CITIZEN SCIENCE:
THEORY AND PRACTICE

Straub, M C P 2016 Giving Citizen Scientists a Chance: A Study of Volunteer-led Scientific Discovery. *Citizen Science: Theory and Practice*, 1(1): 5, pp. 1–10, DOI: http://dx.doi.org/10.5334/cstp.40

RESEARCH PAPER

Giving Citizen Scientists a Chance: A Study of Volunteer-led Scientific Discovery

Miranda C. P. Straub*

The discovery of a class of galaxies called Green Peas provides an example of scientific work done by volunteers. This unique situation arose out of a science crowdsourcing website called Galaxy Zoo. It gave the ability to investigate the research process used by the volunteers. The volunteers' process was analyzed in terms of three models of scientific research and an iterative work model to show the path to this discovery. As has been illustrated in these models of science, the path was iterative, not predetermined, and driven by empirical data. This paper gives a narrative of the 11-month, volunteer-led discovery process of the Green Pea galaxies and the transition to the Galaxy Zoo science team's involvement to analyze and report on a new class of galaxy. This study identified the cycles of the volunteers' work and situated them in the proposed integrated model of the scientific research process to show that the volunteers engaged in meaningful scientific research.

Keywords: citizen science; citizen science discovery; science research process; Zooniverse

Introduction

The main purpose of most citizen science projects is for volunteers to make a contribution to a scientific endeavor. From the Christmas Bird Count of 1900 to the suite of projects offered under the umbrella of the Zooniverse project, the goal is to advance scientific understanding. The primary tasks of the various projects depend on the data or need of the project and usually must be simple enough for volunteers to accomplish without extensive training or background knowledge. This method of data collection or data processing has contributed to numerous scientific publications in a wide variety of fields in the Zooniverse project alone (Zooniverse Team 2016). More extensive examinations of a wide variety of citizen science projects have proposed typologies of projects and discussed the possibilities for the future of the field of citizen science (Bonney et al. 2015; Wiggins and Crowston 2010). These reviews have made it clear that more research needs to be done on how people learn through citizen science and about ways of engaging people who would not otherwise participate.

The success of many citizen science projects has led to initial research on the volunteers and their motivations for participation. A study of some Zooniverse community members showed that motivations included helping, contribution to science, community with other citizen scientists,[1] and simply having fun (Raddick et al. 2010). Studies of other projects have explored whether citizen scientists learn about science through their participation. These studies suggest that, not only do participants learn about science, they also step beyond the primary task of the project and have the ability to contribute at a deeper level (e.g., Cohn 2008; Trumbull et al. 2000). These studies have noted that some volunteers are both willing and able to offer ideas regarding their experimental design and decisions they made about data collection. Even though these particular projects–or citizen science projects in general—typically do not expect this kind of contribution from their volunteers, there are cases where participants engage in significant work that merit deeper study. This raises some questions: To what extent are participants contributing to science? Is their work sufficient to constitute research? This paper proposes a model of the scientific research process that can be used to analyze the work of citizen scientists. It then employs this model to analyze an instance of discovery of the Green Pea galaxies to establish the work as authentic scientific research.

Models of science

To analyze the work of citizen science participants, a model or definition of scientific research must be chosen. Because there is no single, widely accepted model for how science is done, I have chosen to use three proposed models of scientific work and reasoning to build an integrated model of the scientific research process. First, the National Academy of Sciences (2012) created a model for the activities of science as part of their report "Frameworks for K–12 Science Education: Practices, Cross-Cutting Concepts, and Core Ideas." In this report, the authors explained that they wanted a model that included the broader practices of science rather than just a focus on experimental proce-

* Curriculum and Instruction Department, University of Minnesota, Minneapolis, Minnesota
mirandacpstraub@gmail.com

dures, which limits the scope of scientific research. The model also emphasizes the iterative and social aspects of scientific work. It presents the work of science in three "spheres of activities," which are Investigating, Evaluating, and Developing Explanations and Solutions. Each of the spheres contains three to six activities that scientists or engineers use in their work. The report stresses that scientific work does not happen in any predetermined path and that iteration is an integral part of the process. This iterative component is represented on a diagram as two-way arrows between each of the three spheres.

The second model, from UC Berkeley's Museum of Paleontology (2015), provides an extensive explanation of science and how it is done. The model is an interactive diagram that has four main components of science: Testing Ideas, Benefits and Outcomes, Exploration and Discovery, and Community Analysis and Feedback. Within each of these components, the authors categorize activities and micro-processes that are part of the overall process of science. The model also emphasizes the iterative, not predetermined path of science work, both via arrows that connect every component of the model and in the support material explaining the model.

The third model is by Ronald Giere (1991), who developed a reasoning process for use in teaching scientific reasoning. This model is based on historic developments in science and is useful for assessing current scientific claims and results. The diagram of the reasoning model is presented in steps; however, two-way arrows point between several steps depicting iteration.

Because this study focused on the process of scientific research, only the parts of the models that pertain to activities of the individuals conducting the research were included in the integrated model of scientific research (**Table 1**). For example, an element from the NAS and Giere models that was included was "experiment," which corresponds to an activity or action of a scientific researcher. An element that was not included was "real world," which corresponds to an element of the broader context of science rather than an activity.

