even in the absence of participation in the project—or it might not have. Use of control groups can assist in disentangling such interactions. In projects with prolonged engagement, controlled studies and robust longitudinal data that track participant trajectories are especially important for making strong inferences about learning (Masters et al. 2016).

Waitlist controls have been used to good effect in health intervention studies, but they have seen little use in citizen science projects (Wells and Lekies 2012). Waitlist controls involve random assignment of participants to learning treatment(s) or a waitlist control group immediately after they have joined a project and can help to separate learning that occurs due to the learning intervention (participating

Cornell University, US
Corresponding author: Janis L. Dickinson (jld84@cornell.edu)

in a project) from temporal changes in knowledge due to other factors (Ferguson et al. 2012; Roeser et al. 2013). In citizen science research, this means that the waitlist control participants are excluded from participating in the project until after they have taken both the pre- and post-tests. If a project's internal content and activities are driving learning, we predict that treatment participants will demonstrate greater gains in learning outcomes than do waitlist control participants (**Table 1**). On the other hand, if the motivation that brings participants to citizen science is sufficient to drive learning without actual engagement in the practice of science (Jennett et al. 2016; Raddick et al. 2010; Rotman et al. 2014), we should see no difference in learning outcomes between treatment and control groups, meaning that they should show a similar shift in conceptual or content knowledge between the pre- and post-test.

Projects delivered on the Web have additional issues. Learning resources are widely available on the Web, both within and outside of Web-supported citizen science projects. If participants in Web-supported citizen science projects are already seeking out information online before joining, it may be especially important to use waitlist controls. Like other controlled studies, waitlist-controlled experiments can be designed with multiple treatments to test which aspects of a project influence learning (e.g., data collection and reporting, engaging with others in the project, and access to learning materials). For example, one observational study concluded that learning varied with involvement in the social aspects of a project but was not affected by actual participation in science practice, suggesting that participants who merely engaged in a supported and scientifically oriented forum learned as much as those participating in the actual citizen science activity (Jennett et al. 2016). Given that a distinguishing premise of citizen science projects is that practicing science increases the potential for learning (Bonney et al. 2016; Phillips et al. 2014), it is important to test the hypothesis that the citizen science activities themselves play a critical role in learning (Phillips et al. 2012; Wells and Lekies 2012). We argue that while experiments are difficult and costly (they take large amounts of Web-programmer and Web-designer time to carry out), online citizen science is ripe for robust tests of learning impacts, such as can be derived from experimental studies.

Waitlist controls have been used to examine online learning in psychological contexts; however, the studies we found tended to involve emotional, cognitive, or behavioral outcomes rather than measures of content learning or conceptual learning. Examples include studies of (1) the effects of learning modules on taking medication, managing stress, and sleep quality in people with epilepsy (Dilorio et al. 2011), (2) the effects of online tutorials on the effectiveness of doctor-patient communications (Heiman et al. 2012) and arthritis patients' self-care (Fary et al. 2015), and (3) the effects of computerized, cognitive remediation on verbal learning and processing speed in schizophrenics (Sartory et al. 2005). In the outdoor education domain, an 18-month controlled study provided modest evidence of learning in school gardens; however, this study did not involve online learning, nor did it involve self-selected participants who were engaged in the practice of citizen science data collection (Wells et al. 2015). Controlled studies and robust longitudinal data are especially needed for strong inference about content and conceptual learning, which can be supported through many outlets (Masters et al. 2016). We have found no controlled studies of conceptual or content learning in citizen science projects that use the Web to crowdsource data collection, despite a proliferation of such projects over the past 10–15 years.

Here we examine data from a randomized online experiment that varied both the mode of participation in the Cornell Lab of Ornithology's YardMap citizen science project (two different learning treatments) and also used a waitlist control. YardMap (more recently renamed "Habitat Network") is a socially networked mapping project that engages participants in learning about and creating visual maps of a wide range of sustainable practices they undertake to help wildlife and reduce carbon emissions, usually in and around their own homes (**Figure 1**). At the time of this study, the project had internal content focused on

Table 1: A priori hypotheses and predictions driving experimental treatments or inclusion of additional explanatory and control variables.

	Hypothesis	Prediction
1)	Social interaction fosters engagement in YardMap.	Participants in the social version of YardMap will be more active in the project than participants in the non-social version and will login more times.
2)	Based on activity theory (Krasny and Roth 2010), mapping and identification of sustainable practices increases content knowledge	Post-pre-test differences will increase in the mapping (only) and social mapping treatments compared to the waitlist control.
3)	Activity in the project (a measure of effort) will increase learning.	The number of logins into the project will be positively associated with the post-pre test difference.
4)	Based on theories of social learning, social interaction within YardMap will increase content knowledge more than will non-social mapping.	Post-pre-test differences will be greater for participants using the fully social mapping application than for participants using the mapping application stripped of social tools.
5)	The amount of learning that can be detected is lower for higher pre-test scores.	Pre-test score will be negatively associated with post-pre differences in content knowledge.

managing residential habitat to support birds and pollinators and was connected to eBird, a worldwide bird monitoring tool, for analysis of impacts of backyard practices on bird occupancy. Its social components included a visible newsfeed, a forum, and tools that enabled participants to view and comment on their own and each other's maps.

We were interested in measuring learning occurring within the project as a whole, but also sought to explore the possibility that the social interactions enabled by YardMap enhanced learning. **Table 1** lists our a priori hypotheses together with predictions about how different forms of participation would influence the changes in content knowledge and conceptual knowledge of participants. Because there is an inherent ceiling effect in which participants with high pre-test scores cannot have as large a knowledge gain as those with low pre-test scores, we predicted that it would be necessary to control for pre-test score for each content area tested.

We present the results of this study using measures of factual knowledge (Bird- and Tree-IDs) and conceptual learning (ecological concepts). Based on the results, we conclude with recommendations that should prove useful to future researchers tackling the hard problems of designing controlled studies and online experiments to investigate learning in citizen science projects.

Methods

Experimental design

This study was conducted under the guidance and approval of the Institutional Review Board for Human Participants (IRB) at Cornell University under protocol #0906000455. The main hypotheses (**Table 1**, items 1 and 2) led us to use a true experimental design that involved a waitlist control and delivered two versions of the project to treatment participants: A social version and a non-social version. The social version allowed participants to view others' map summaries, read participant comments in a newsfeed, access all prior comments in a forum, and comment on specific features of their own and each others' maps. As shown in **Figure 1**, participants could view different aspects of each other's YardMap practices: (A) examine in detail each other's maps to see how they were modifying habitat, (B) view statistics about each others' habitat breakdown, (C) examine site characteristics, including practices such as pesticide use and herbicide use, and (D) view the list of birds recorded in other participants' mapped areas (see **Table 2**). In contrast, the non-social version was created by disabling the project's commenting, site-viewing, and newsfeed/forum features so that participants were mapping in isolation from others. The two treatment groups and the control participants had access to the "Learn" pages and infographics on the YardMap site as well as the rest of the Lab of Ornithology's Web properties, except for those few properties (e.g., FeederWatch) that resided behind a paywall.

To drive recruitment to the YardMap citizen science project, we arranged for as much publicity as possible during the two months of recruitment into the study (May 2014 through June 2014). Publicity included articles highlighting the project on social media for *Birds & Blooms* magazine and the Cornell Lab of Ornithology as well as paid "boosts" of Facebook posts. Individuals who created a new

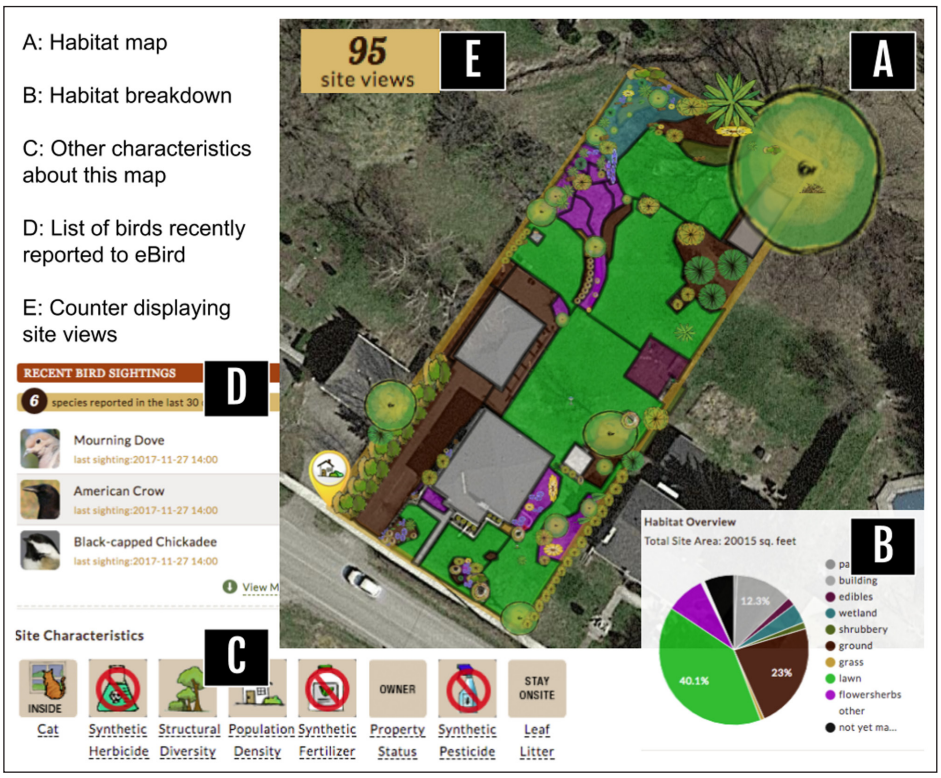

Figure 1: Example of participant-generated YardMap. Participants in the social version can view each other's yard practices (site characteristics), habitat types, and birds seen as well as forum comments and the news feed of all comments.

login, and thus were new to YardMap during the dates of the recruiting period, were invited to participate in the study via a pop-up that appeared after account creation. Although individuals joining YardMap were not required to participate in the study, they were informed about the study and encouraged to participate with incentives consisting of a chance to win one of ten $100 Amazon gift cards or an iPad mini from Apple.

Once study participants gave their informed consent, they were asked to fill out a pre-test consisting of 56 questions, including questions about content (Bird-ID and Tree-ID) and ecological concepts. Our Web application assigned participants at random to one of the two "active" treatment groups (social/non-social) or the waitlist control condition. The social treatment group was exposed to YardMap with the social networking features enabled while the non-social treatment group was exposed to YardMap with the social networking features disabled (see **Table 2**). Members of the waitlist control group were required to wait for eight weeks before participating in the project, while members of each of the two treatment groups were immediately directed from the online survey into the YardMap Web application and encouraged to participate. Those assigned to either the social treatment or the non-social treatment were placed in their respective versions of the YardMap Web application, while those assigned to the waitlist control were directed to a Web page telling them about the waitlist control and showing them a persistent countdown clock indicating how long their participation in YardMap would be delayed (**Figure 2**).

Two months after the date on which they completed their pre-test, active and waitlist control participants were sent an email invitation to take a post-test consisting of 43 questions (without the demographic questions that were

Table 2: Different kinds of experiences available to participants in the social vs. non-social version of the YardMap Web Application.

Social treatment	Non-social treatment
Participants can peek at others' maps by browsing and clicking items in a shared map interface.	Participants can view points representing maps in a shared map interface but cannot peek at others' maps.
Participants can comment directly on each other's maps and items within those maps.	Participants cannot see maps or objects and cannot comment.
Participants can access the forum in YardMap (i.e., can post, like, comment, share, follow) and see the newsfeed in their own map page.	No forum access.
Participants can access learn articles and infographics.	Participants can access learn articles and infographics.

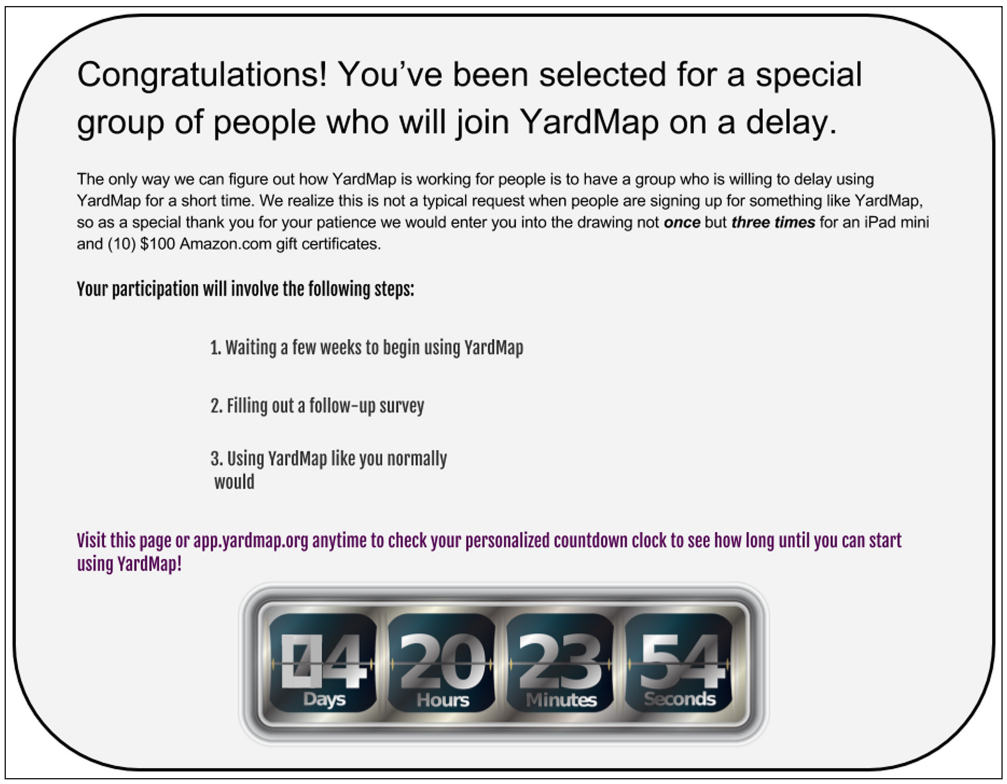

Figure 2: The message that waitlist control participants saw on the YardMap Website after they took the pre-test. This message remained when they signed in until they completed the post-test about 8 weeks later.