Each model provides activities of science and depicts a non-linear, iterative process. I have organized the individual activities from the three models into four categories: Questions and Hypotheses, Data, Analysis, and Communication. The Questions and Hypotheses category encompasses activities that define the questions being asked and which set criteria for confirming or refining proposed scientific models. The Data category contains the activities of gathering empirical evidence, which includes experimentation, data collection, and measurements. The Analysis category includes activities that analyze the connections among data and look for relationships; this is where data are assessed to determine whether they will support a scientific model. The Communication category includes peer review, arguing, critique, and publication; these activities can occur within research groups, in response to journal articles, at conferences or other presentations, or in the form of competing work. While these four categories organize and associate the activities of science research from the three models of science, this presentation has no way of including the iterative nature of scientific work or a framework for the incremental progress that occurs.

Model of iteration

To study the iterative progress of research I used the IMOI (Input, Mediator, Output, Input) model (Ilgen et al. 2005), which provides a structure to analyze iterative work by identifying micro-cycles within a global project. The model shows that the work is accomplished in small steps and moves to a new state after each cycle. This iterative refinement over time is identified in four stages: Input, mediator, output, and input. The input depends on the work of the research group and could include a project goal, information, or a previous revision. The group takes the input and acts as a mediator for the information. The group must decide what to do with the input, how it may change their understanding, and what it means in reaching their goal. The output of the mediation could be an answer to a question, a new procedure, or a new understanding of the task. The group takes the output or other new information as a new input for the next cycle. This work continues until the group has reached its goal.

Research context and discovery narrative

Galaxy Zoo and GPAC thread overview

Galaxy Zoo 1 began in July 2007 with an appeal to the public to help classify images of galaxies according to morphology. The goal of the original project was to have volunteers

Questions and Hypotheses	Data	Analysis	Communication
Ask questions [1, 2]	**Data collection [1, 2]**	**Analyze [1]**	**Share data and ideas [2]**
Explore the literature [2]	Experiment [1, 3]	Interpret data [2]	Feedback [2]
Predict [1, 3]	Observation [1, 2, 3]	Calculation [1, 3]	**Critique [1]**
Formulate Hypotheses [1]	Measurement [1]	Test solutions [1, 2]	Argue [1]
Propose solutions [1]		Reason [1, 3]	Peer review [2]
Negative/positive evidence [3]			Replication [2] **Publication [2]**
Accept/reject model [3]			
Revision [2]			
Imagine [1]			
Build theory [2]			

Table 1: The activities of the scientific research process. The number by each activity indicates the model from which it came (NAS 2012 [1]; Berkeley 2015 [2]; Giere 1991 [3]). The bold text in the table indicates the activities used in the Green Peas discovery.

visually inspect approximately one million galaxies from the publicly available Sloan Digital Sky Survey (SDSS) data and classify each galaxy according to predetermined morphology choices. In the original project, the volunteer was shown a picture of a galaxy and given the choices of spiral, elliptical, or other. Once they selected their choice and submitted it, another randomly assigned galaxy appeared. The volunteer could use the galaxy ID to go to the SDSS page to get more information about the object or search the database for other similar objects. Each galaxy was viewed an average of 38 times (Lintott et al. 2008). The volunteers were able to post questions, comments, and images to a forum that was monitored by project scientists.

The public response was unexpected and overwhelming, with 100,000 people joining and contributing 25 million classifications during the first 50 days (Raddick et al. 2010). The initial project finished in half the expected time and established a protocol for crowdsourcing data analysis and online citizen science (Fortson et al. 2012; Lintott et al. 2008). Galaxy Zoo was not originally intended to be outreach-focused or educational; however, it was quickly apparent that support for the volunteers was needed. The astronomers found themselves swamped with questions from the army of participants, so they created an online forum to help with disseminating answers. The forum took on a life of its own as the volunteers talked with each other, answered questions, and were able to create opportunity for both community and deeper scientific contribution.

The focus of this paper is the volunteer-led research process that took place in one thread from Galaxy Zoo in 2007–08. The thread, entitled "Give Peas a Chance" (GPAC), was started very early and chronicles the story of how a new class of galaxy was established through volunteer work. The Green Peas discovery illustrates how willing volunteers can engage in a research process and produce a scientifically valuable output. The thread is still available for contribution; however, this study focuses on the first 11 months, from August 2007 to July 2008, which incorporates the first 1,193 posts. At that time the lead scientist on the study, Carolin Cardamone, started a separate thread titled "Peas Project," which is briefly summarized. The Galaxy Zoo forum and the GPAC thread were open to anyone who had an account with the project. Posts were displayed chronologically, with 15 posts per page. The users had the ability to include internal links back to previous posts or to posts on other threads within the forum or external links to relevant information. Many factors are interesting and important in this discovery process, including community functions, scientific discourse, and establishing credibility and authority; however, this paper focuses on the activities of the scientific research process that the volunteers exhibited.

Participants

In the 11-month volunteer-organized portion of GPAC, 105 volunteers contributed to the project. Given the nature of the project, the only characterization of the volunteers that can be made is their contribution to the forum. The barrier to entry is kept very low to attract sufficient volunteers for crowdsourcing to be effective. The project requires only a username and an e-mail address to sign in and begin contributing. Occasionally volunteers shared information about what they do, where they are from, or special skills they have; however, for the purposes of this study, the careers or educational backgrounds of the volunteers are not considered. The analysis shows the distribution of the work done by these volunteers. The participants are identified by their usernames on the forum, which is how they will be identified here.

Narrative construction

To understand the unfolding narrative over the initial 1,193 posts of this thread, I used qualitative data analysis methods in line with the practices of grounded theory to look for the interactions and events that highlighted the evolution of the work (Corbin and Strauss 2008). Grounded theory is a systematic approach to qualitative data that seeks to establish patterns or connections in the data before applying a theory to or building a theory from the data. This approach is particularly applicable to areas of study that are new and which do not have established frameworks to use for analysis. To construct the narrative presented in this paper, I first read the GPAC thread, taking note of the frequent posters, content of posts, and interactions of the participants. I kept a journal of interesting posts and wrote observer comments as I read the thread to make note of impressions, potentially significant contributions, and events that seemed to move the group to a better understanding of its work. These posts were documented with both post number and dates to review the posts and surrounding comments. I then highlighted 10 major events in the thread that provided an overview of the cumulative work of the volunteers.