on the pre-test). Questions about Bird-ID, Tree-ID, and ecological concepts were identical to those given on the pre-test. In the Web application, the invitation to complete a post-test overlaid the YardMap Web pages ensuring that participants did not go on using YardMap longer than the eight-week study period. Participants could not dismiss the invitation until the post-test was completed. Once participants completed the post-test, they were sent into the standard social version of the YardMap application.

Description of survey participation

Pre-test

A total of 4,390 individuals signed up for YardMap during the period from May 1 through June 30, 2014. Of those, 2,371 (54.1%) opted into the study and started the pre-test. A total of 1,611 (or 68% of those opting in) completed the pre-test, which had 56 questions, some with multiple items. For a period of about 2 weeks (May 20, 2014 14:46 EST to June 4, 2014 15:24 EST), Qualtrics, the survey software used to conduct both the pre- and post-test, experienced a bug in which 386 participants opting into the study were given the post-test survey instead of the pre-test survey and, if they completed the survey (n = 356), they were assigned non-randomly to the social version of the app. Removing participants who did not complete the pre-test or were not assigned randomly (356) left a total of 1,255 participants, 427 individuals in the waitlist control, 398 in the non-social treatment, and 430 in the social treatment.

Post-test

The mean interval between starting the study (finishing the pre-test) and finishing the post-test was 59.79 ± 0.6 (SEM) days. Seventeen of the 43 questions asked on the pre- and post-surveys are included the analyses presented in this paper (Appendix I, Supplementary File). A total of 650 (51.8%) of the 1,255 randomly assigned participants finishing the pre-test within 50 minutes also filled out the post-test (control n = 296, non-social n = 158, social n = 196). These individuals constituted the sample that we used to analyze the post-pre test outcomes. Missing data for one or more explanatory variables of interest for particular analyses resulted in additional variation in sample sizes.

Learning measures

We used three sets of measures as response variables to assess learning outcomes over the study period. The first two sets comprised questions about bird and tree identification; the third set involved more complex ecological concepts related to backyard conservation practices. Educational content on ecological concepts was featured explicitly within the YardMap citizen science project on the "Learn" pages and was accessible via a prominent tab on the project's Webpages. The "Learn" pages contained searchable information about habitat and residential ecology and were written for the public in a simple and straightforward manner. Participants could search for Bird-ID information on the Cornell Lab's Website (primarily in a Web property called "All About Birds"); Tree-ID information would have to be accessed elsewhere. The reason for testing learning in these different contexts was to differentiate the impacts of project-specific learning materials (ecological concepts), content that could be gleaned from the Lab's larger set of Web properties (Bird-ID), and content that participants would have to find in other ways (Tree-ID).

Additionally, Tree- and Bird-ID skills, as well as ecological concepts, could be a part of the shared intellectual capital of the community and might be learned from others via exposure to information and social norms present in the social version, but not available in the non-social version. By delivering two treatments, including one that had social tools (**Table 2**), we sought to determine whether this kind of social interaction was driving learning.

Analysis

Data were analyzed in Program R version 3.1.3. We performed analyses as follows: (1) we examined key characteristics of the data, including variation in participant effort (measured as the number of logins to the site) and the potential for biases in which treatment or experience with the test influenced participant effort; (2) we examined variation in pre-test scores and the response variable, post-pre difference in test scores; (3) we examined bias in participation and taking of the post-test; (4) we used General Linear Models (GLMs) and significance testing (*alpha* = 0.05) to test the *a priori* predictions in **Table 1**; (5) we further examined the relationship between pre-test score and post-pre difference in test scores using linear regression. For both types of statistical models (3 and 4 just above), the response variables were the post-pre difference in total learning score for each of the three types of learning. When participants did not answer some of the questions within a set (e.g., on birds, trees, or ecological concepts) or selected the answer, "I don't know," we coded their response as incorrect. Explanatory variables were standardized using the scale() command in R to allow for comparison of estimated effect sizes (in the GLMs) between different sets of content knowledge questions (tree-facts, bird-id, ecological concepts).

We first explored key characteristics of the data, including variation in levels of effort in the project and whether effort varied among treatments (**Table 2**). Testing for an effect of social/non-social treatment on effort in the project was necessary because, if found, such a difference would mean that the two treatments differed in two ways: Social exposure and effort, both of which might be positively associated with learning.

We also examined variation in the three measures of content learning, variation in participation in the online surveys, potential biases in survey response as a function of treatment at each stage of the experiment (pre-test and post-test), and accuracy of participants' self-estimates of their knowledge about birds and gardening.

Results

Description of participant effort

As in other online, user-generated content systems (Stewart et al. 2010), participant effort in YardMap approximated the Zipf curve, such that the majority of treatment group partic-

ipants (~75%) only had one login after completing the pre-test, and the rest showed wide variation in effort, producing a long tail (**Figure 3**). Because participants in the non-social treatment had fewer ways to engage with the project and were not exposed to the social parts of the project, we predicted that participants in the non-social treatment would be less active than those in the social treatment, meaning they would log in fewer times (**Table 1**). To test this prediction, we divided participants into two categories, people who visited just once and those who returned to the project after their initial visit. Our prediction was not supported: participants in the social treatment were not more likely to return to the project after a first visit than were participants in the non-social treatment (22% (nonsocial, n = 158) v. 24% (social, n = 168), Chi-square = 0.183, $df = 1$, $p = 0.67$).

Pre-test score variation

Table 3 presents descriptive statistics for learning scores. Bird-ID knowledge was variable and skewed slightly left (**Figure 4**). About half of the participants answered all of the Tree-ID questions correctly, skewing the distribution of scores to the right (**Figure 4a–c**). The mode for Tree-ID scores was 6, the highest possible score (**Table 3**). Only the total score for ecological concepts was distributed normally (**Figure 4c**). As expected, subtracting pre-test scores from post-test scores produced post-pre differences that were approximately normal (**Figure 4c–e**).