Volunteer discoveries

The first two major discoveries of Galaxy Zoo occurred only weeks after the website launched and were within days of each other. Both serendipitous discoveries were made by the same user. Galaxy Zoo volunteer Hanny commented on both the Voorwerp (Dutch word for "thing") and the Green Pea Galaxy (**Figure 1**) on the forum and piqued the interest of other users and astronomers alike (Cardamone et al. 2009; Jozsa et al. 2009). Amid the flood of processed data, user comments, and questions, the Voorwerp almost immediately caught the attention of the astronomers working on the project. The Green Peas, however, remained a curiosity of the volunteers for almost a year before they garnered the astronomers' interest. By the time the members of the science team were able to study the Peas more closely, the group of volunteers had made several collections of individually inspected candidates for Green Pea galaxies. The science team members who studied the Peas, led by Dr. Cardamone, were able to use the volunteer-generated collection of candidates to set the criteria for the population of galaxies. The scientists found that the galaxies have a relatively high rate of star formation for their mass (Cardamone et al. 2009).

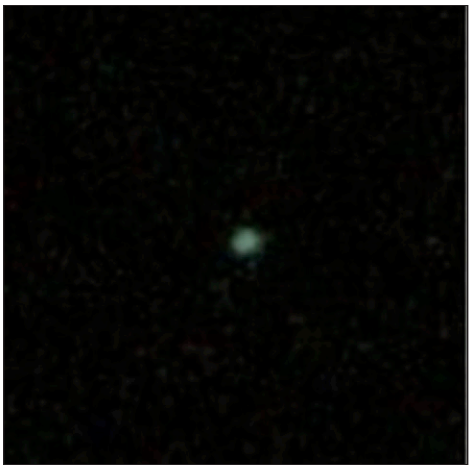

Figure 1: Green Pea: The first Green Pea galaxy posted on the GPAC forum by Hanny on August 12, 2007 (SDSS 2015a).

The Green Peas discovery

The initial poster on the GPAC of the thread, Hanny, gave a short description of the genesis of the thread by saying,

> "I found the first pea and posted it as a joke on the forum, with the topic title: 'Give peas a chance' (obviously after Lennon's song 'Give peace a chance'). Then others started posting them too and we talked (joked) about me making soup of them for a while. Later on people started collecting other fruits (i.e., not the green ones) and I think you know the rest. Hope that helps."[2]
>
> Peas Project thread, Nov. 05, 2008

Hanny's comment illustrates that, while the discovery of Green Pea galaxies was astronomically significant, it began as a light-hearted joke. The seriousness of the work evolved gradually as more objects were posted and the population of galaxies emerged. Hanny posted the first Peas on the thread and was one of the main moderators of the thread for the 11-month discussion. Other volunteers joined in the joke, contributed their Pea findings, and participated in the process of classifying the characteristics of a Pea. This collaboration was fully self-organized and self-regulated, and the volunteers relied on each other's various expertise and work to find the objects and decide what made the cut as an "official" Pea and what was just a green, fuzzy object. **Figure 2** is a graph of the cumulative posts for the GPAC thread over the 11-month, volunteer-led discovery process along with numbers noting the major events throughout the process. A short description of the events, dates, and contributing users is found in **Table 2**. I refer back to these events throughout the description of the volunteers' work.

Early work and explanations

At first, the criteria for a "Pea" were completely subjective and the original joke was maintained as the volunteers posted mostly green, compact objects. The thread grew in popularity as other users contributed to Hanny's collec-

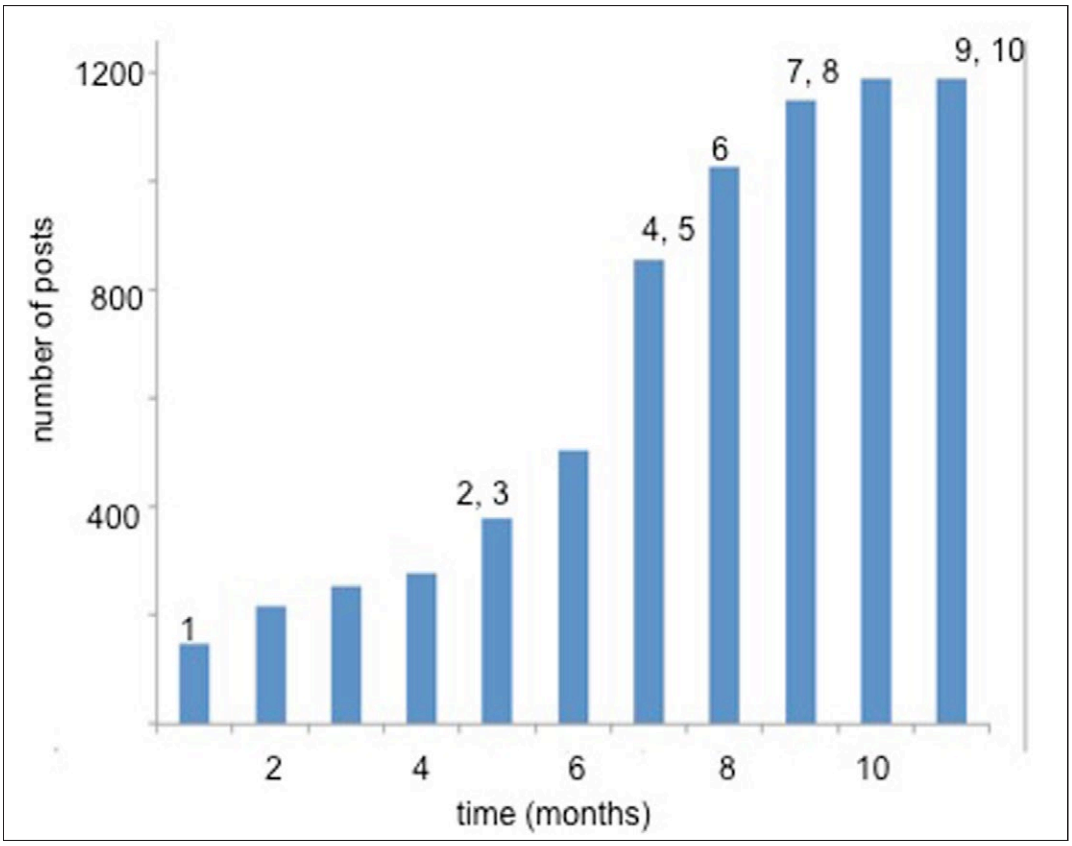

Figure 2: Timeline of GPAC thread: The numbers on the graph indicate the events identified in **Table 2**.

Event #	Date	Summary	Contributing user
1	08/12/07	Start of the “Give peas a chance!” thread	Hanny
2	12/14/07	First comment on a Pea by science team member	zookeeperkevin (science team)
3	12/24/07	First spectrum posted with object image	Rick Nowell
4	02/02/08	First Pea confirmed by established criteria	starry nite
5	02/03/08	First list of formalized objective and subjective criteria	starry nite
6	03/14/08	Instructions for querying SDSS	FermatsBrother
7	04/02/08	Summary of Green Pea galaxies	Rick Nowell
8	04/09/08	Comprehensive list of Pea candidates	laihro
9	07/08/08	Peas Project thread for research started	ccardamone (science team)
10	07/13/08	First “Pea Picker” program written	waveney

Table 2: Ten events from the GPAC thread.

tion and directed users from other threads to post their findings. In the first four months, the participants made over 300 comments and posted images that they came across either in their personal classifications or found posted in other discussion threads. Even though the conversation began as mostly jokes about peas, it evolved to a point at which volunteers offered more scientific comments or speculation about what these objects were.

At the end of the fourth month of the thread, one of the science team members, Kevin Schawinski (zookeeperkevin), responded to another forum post entitled “What’s this green thingy?” with a preliminary explanation about the object. He said the Peas were a type of emission line galaxy (ELG) with powerful doubly ionized oxygen (OIII) emissions, which contributed to the apparent green color. This explanation was linked back to the GPAC thread, as well as a similar post in the “Objects of the Day” thread, giving the Peas wider attention. Schawinski included an image of one of the Peas and the spectrum from the Sloan Digital Sky Survey (SDSS) page (**Figure 3**) to illustrate the OIII emission line (**Table 2**, Event 2). The abundance of doubly ionized oxygen seen in the spectra was the initial identifier of the Green Peas (see **Figure 3**). This post led user Rick Nowell to conduct his own literature search on OIII galaxies and report back to the thread with an article he found that mentioned oxygen-rich galaxies and information regarding citations in the NASA/IPAC Extragalactic Database (NED) catalogue. He also quoted a Wikipedia page regarding doubly ionized oxygen and other useful information resource pages. Rick Nowell reviewed the spectra of the objects posted in the thread up to that point and began a collection of the objects with the characteristic OIII abundance (**Table 2**, Event 3). He was the first to post a spectrum on the GPAC thread, which marks a change in the posting norm of objects. Once several of the more prolific posters also started putting the images, SDSS links, and spectra in their posts, this became the norm for presenting the potential Pea candidates. Since “real” Peas needed to have both the pea-like appearance and the characteristic emission spectra (**Figure 3**), this allowed for easier inspection of the posted objects.

Confirmation criteria

As volunteers posted their finds on the thread, other users evaluated the objects and usually discounted them as a candidate citing either the lack of a pea-like appearance or the incorrect spectra. In the beginning of January 2008 (Month 6), starry nite, who became the most frequent poster, asked, “Could someone ‘in the know’ post the criteria we should be looking for when searching for peas?” Rick Nowell responded with the initial criteria being “a lot of ‘doubly ionized oxygen’” and a green appearance. After engaging in a separate search and finding several objects that had borderline characteristics, starry nite found a true pea and posted,

> “A pea! A pea! Green, galaxy ID, correct spectral chart, and not posted before!
> $^{\text{PEA!}}$ *PEA!* $_{\textbf{PEA!}}$ *PEA!* $^{\text{PEA!}}$”
> starry nite, Feb. 02, 2008 (**Table 2**, Event 4)

starry nite was able to use the criteria of appearance and emission line abundance to positively identify a Green Pea Galaxy without needing confirmation from anyone else. The objectivity of this identification shows the initial establishment of the Pea criteria; however, the criteria continued to develop over the course of the thread. starry nite performed a significant search for pea-like objects that had previously been posted in other threads of the forum. This search for Peas led starry nite to postulate stricter, more objective criteria for the characteristics of a Green Pea Galaxy:

> “Characteristics of a ‘Pea’ galaxy A.K.A. OIII galaxy:
>
> 1. Mostly-flat spectral chart except for:
> 2. An extreme peak at OIII (double-ionized oxygen emission line),
> 3. If there is a peak at OII, it must be shorter than the OIII peak,
> 4. Other peaks must all be smaller than the OIII.
> 5. Any H-peaks should be narrow, not wide-based which might indicate a quasar (with a redshift z < 0.3).
> 6. Redshift range (z) of approximately 0.14–0.35 for a green color on SDSS.”
>
> starry nite, Feb. 03, 2008 (**Table 2**, Event 5)

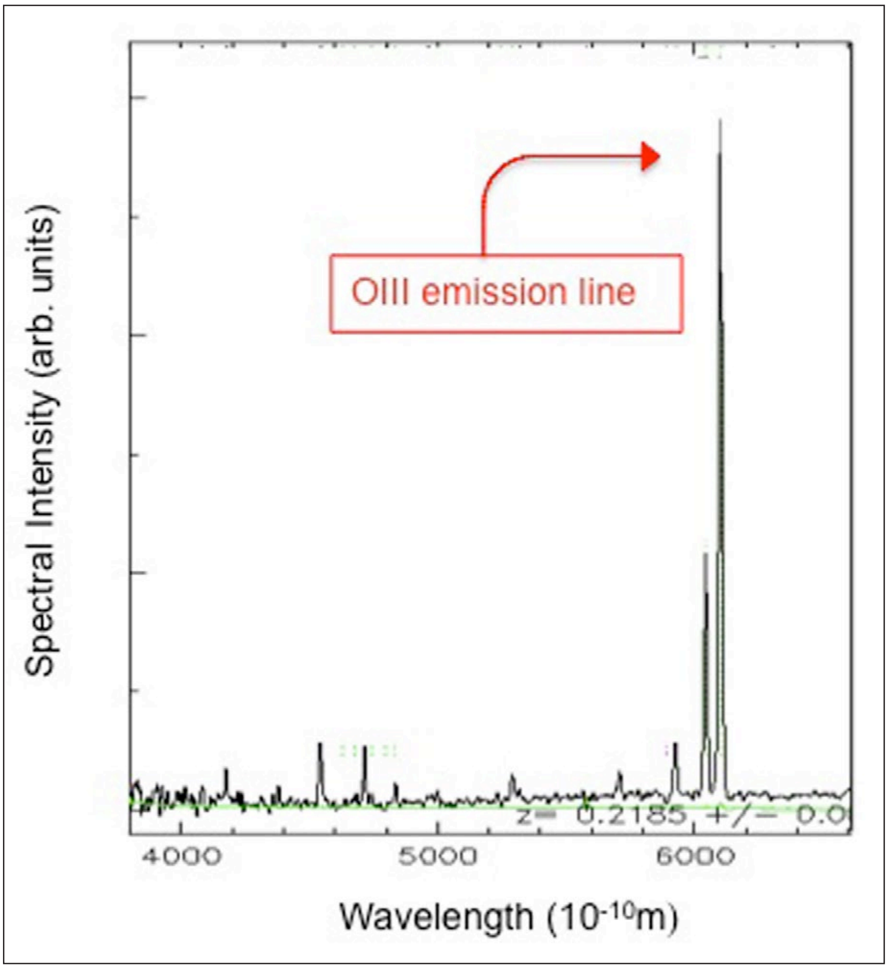

Figure 3: OIII emission spectrum: Spectrum posted by zookeeperkevin in the "Object of the Day" thread on Dec. 23, 2007 (Month 5) (SDSS 2015b). Original image was edited for emphasis of the OIII emission line.

These criteria allowed the confirmation of a "true" OIII object to be more quantitative and objective. starry nite maintained a list of Peas that fit the above criteria. By the end of the 11-month search, it contained 248 Pea candidates along with the redshift of each object and its apparent color.

Pea hunting strategies

The Pea collecting continued through Months 6–8 in four distinct ways: Searching while doing personal classifications; searching previously posted images; evaluating galaxies in published work; and running SQL queries on SDSS. The first three strategies are considered manual while the fourth is automated.

Manual strategies

The first strategy, utilized by most of the users, was looking for Peas as they classified galaxies in the primary task for Galaxy Zoo. Given the rarity of the Peas, users came across them infrequently and discovery was a cause for celebration, as seen in starry nite's quotation (**Table 2**, Event 4). The second strategy was searching the other threads in the Galaxy Zoo forums for previously posted Pea candidates. The Pea galaxies caught other people's attention for their shape and/or color, and they posted them as interesting objects in various thread conversations. The two participants to use this strategy prominently were starry nite and ElisabethB, who became the two most frequent posters. When a user brought a Pea-like object from another thread to GPAC, it was common for the post to acknowledge the first poster with an internal Galaxy Zoo link back to the original post or include their username. Sometimes more prominent users, like Hanny, would send volunteers to the thread with their object. This accounts for most of the users who posted only once on the GPAC thread (48 participants). The third manual strategy was reviewing either papers or catalogues in NED that contained confirmed Peas. Rick Nowell posted a paper on Wolf–Rayet galaxies that included several confirmed Peas, so another user, Galaxy Hunters Inc, reviewed the approximately 170 galaxies to check if there were Pea candidates in the paper.

Automated strategy

Finding Peas was a slow process for one person and many of the almost 600 objects posted did not fit the Pea criteria. This changed in Month 8 when user FermatsBrother posted,

> "Hi Starry – How's this 1/2 dozen for you?"
> Mar. 13, 2008

along with 6 images of "perfect" Peas. Galaxy Hunters Inc and starry nite responded with shock and amazement that FermatsBrother was so "lucky" to have "cultivated" so many Peas on his own. starry nite urged, “share with the rest of us. . . . How are you searching for these?” In March (Month 8), FermatsBrother posted the SQL query for searching SDSS (**Table 2**, Event 6). The query was intentionally limited to small ranges in the redshift in order to limit the number of galaxies returned and the runtime of the query. starry nite was the only user at the time to explicitly modify the query and use it. Another volunteer, laihro, wrote an inclusive query of SDSS and produced a large list of objects that fit certain Pea criteria (**Table 2**, Event 8); however, the volunteers’ queries were not refined enough to pull all Green Pea galaxies from the database without needing to confirm them by visual inspection.