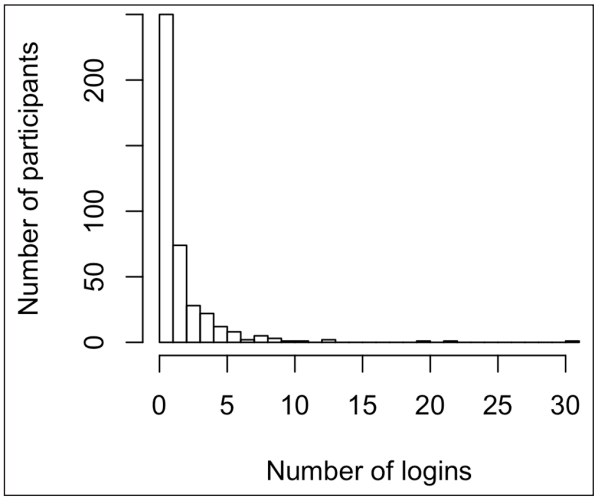

Figure 3: Number of logins by individuals over 8-week study period. Logins are distributed as a Zipf curve.

Table 3: Pre-test scores. Scores on pre-test for the three categories of tests using all participants who completed the pre-test regardless of whether they completed the post-test.

Measure	Bird-IDs	Tree-IDs	Ecological concepts
Number of questions	8	6	5
Mean ± SEM	3.7 ± 0.08	4.75 ± 0.06	3.07 ± 0.05
Median	4	6	3
Skew	0.1	–0.53	–0.1
Kurtosis (sharpness of peak)	–0.69	–0.95	–0.75
Sample size	591	586	580

Testing for sample bias

The first assumption inherent in controlled studies is random allocation to treatments, which predicts that pre-test scores will not differ among treatments at the start of the study. Among the 555 participants who completed the pre-test within 50 min, Tree-ID scores were higher on the pre-test for the combined treatments than for the waitlist control (**Table 4**). The pattern was the same for the approximately 447 such participants who took both the pre- and the post-test. We found no differences among treatments for the total score on the Bird-IDs or Ecological Concepts questions (**Table 4**).

We also asked whether there was a post-test bias in which a disproportionate share of people with lower scores on the pre-test dropped out of the study and failed to take the post-test. This finding would constrain the ability to detect learning because it would disproportionately remove the participants who had the most to learn. Participants who scored lower on the Bird-ID questions were less likely to take the post-test than were participants who scored higher (**Table 5**). We did not find such an effect for Tree-ID or Ecological concepts scores.

Post-test bias can also occur if treatment/control assignments influence the frequency with which participants who complete the pre-test subsequently elect to take the post-test. Our results showed a difference in drop-out rate in which participants in the two treatment groups, combined, were less likely to take the post-test than were participants in the waitlist control (**Table 5**).

Results of General Linear Models to test for treatment effects

We used general linear models to test predictions regarding post-pre differences in knowledge of Bird-IDs, Tree-IDs, and ecological concepts. We tested for an effect of participation in the project on learning outcomes by comparing the two treatment groups combined with the control; we then asked whether the social components of the mapping application enhanced learning, comparing the social with the non-social treatment (**Table 1**).

Treatment participants (those participating in either version of the mapping project) did not learn more than did those in the waitlist control (**Table 6**). This was the case for all three types of learning content. The number of logins (effort) was not a significant predictor of the post-pre learning difference. On the other hand, the pre-test score was significantly negatively associated with post-pre score difference for all comparisons (**Table 6**). The relationship between pre-test score and knowledge gain (post-pre difference) was always negative, and the estimated effect size ranged from –0.41 to –0.57. This means that the less participants knew on the pre-test, the larger their gain in knowledge in the interval between the pre- and post-test. These same findings were supported when we increased the potential to detect learning by removing

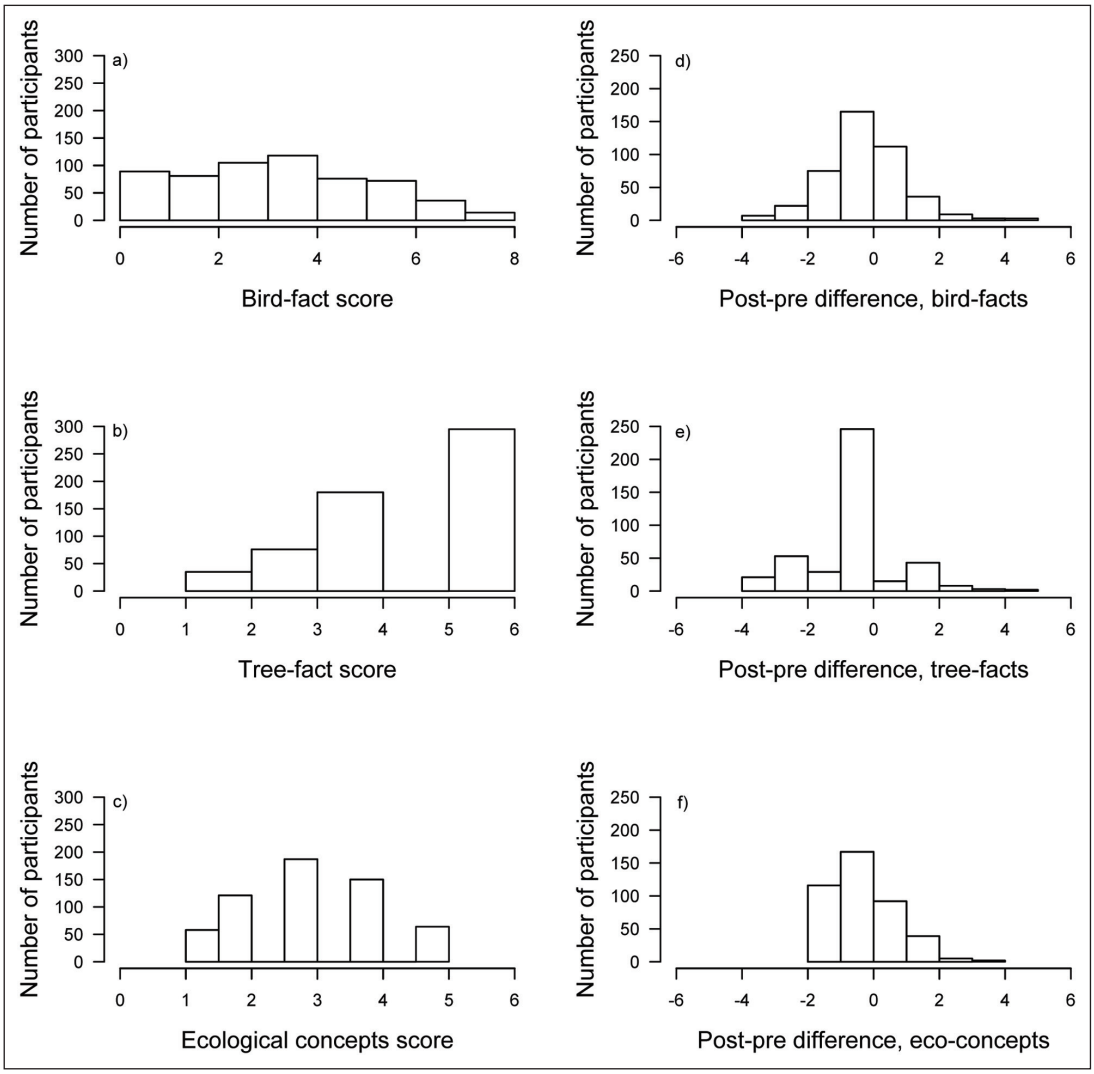

Figure 4: Distributions of pre-test scores **(a–c)** and post-pre differences in scores **(d–f)** for the three types of questions.