The summary

At the beginning of April (Month 9), Rick Nowell gave a comprehensive summary of the Peas at a newcomer’s request (**Table 2**, Event 7). The summary included an overview of the discovery starting with Hanny’s first post and highlighted the contributions of several of the volunteers. It included an explanation of doubly ionized oxygen, where it typically is found, and what happens when there is an interaction with another particle. It gave a description of the characteristic spectrum of the Green Pea and some comments on the appearance. It noted that the color selected was originally green, but that more colors had been found and collected. It mentioned that the group had figured out that the Green Peas are “really red according to Fermat’s Brother’s extensive studies.” It noted that an explanation of the galaxies was “not a question anybody has yet answered satisfactorily” and that the galaxies “really are pretty rare.” It mentioned the collections that had been made and that “there was talk of getting some research together, but that seems to have dwindled away. . . .”

In the last two months of the volunteer-led portion of this research process, the contributions essentially stopped. Many of the main threads in the Galaxy Zoo forum and astronomical catalogues had been searched, and the participants had exhausted their knowledge about emission lines, filters, and SQL searches. On July 08, 2008, ccardamone (science team member) posted,

> “I’ve started a new thread called ‘Peas Project’ to collect a sample of Peas for further investigation. Hopefully we can gather enough of these interesting new objects to find out what they are and why they’re so special.”
>
> Jul 8, 2008 (**Table 2**, Event 9)

In the Peas Project thread, Dr. Cardamone introduced herself to the group and requested help for collecting the candidates that were green, compact, had a “bright” OIII emission line and had a redshift in the range of 0.15–0.45. The volunteers were eager to help collect the Peas they had been studying for the last year. Because one of the lists contained thousands of objects that had not been individually inspected, one of the users, waveney, wrote a computer program allowing the volunteers to sort all of the objects by color to ensure that they had only the green galaxies (**Table 2**, Event 10). The research work shifted from the volunteers to Cardamone, who was the primary science team member conducting the data analysis of the population. She continued to update the group on what she was doing with regard to the various statistical techniques and her initial findings. Even though the volunteers were unable to contribute to this work, they continued to ask questions and follow the process until the paper was written and published in 2009. A small group of the volunteers were thanked by name in the paper, and the names of all volunteers who wished to be mentioned as participants in Galaxy Zoo were provided on Galaxy Zoo publications via an HTML link on the authorship page (Cardamone et al. 2009). (Carolin Cardamone was a graduate student at the time of her involvement in this project; she has since earned her PhD.)

Analysis

In order to analyze the GPAC thread, I used the integrated model of research and the Input, Mediator, Output, Input (IMOI) model (Ilgen et al. 2005). The combination of these models provides both the activities and the iteration of scientific work, which can be used to connect the work and interaction of the volunteers with the scientific research process. I identified four distinct IMOI cycles in the selection of the thread that exhibited activities of the scientific research process. Cycles were chosen only if there were clearly identifiable input, output, and clear mediation by users. For each of the four cycles analyzed below, I identified the input, mediation, and output and situated it within the context of the scientific research process (see **Table 3**).

Cycle 1

The first major **input** of the discovery process was the first image of a Green Pea (**Figure 1**). The intent was not serious, and neither was the first cycle of **mediation** as the social group surrounding Hanny responded by also posting their own version of Peas. The **outputs** of the first four months of joking yielded about 300 images of assets dubbed “peas” of some sort. This cycle depicted the ask questions and data collection activities of the research process.

Cycle 2

The second cycle’s **input** was Schawinski’s post of a spectrum illustrating the abundance of doubly ionized oxygen. Rick Nowell **mediated** this input by reviewing the previously posted images, and he found 67 assets had the characteristic spectra. The **first output** of this cycle was the initial characteristics of a Green Pea candidate. Rick Nowell looked for publications mentioning the doubly ionized oxygen in galaxies and communicated his findings as well as the explanation given about the Green Peas. Rick Nowell analyzed the spectra of the Pea posts to that point and decided to collect a subset of them based on the characteristic OIII emission. The **second output** of the cycle was his collection of candidates that fit the spectra criteria and were visibly compact and green. This

Cycle	Input	Mediation	Output	Research category and activities (Table 1)
1	First posted Pea	Group posted Pea-like objects	300 images Question: What is this?	Questions and hypotheses – Ask questions Data – Data collection
2	First spectrum	Analysis of 300 objects Review of literature	Characteristics of a Pea First Pea collection	Analysis – Data analysis Questions and Hypotheses – Explore the literature Communication – Share data and ideas
3	Output from Cycle 2	Search of GZ forum for posted Peas Disconfirming cases of Peas	Confirmed Pea candidate Six Pea criteria	Data – Data collection Communication – Critique Questions and Hypotheses – Negative/positive evidence
4	Three lists of Pea candidates	Science team mediated work Some volunteer contributions	Green Peas paper	Analysis – Data analysis Communication – Sharing data and ideas – Publication

Table 3: IMOI cycles with science research activities.

cycle contained the activities of exploring the literature, communication, and analysis.