Table 4: Pre-test bias among ~560 participants completing the pre-test. Results of GLMs (for Bird-ID and ecological concepts) and non-parametric analyses for Tree-ID to determine whether pre-test scores were random with respect to treatment. Sample included all participants who completed the pre-test within 50 minutes, including those who did not take the post-test.

Scores on pre-tests	Explanatory variable	Effect size	Test statistic	*P*-value
Bird-IDs (GLM, Negative binomial, n = 591)	Control v. Two treatments combined	0.06 ± 0.04	$t = 1.29$	0.20
	1 control and 2 separate experimental treatments	0.04 ± 0.03	$t = 1.77$	0.08
Tree-IDs (Mann-Whitney U, Kruskal-Wallis, n = 555)	Control v. Two treatments combined	0.38 ± 0.17	$W = 37{,}886$	**0.01***
	1 control and 2 separate experimental treatments	–	Chi-square = 6.67	**0.036***
Ecological Concept questions, (GLM, Gaussian, n = 580)	Control v. Two treatments combined	–0.02 ± 0.09	$t = -0.19$	0.85
	1 control and 2 separate experimental treatments	0.03 ± 0.05	$t = 0.48$	0.63

all participants who had perfect scores on the pre-test. The r-square values for the general linear models were low, ranging from 0.10 to 0.27.

Based on the results, we conducted an *a posteriori* analysis of the relationship between pre-test score and post-pre score difference to see if the slopes of the lines differed between treatments and controls. Linear regression corroborated the GLM results in finding a significant negative relationship between pre-test score and post-pre differences for the two treatments and the control

Table 5: Post-test bias. Results of Generalized Linear Models to determine the effect of treatment and pre-test score on the tendency for participants to take (1) the post-test or not (0) (binomial response variable). $N = 560$ participants.

Explanatory variable	Effect size	z	*P*-value
Control (0) v. Two treatments (1)	–0.72 ± 0.19	–3.83	**≤0.001***
Three separate categories: Control, non-social, social (0, 1, 2)	–0.38 ± 0.10	–3.62	**≤0.001***
Bird-IDs pre-test score (0–8)	0.15 ± 0.05	3.15	**≤0.002***
Tree-IDs pre-test score (0–6)	–0.002 ± 0.068	–0.03	0.97
Ecological concepts pre-test score (0–5)	0.01 ± 0.08	0.15	0.88

Table 6: Results of General Linear Models to test predictions regarding learning as measured by post-pre differences. The response variable was post-pre difference in scores for Bird-ID, Tree-ID, and Ecological concepts (analyzed separately). The explanatory variables included experimental treatment (waitlist control was coded as zero; the non-social treatment was coded as 1; and the social treatment was coded as 2), number of logins as a measure of a participant's activity in the project, and pre-test score (birds, trees, or ecological concepts). The r-square for these analyses ranged from 0.10–0.27.

Explanatory variable	Bird-ID			Tree-ID			Ecological concepts		
	Estimated effect size ± SEM	*t*	*p*-value	Estimated effect size ± SEM	*t*	*p*-value	Estimated effect size ± SEM	*t*	*p*-value
Treatment (0 vs. 1 and 2 combined)	0.11 ± 0.14	0.77	0.44	–0.03 ± 0.09	–0.17	0.86	0.08 ± 0.07	0.49	0.49
N logins	–0.01 ± 0.07	–0.20	0.84	0.02 ± 0.03	0.82	0.41	0.08 ± 0.05	0.14	0.14
Pre-test score	–0.41 ± 0.06	–7.14	**<0.001*** **	–0.55 ± 0.06	–8.64	**<0.001*** **	–0.57 ± 0.05	–12.47	**<0.001*** **
Treatment (1 vs. 2)	0.17 ± 0.17	0.99	0.32	0.12 ± 0.18	0.67	0.50	–0.14 ± 0.13	–1.09	0.28
N logins	–0.01 ± 0.07	–0.10	0.92	0.02 ± 0.03	0.85	0.39	0.08 ± 0.05	1.44	0.15
Pre-test score	–0.41 ± 0.09	–4.55	**<0.001*** **	–0.54 ± 0.09	–5.97	**<0.001*** **	–0.53 ± 0.06	–8.79	**<0.001*** **

(**Figure 5a–c**). The slopes of the regression lines were similar among the two treatments and control for all three types of learning content.

Discussion

Evidence for learning

Our main result is that there was no detectable difference in learning among the treatments and the waitlist control group, nor was there any statistical difference in learning outcomes between the social and non-social versions of the YardMap application. However, if no learning occurred among the participants in the study, we would expect to find no relationship between pre-test score and post-pre differences in test scores (**Figure 6**); instead we observed a line with a negative slope (**Table 6**). The similarity of the slopes of the regression lines for the three treatments is consistent with the conclusion that learning occurred but did not differ among treatments nor between the treatments and the waitlist control (**Figure 5**). People learned as many Bird-ID's, Tree-ID's, and ecological concepts by merely intending to participate in YardMap (and/or by taking the pre-test) as they did by actually participating. This learning was not influenced by effort in the project (waitlist control participants had zero logins between the pre- and post-test), however, learning may have been influenced by things we could not measure, such as access to "Learn" pages in YardMap, access to other Cornell Lab of Ornithology Web properties, or access to other learning resources on the Web.

Alternative explanations for findings and implications for experimental Citizen Science research

Our findings, although they did not support our hypotheses about activity-based learning and social learning in YardMap (**Table 1**), provide new insights about the efficacy of using waitlist controls in online learning environments and enable us to specify how the field can move forward (**Table 7**). We cannot necessarily conclude from this study that participation or social interaction within YardMap does not foster learning; instead, it is still possible that learning occurred in the treatment and control groups for different reasons. Here we discuss alternative explanations for the finding of no difference.