Cycle 3

The second cycle's outputs became the third **input** of the work of user starry nite, who began major work searching the greater Galaxy Zoo forum to find candidates. He reposted objects along with the spectra on the GPAC forum, citing the original poster. This **mediation** led to **two early outputs**, which were a confirmed Pea candidate and a more detailed set of criteria for Pea candidates. The **longer-term output** was the second collection of Pea candidates, which fit the six Pea criteria (**Table 2**, Event 5). These candidates presented some disconfirming cases of the characteristics of the Green Peas. There were several objects that looked like a Pea candidate but did not have the other criteria. starry nite's list contained a wider range of objects than fit into the "classic" description. The disconfirming cases made the participants discuss their understanding of the candidates and why there would be objects that fit some criteria like the correct spectra but did not have the appearance of the Green Pea. The group had to decide how to navigate this ambiguity, and they continued to collect objects that were closest to the Green Pea criteria. The group was never able to resolve this issue, presumably because of the lack of astronomical and scientific knowledge. This cycle fit into several activities of the research process: data collection, critique, and negative/positive evidence.

Cycle 4

The last cycle began with Cardamone's introduction into the research process. The **input** was the collective work of the volunteers, but mainly, the lists they generated. Cardamone **mediated** the Pea candidate lists by setting bounds on the population to be studied and by running various analyses. This work was done semi-publicly, as Cardamone kept the "Peas Project" thread informed of her work. The volunteers were unable to contribute to this cycle because of lack of training or knowledge about galaxy analysis. The final **output** of this work was the publication of the study (Cardamone et al. 2009). Cardamone's involvement effectively finished the study of the Green Peas. This cycle exhibited the activities data analysis, communication, and publication of the scientific research process.

Parallel work and incomplete cycles

As the volunteers compared each Pea candidate with the established criteria, they adjusted their understanding of the objects. Each cycle incorporated multiple activities of the research process and showed that those activities happened in parallel with each other. Even though group members did not explicitly identify what they were doing, they worked together through each cycle as they responded to the posts of other users. Because there were multiple people contributing at once, they could work through several activities of research at the same time. For instance, the literature review was happening while more Pea candidates were posted and participants were sharing what they understood about the objects. At several points in the thread, the group adjusted the parameters of objects they were collecting as they incorporated the outputs of the previous cycles of work.

This analysis did not consider every potential cycle in the thread, nor did it identify unproductive cycles or work that did not explicitly relate back to the Green Pea galaxies. Not all of the events highlighted in **Table 2** explicitly fit into a full IMOI cycle; however, they still were important in the progress of the work. Given the collective nature of this work and amount of time over which it took place, all the inputs and mediations could not be identified, though several events could be considered outputs from implicit cycles. One such event was the summary given by Rick Nowell (**Table 2**, Event 7). This communication was one of the overall outputs of the thread as a whole. Another such event was the Pea-picker program written by waveney, as it provided one last refinement of the data before they were given to Cardamone for analysis. This again did not fit into an explicit cycle, as the group was not learning anything new or adding understanding.

Quantitative description

I employed quantitative methods to provide a fuller description of the distribution of the posts over the 11-month period and the contributions of the individual volunteers. All posting data, including volunteer IDs, galaxies viewed, and answer choices, are stored in a structured query language (SQL) database. This allowed the project scientists to aggregate the volunteers' answers to provide morphological data from the images. The SQL database was queried to count the number of posts in each month, the volunteers who had posted in the 11-month period, and the number of images in the posts.

Table 4 illustrates the distribution of the contributions of the 105 participants in the selected portion of the GPAC thread. It compares the contributions of the top 13 volunteers and all other volunteers and illustrates a reason for looking extensively only at the few volunteers who made the largest contributions. People who made only a few contributions were helpful in the overall goal of the work, but the top 13 volunteers facilitated a majority of the discovery process.

Limitations of the volunteers

Throughout the thread, the users acknowledged their lack of training and expertise in astronomical research. While they knew they were contributing in significant ways to the project, they were limited both in their knowledge and by the tools they could find. Even though the six characteristics of a Green Pea Galaxy were clearly laid out, the volunteers did not have an authoritative way of discounting edge cases for appearance or other objects that looked like peas but lacked the characteristic spectrum. Because none of the volunteers had formal astronomical training, it took time to identify the population as novel. Rick Nowell used the strategy of looking up confirmed candidates in a literature database. He found several Peas were mentioned in papers, but not as a class of galaxies. After exhausting the resources available to him, he and others assumed that the population of galaxies had not been studied; however, the volunteers stopped short of declaring they had made a discovery. Once the criteria of the galaxies were defined by the science team, the participants were again able to assist the research.

Conclusion and Implications

To answer the question of whether the discovery of the Green Pea galaxies constituted authentic scientific research, I constructed an integrated model of the scientific research process. This model included three separate models of science, science research, and scientific reasoning, and a model of iterative work. The three models of science provided a robust list of 26 activities of science, and the IMOI model provided a framework to identify how the volunteers worked collectively to identify this unique class of galaxy. This proposed model of scientific research could be used as a framework to study the collective work of other citizen science projects or in science education as a way of explaining the process of science research.

The discovery of the Green Pea galaxies shows that citizen scientists engaged in meaningful scientific research. This work used the IMOI cycle as a framework for analyzing the group's progress in collecting evidence for the existence of a new class of galaxy. The integrated model of scientific research allowed the conversation and work of the volunteers to be characterized in terms of various activities of scientific research. Four distinct IMOI cycles were identified, with multiple instances of the activities of scientific research. This identification and the resulting discovery confirm that the work of the citizen scientists constituted meaningful scientific research. It should be noted that having a confirmed discovery or result is not necessary to establish that scientific research has been done—the process of scientific research will necessarily include unhelpful results or leave scientists with more questions than answers.

Understanding the capabilities of the volunteers and establishing their abilities to engage in meaningful scientific work is important in the future of this and similar projects. In this case, despite the fact that the volunteers collected dozens of Green Pea galaxies, they were limited in their claims about the discovery because of the lack of tools available and the lack of astronomical knowledge and training. To complete this work, a science team member needed to do the data analysis and present the work to the science community. This teaches us that even though the contributions made by citizen scientists at the entry level are valuable, some participants are willing to do more if given the opportunity and access. Understanding

	Posts (%)	Images (%)	Spectra (%)	SDSS links (%)
Top 13 Posters	942 (79)	479 (81)	352 (95)	473 (79)
All other posters (n = 92)	251 (21)	112 (19)	20 (5)	122 (21)
Total (N = 105)	**1193**	**591**	**372**	**595**

Table 4: Distribution of contribution: This table compares the contributions from the top 13 posters to all other posters (n = 92). The number indicates how many posts contain the object identified.

the tools and support they need to do valuable work for science is important in utilizing their willingness to contribute and learn.

Note

[1] In this paper, the term "citizen scientist" refers to people who have contributed to a citizen science project and is synonymous with "participant," "volunteer," and "user."

[2] The text of the users' posts will be kept in original form with regard to spelling, content, and grammar. The communication style of the users shows their personality and level of understanding of the subject as well as the level of formality used.

Acknowledgements

This work was funded in part by the National Science Foundation award DRL-0941610 under the leadership of Lucy Fortson. I thank her and Kyle Willett for informative discussions in the early stages of this work. I also thank Carolin Cardamone for her insight into this fascinating discovery and Leon Hsu, Fred Finley, Dan Endean, Roxanne Radpour, Charity Pihlaja, and Allan Straub for their helpful editing comments at various stages of this work.

Competing Interests

The author is an advisee of a member of the editorial board.

References

Bonney, R., Phillips, T.B., Ballard, H.L. and Enck, J., 2015. Can citizen science enhance public understanding of science? *Public Understanding of Science*, 25(1): 2–16. DOI: http://dx.doi.org/10.1177/0963662515607406

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S.P., Bennert, M., Urry, C.M., Lintott, C., Keel, W.C., Parejko, J., Nichol, R.C. and Thomas, D., 2009. Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 399(3): 1191–1205. DOI: http://dx.doi.org/10.1111/j.1365-2966.2009.15383.x

Cohn, J.P., 2008. Citizen science: Can volunteers do real research? *BioScience*, 58(3): 192–197. DOI: http://dx.doi.org/10.1641/B580303

Corbin, J.M. and Strauss, A.L., 2008. *Basics of Qualitative Research 3e*, Sage Publications, Inc., Thousand Oaks, CA.

Fortson, L., Masters, K., Nichol, R., Edmondson, E.M., Lintott, C., Raddick, J. and Wallin, J., 2012. Galaxy Zoo. *Advances in machine learning and data mining for astronomy*, 2012: 213–236

Giere, R.N., 1991. *Understanding Scientific Reasoning* (3rd ed.) Holt, Rinehart and Winston, Fort Worth, TX.

Ilgen, D.R., Hollenbeck, J.R., Johnson, M. and Jundt, D., 2005. Teams in organizations: From input-process-output models to IMOI Models. *Annual Review of Psychology*, 56: 517–543. DOI: http://dx.doi.org/10.1146/annurev.psych.56.091103.070250

Józsa, G.I.G., Garrett, M.A., Oosterloo, T.A., Rampadarath, H., Paragi, Z., van Arkel, H., Lintott, C., Keel, W.C., Schawinski, K. and Edmondson, E., 2009. Revealing Hanny's Voorwerp: radio observations of IC 2497. *Astronomy and Astrophysics*, 500(2): L33–L36. DOI: http://dx.doi.org/10.1051/0004-6361/200912402

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D. and Murray, P., 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3): 1179–1189. DOI: http://dx.doi.org/10.1111/j.1365-2966.2008.13689.x

National Academy of Sciences, 2012. A framework for K-12 science education: Practices, cross-cutting concepts, and core ideas. National Academy Press: Washington, D.C.

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

Trumbull, D.J., Bonney, R., Bascom, D. and Cabral, A. 2000. Thinking scientifically during participation in a citizen-science project, *Science Education*, 84(2): 265–275. DOI: http://dx.doi.org/10.1002/(SICI)1098-237X(200003)84:2<265::AID-SCE7>3.0.CO;2-5

University of California Museum of Paleontology, 2015. The *real* process of science. Understanding Science. Available at http://undsci.berkeley.edu/article/0_0_0/howscienceworks_02. [Last accessed 18 August 2015].

Wiggins, A. and Crowston, K., 2010. From conservation to crowdsourcing: A typology of citizen science. In *Proceedings of the Forty-fourth Hawai'i International Conference on System Science (HICSS-44)*, Koloa, HI, pp. 4–7. DOI: http://dx.doi.org/10.1109/HICSS.2011.207

Zooniverse Team, 2016. Publications, 2016 Available at https://www.zooniverse.org/about/publications [last accessed 12 January 2016].

How to cite this article: Straub, M C P 2016 Giving Citizen Scientists a Chance: A Study of Volunteer-led Scientific Discovery. *Citizen Science: Theory and Practice*, 1(1): 5, pp. 1–10, DOI: http://dx.doi.org/10.5334/cstp.40

Submitted: 20 August 2015 **Accepted:** 21 January 2016 **Published:** 20 May 2016

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