Bias

We found three kinds of bias that may have influenced the results, and in each case, the bias would have worked against finding that treatment individuals learned more than controls. First, for one of our learning measures (Tree-ID), the score on the pre-test for individuals in the

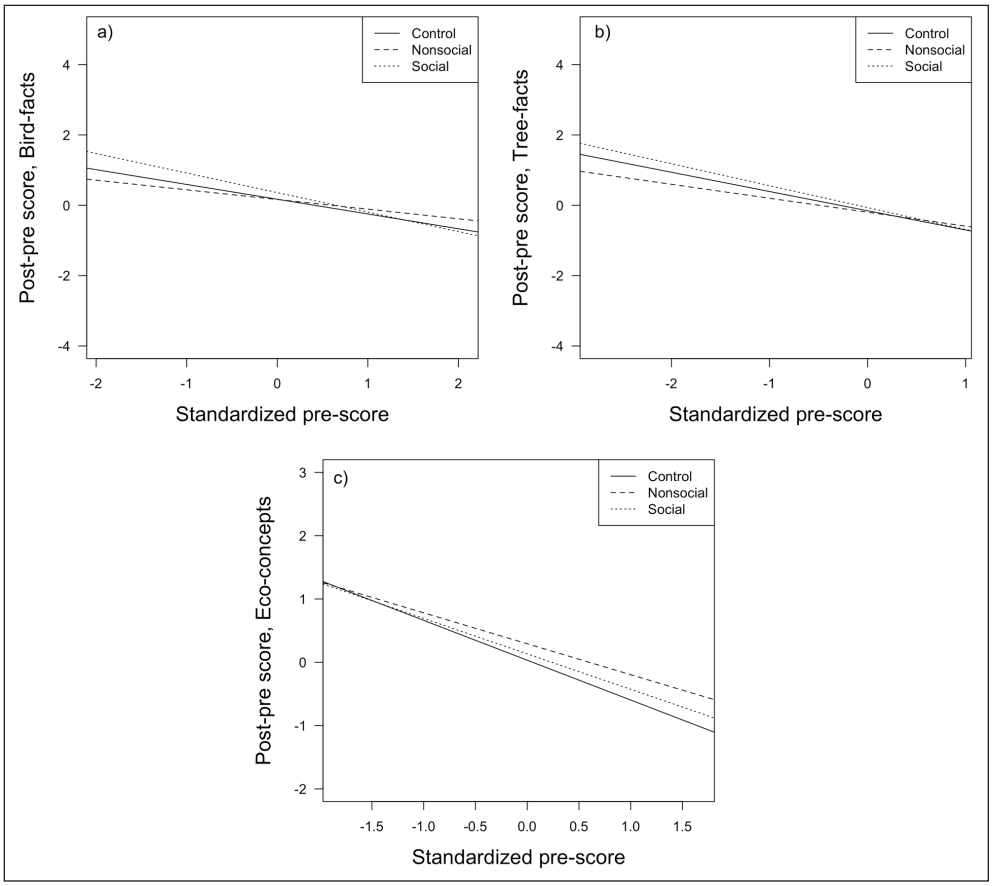

Figure 5: Linear regression lines for the relationship between pre-test score and post-pre-test difference. R^2 values ranged from 0.11 to 0.75 and were lowest for birds and highest for ecological concepts.

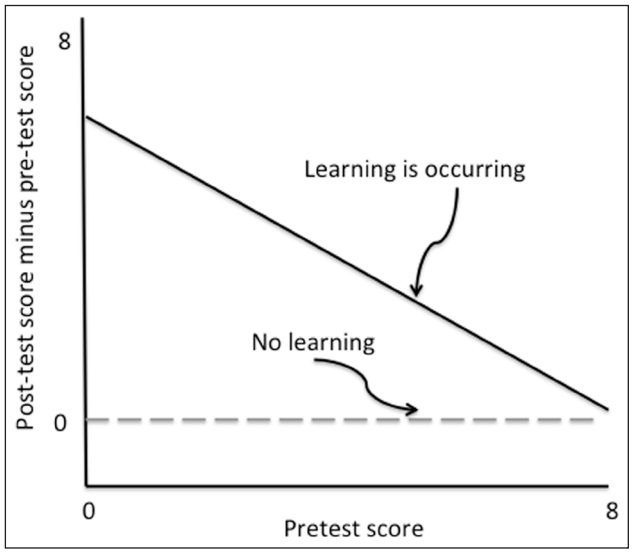

Figure 6: Expected relationship between pre-test score and learning difference (post minus pre-test) when learning occurs *versus* when it does not occur. If learning occurs, we predict that the learning difference is highest for participants with low pre-test scores, and declines to zero for participants with perfect pre-test scores. Participants with perfect pre-test scores cannot demonstrate learning based on the questions asked. In contrast, if there is no learning, we no relationship between pre-test scores in the study and the post-pre learning difference (a line with zero slope). A steeper negative slope will tend to show that more learning has occurred.

two treatment groups was higher than for the control. Consequently, the range of learning that the Tree-ID measure could detect was smaller in the experimental groups than in the waitlist control group. We did not find such a bias for the Bird-ID or ecological concepts questions. Second, when we looked at dropout rates, we found that participants who scored lower on the Bird-ID pre-test were more likely to drop out and not take the post-test than were high-scoring participants, constraining the ability to detect learning of Bird-IDs in all groups. Third, treatment participants were less likely to take the post-test than were the control participants, a perhaps unsurprising outcome given that participants in the waitlist control were unable to get into YardMap and participate until they took the post-test. This finding suggests that the use of waitlist controls in learning studies that involve self-selection into desired activities may have unwanted consequences: Making people wait to do an activity they are interested in likely increases their motivation to complete the post-test. It may also increase their desire to learn during the waiting period.

Low activity in the project

The distribution of activity in the project approximated a Zipf curve (**Figure 3**), such that only a small share of the participants in the study were active in the project. This activity distribution typifies online citizen science projects and suggests that we may need to generate much larger sample sizes and restrict analysis to highly active participants to better assess the effects of social interaction and

Table 7: Recommendations for how to proceed with controlled studies of online learning.

Potential problem	Recommendation
Increased learning in waitlist control	Design study with two waitlist controls and unseen questions: 1) Control for participation: Study participants take the pre-test with active participants and are blocked from the project until after they take the post-test. 2) Control for pre-test effect with second control or use Solomon Four Group Design: Study participants do not take a pre-test, wait to start the project with the other waitlist controls, and only take the post-test [If there is a pre-test effect, their post-test scores will be lower than those of the waitlist controls who took the pre-test]. 3) Use new questions to test conceptual knowledge and facts in the post-test to ameliorate the pre-test effect on learning.
Insufficient variation in pre-test scores	When using instruments that have not been validated, test the pre-test with 50 random participants to make sure there is enough variation in pre-test scores to detect an increase.
Differences among treatments in pre-test scores	Increase the sample size to allow segmentation of the pre-test data in a way that homogenizes pre-test scores among treatments.
Learning potential declines with pre-test score	Include pre-test score as an explanatory variable in analyses of learning.
Increase number of high-effort participants in sample	Increase sample size to provide a more robust sample of high effort participants, allowing segmentation of data to study effects of activity in the project on learning.

citizen science participation on learning. Restricting analysis, *a priori*, to comparisons of high-effort participants (e.g., those logging in 5 or more times) has important implications for future studies: Unless participant effort in projects can be increased, obtaining sufficiently large samples of high-effort participants is likely to require longer time periods to boost sample sizes, which may require allowing participants to enter the study for a year or more instead of the two months that we allowed. Given the costs of setting up online experiments in terms of programming time and project delivery effort, longer studies are likely to be more definitive and thus cost-effective.

Exposure to the pre-test may have increased learning in the waitlist control

Our results provide insight into how the pre-test itself might disproportionately increase learning in the waitlist control. In our study, the pre-test had specific questions about learning content, and these same questions were asked in the post-test eight weeks later. These questions may have canalized and motivated learning after participants completed the test. Such "test-driven" learning may have been more likely to occur among waitlist control participants than active participants. Waitlist control participants who finished the pre-test were immediately told that they had been selected to wait two months before participating in the project. They left the platform just after a test of their content knowledge, and, unlike active participants, were immediately free to search for answers to questions they had been unable to answer on the pre-test (e.g., What was that bird?). This difference between waitlist control and treatment participants could lead to a "pre-test effect" in which the waitlist control group learns as much or even more about the tested facts and concepts than does the group of active participants. Such a bias in the waitlist control would militate against detecting increased learning in the treatment groups even when it is occurring.

If participants are primed to learn specific content due to experiencing questions about that content in a pre-test, this suggests that a second kind of waitlist control is needed wherein participants are required to wait to do the project, but are not given the pre-test alongside treatment group participants (**Table 7**). Such control participants would fill out other survey questions (e.g., demographic questions), then wait and answer the learning questions for the first time when other participants are taking the post-test (**Table 7**). Scores for this second control – an untested waitlist control – could then be compared with post-test scores for active participants to test for learning and could also be compared with scores of the pre-tested waitlist control to look for pre-test effects on learning. A second alternative would be to use the Solomon four-group design, which would add both treatment and control groups that are not exposed to the pre-test (Solomon 1949). Both of these alternatives would require increased sample sizes. A third alternative would include in the post-test novel content questions that could not have been influenced by the pre-test (**Table 7**). The challenge of managing all of these treatment and control groups may explain why the field of computer science uses A/B testing instead, where they introduce a new feature and test the cohorts that entered the project a few weeks before and a few weeks after the new feature was launched.

Lack of power to detect learning

Tree-ID questions had high scores on the pre-test such that the maximum score, 6, was also the modal score; in the case of Tree-ID, high pre-test scores with low variabil-

ity made it difficult to find differences between pre- and post-test scores. Validating questions on a separate population of project participants prior to starting the experiment could help to ensure sufficient variation in pre-test scores (**Table 7**).

Pre-test scores are an important explanatory variable for studying learning

Our finding that the pre-test score was negatively associated with knowledge gain (the post-pre difference) for all three sets of learning questions (birds, trees, ecological concepts) has implications for post-pre test comparisons. Such comparisons, when they fail to include an individual's pre-test score as an explanatory variable, may fail to detect learning when learning is occurring, simply because they fail to take into consideration variation in the potential to achieve a score increase. Given that high pre-test scores constrained our ability to detect learning, it could be helpful to select questions that produce low starting scores on pre-tests, although, based on our evidence of bias, this can also have adverse effects on participants' willingness to stay in the study.

We suggest that the relationship between pre-test score and post-pre difference is itself a measure of learning because there would be no relationship if learning had not occurred (**Figure 6**). The most commonly used measure to compare learning across studies is the normalized learning gain (Hake 1998), which is problematic because it involves ignoring scores that go down between the pre- and post-test, replacing what would be a negative difference with zero (Miller et al. 2010). We suggest that focusing on study designs that use the slope of the relationship between raw pre-test scores and the post-pre difference could lead to a new means of comparing learning outcomes across projects.

Conclusions

This research provides new insights into the nuances of studying learning in online citizen science environments. Some of these insights will also apply to offline studies, especially controlled experiments and observational studies that involve quantitative analysis of post-pre differences in test scores based on surveys or exams (**Table 7**). It is important to note that content-learning is only one of several desired outcomes for citizen-science projects; the field of citizen science has articulated many other learning and behavioral outcomes of interest (McCallie et al. 2009; Phillips et al. 2018). Yet, content-learning remains one of the most commonly measured outcomes in research and evaluation of citizen science projects (Phillips et al. 2018). Because of its persistence as an outcome of interest to project leaders and stakeholders and because the impact that citizen science participation appears to have on content-learning is still poorly understood (in part due to the complexities described here), research methods need to be examined thoroughly. The efficacy of methodologies is nuanced, and the problem of experimental research design is worthy of investigation in its own right, especially with respect to sample sizes, biases, robust controls, and variability in measures of learning. In order to adopt standardized frameworks for measuring learning outcomes (Phillips et al. 2018), we must develop rigorous, well-tested methodologies and standardized methods for articulating results. We suggest that using the "negative slope to detect learning" as presented here could allow researchers to compare within and among projects to help build a corpus of evidence around learning outcomes (as well as behavioral and other outcomes). We also suggest that we should not avoid experiments simply because they are hard; improved experimental tests of learning are still the only robust way to measure learning in informal settings. Progress in this area will undoubtedly be of interest to the fields of citizen science and informal science learning, their funders, and their public stakeholders.

Supplementary File

The supplementary file for this article can be found as follows:

- **Appendix I.** Learning questions used on the pre- and post-surveys. DOI: https://doi.org/10.5334/cstp.218.s1

Ethics and Consent

This study was conducted under the guidance and approval of the Institutional Review Board for Human Participants (IRB) at Cornell University under protocol #0906000455.

Acknowledgements

This research was conducted in 2014 under IRB # 1005001414. We are grateful to the editors and reviewers for important comments that greatly improved the manuscript. We thank Chris Marx for application programming critical to the study, Kevin Ripka for Web design work, Steve Yalowitz and Tammy Cherry for helpful discussion and preparation of an evaluation report based on this study, Robyn Bailey, Suzanne Treygor, Becca Rodomsky-Bish, and Jacob Johnston for helping to answer participants' questions in YardMap (now Habitat Network). This research was supported by NSF grants #0917487 from the Division of Research on Learning (DRL-AISL) and #1441526 from the program for Computer and Information Science Education (CISE-Cyberlearning). We thank The Nature Conservancy and its donors for funding to support revisions to the project after this study.

Competing Interests

The authors have no competing interests to declare.

References

Bonney, R, Phillips, TB, Ballard, HL and **Enck, JW.** 2016. Can citizen science enhance public understanding of science? *Public Understanding of Science*, 25: 2–16. DOI: https://doi.org/10.1177/0963662515607406

Dilorio, C, Bamps, Y, Reisinger, E and **Escoffery, C.** 2011. Results of a research study evaluating WebEase, an online epilepsy self-management program. *Epilepsy & Behavior*, 22(3): 469–474. DOI: https://doi.org/10.1016/j.yebeh.2011.07.030

Fary, RE, Slater, H, Chua, J, Ranelli, S, Chan, M and **Briggs, AM.** 2015. Policy-into-practice for Rheumatoid Arthritis:

Randomized controlled trial and cohort study of e-learning targeting improved physiotherapy management. *Arthritis Care & Research*, 66(7): 913–922. DOI: https://doi.org/10.1002/acr.22535

Ferguson, R, McDonald, B, Rocque, M, Furstenberg, C, Horrigan, S, Ahles, T and **Saykin, A.** 2012. Development of CBT for chemotherapy-related cognitive change: results of a waitlist control. *Psychooncology*, 21: 176–186. DOI: https://doi.org/10.1002/pon.1878

Hake, RR. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1): 64–74. DOI: https://doi.org/10.1119/1.18809

Heiman, HL, Heiman, HL, Uchida, T, Adams, C, Butter, J, Cohen, E, Persell, SD, Pribaz, P, McGaghie, WC and **Martin, GJ.** 2012. E-learning and deliberate practice for oral case presentation skills: A randomized trial. *Medical Teachers*, 34(12): e820–e826. DOI: https://doi.org/10.3109/0142159X.2012.714879

Jennett, C, Kloetzer, L, Schneider, D, Iacovides, I, Cox, A, Gold, M, Fuchs, B, Eveleigh, A, Mathieu, K, Ajani, Z and **Talsi, Y.** 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication*, 15(3): A-05, 1–23. DOI: https://doi.org/10.22323/2.15030205

Krasny, ME and **Roth, W.** 2010. Environmental education for social-ecological system resilience: a perspective from activity theory. *Environmental Education Research.* DOI: https://doi.org/10.1080/13504622.2010.505431

Land-Zandstra, AM, Devilee, JLA, Snik, F, Buurmeijer, F and **van den Broek, JM.** 2016. Citizen science on a smartphone: Participants' motivations and learning. *Public Understanding of Science*, 25: 45–60. DOI: https://doi.org/10.1177/0963662515602406

Masters, KL, Oh, EY, Cox, J, Simmons, B, Lintott, C, Graham, G, Greenhill, A and **Holmes, K.** 2016. Science learning via participation in online citizen science. *Journal of Science Communication*, 15(3): A07, 1–33. DOI: https://doi.org/10.22323/2.15030207

McCallie, E, Bell, L, Lohwater, T, Falk, JH, Lehr, JL, Lewenstein, BV and **Wiehe, B.** 2009. Many experts, many audiences: Public engagement with science and informal science education. *CAISE Inquiry Group Report*. Washington, DC: Center for Advancement of Informal Science Education (CAISE).

Miller, K, Lasry, N, Reshef, O, Dowd, J, Araujo, I and **Mazur, E.** 2010. Losing it: The Influence of Losses on Individuals' Normalized Gains. *American Institute of Physics Conference Proceedings,* 229. Montreal, Canada: John Abbot College. DOI: https://doi.org/10.1063/1.3515208

Phillips, T, Bonney, R and **Shirk, J.** 2012. What is our impact? Toward a unified framework for evaluating outcomes of citizen science projects. In: Dickinson, J and Bonney, R (eds.), *Citizen Science: Public Participation in Environmental Research,* 82–95. Ithaca, NY: Cornell University Press. DOI: https://doi.org/10.7591/cornell/9780801449116.003.0006

Phillips, T, Ferguson, M, Minarchek, M, Porticella, N and **Bonney, R.** 2014. *User's Guide for Evaluating Learning Outcomes in Citizen Science*. Ithaca, NY: Cornell Lab of Ornithology.

Phillips, T, Porticella, N, Constas, M and **Bonney, R.** 2018. A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science: Theory and Practice*, 2: 1–19. DOI: https://doi.org/10.5334/cstp.126

Raddick, M, Bracey, G, Gay, P, Lintott, C, Murray, P, Schawinski, K, Szalay, A and **Vandenberg, J.** 2010. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, 9: 1515–1539. DOI: https://doi.org/10.3847/AER2009036

Roeser, R, Schonert-Reichl, K, Jha, A, Cullen, M, Wallace, L, Wilensky, R, Oberle, E, Thomson, K, Taylor, C and **Harrison, J.** 2013. Mindfulness training and reductions in teacher stress and burnout: Results from two randomized, waitlist-control field trials. *Journal of Educational Psychology*, 105: 787–804. DOI: https://doi.org/10.1037/a0032093

Rotman, D, Hammock, J, Preece, J, Hansen, D, Boston, C, Bowser, A and **He, Y.** 2014. Motivations Affecting Initial and Long-Term Participation in Citizen Science Projects in Three Countries. *Proceedings: Breaking down walls: culture – context – computing 2014; iConference*, 110–124.

Sartory, G, Zorn, C, Groetzinger, G and **Windgassen, K.** 2005. Computerized cognitive remediation improves in verbal learning and processing speed in schizophrenia. *Schizophrenia Research*, 75(2–3): 219–223. DOI: https://doi.org/10.1016/j.schres.2004.10.004

Solomon, R. 1949. An extension of control-group design. *Psychological Bulletin*, 46(2): 137–150. DOI: https://doi.org/10.1037/h0062958

Stewart, O, Lubensky, D and **Huerta, J.** Crowdsourcing participation inequality: a SCOUT model for the enterprise domain. *HCOMP 10*, July 25, 2010; Washington, DC, 30–33. DOI: https://doi.org/10.1145/1837885.1837895

Wells, N and **Lekies, K.** 2012. Children and nature: Following the trail to environmental attitudes and behavior. In: Dickinson, JL and Bonney, R (eds.), *Citizen Science: Public Participation in Environmental Research,* 201–213. Ithaca, NY: Cornell University Press. DOI: https://doi.org/10.7591/9780801463952-021

Wells, NM, Myers, BM, Todd, LE, Barale, K, Gaolach, B, Ferenz, G, Aitken, M, Henderson, CR, Jr, Tse, C, Pattison, KO, Taylor, C, Connerly, L, Carson, JB, Gensemer, AZ, Franz, NK and **Falk, E.** 2015. The effects of school gardens on children's science knowledge: A randomized controlled trial of low-income elementary schools. *International Journal of Science Education*, 37(17): 2858–2878. DOI: https://doi.org/10.1080/09500693.2015.1112048

How to cite this article: Dickinson, JL and Crain, R. 2019. An Experimental Study of Learning in an Online Citizen Science Project: Insights into Study Design and Waitlist Controls. *Citizen Science: Theory and Practice*, 4(1): 26, pp. 1–13. DOI: https://doi.org/10.5334/cstp.218

Submitted: 09 November 2018 **Accepted:** 02 August 2019 **Published:** 24 October 2019

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

OPEN ACCESS