A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

2013, AER, 12(1), 010109, http://dx.doi.org/10.3847/AER2013002

Measuring the Conceptual Understandings of Citizen Scientists Participating in Zooniverse Projects: A First Approach

Edward E. Prather
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
Sébastien Cormier
Department of Physics, University of San Diego, San Diego, California 92110
Colin S. Wallace
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
Chris Lintott
Department of Physics, University of Oxford, Oxford Ox1 3RH, United Kingdom
M. Jordan Raddick
Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218
Arfon Smith
Department of Physics, University of Oxford, Oxford Ox1 3RH, United Kingdom
Received: 01/25/13, Accepted: 06/5/13, Published: 07/10/13

Abstract

The Zooniverse projects turn everyday people into "citizen scientists" who work online with real data to assist scientists in conducting research on a variety of topics related to galaxies, exoplanets, lunar craters, and solar flares, among others. This paper describes our initial study to assess the conceptual knowledge and reasoning abilities of citizen scientists participating in two Zooniverse projects: Galaxy Zoo and Moon Zoo. In order to measure their knowledge and abilities, we developed two new assessment instruments, the Zooniverse Astronomical Concept Survey (ZACS) and the Lunar Cratering Concept Inventory (LCCI). We found that citizen scientists with the highest level of participation in the Galaxy Zoo and Moon Zoo projects also have the highest average correct scores on the items of the ZACS and LCCI. However, the limited nature of the data provided by Zooniverse participants prevents us from being able to evaluate the statistical significance of this finding, and we make no claim about whether there is a causal relationship between one's participation in Galaxy Zoo or Moon Zoo and one's level of conceptual understanding or reasoning ability on the astrophysical topics assessed by the ZACS or the LCCI. Overall, both the ZACS and the LCCI provide Zooniverse's citizen scientists with items that offer a wide range of difficulties. Using the data from the small subset of participants who responded to all items of the ZACS, we found evidence suggesting the ZACS is a reliable instrument ($\alpha = 0.78$), although twenty-one of its forty items appear to have point biserials less than 0.3. The work reported here provides significant insight into the strengths and limitations of various methods for administering assessments to citizen scientists. Researchers who wish to study the knowledge and abilities of citizen scientists in the future should be sure to design their research methods to avoid the pitfalls identified by our initial findings.

1. INTRODUCTION

The Zooniverse projects (www.zooniverse.org), created and maintained by the Citizen Science Alliance, allow anyone to become a "citizen scientist" by volunteering their time to work with actual scientific data and help scientists pursue answers to contemporary scientific research questions. Zooniverse citizen scientists are given analysis tasks for incredibly large, diverse, and/or complex data sets that are (a) impossible for computers to perform, (b) require an excessive amount of computational resources, and/or (c) require an investment of man-hours that exceeds the capacity of the relevant scientific community. Current Zooniverse projects include large data sets from a incredibly diverse array of subjects, such as images of galaxies from the Hubble Space Telescope, light-curves of stars with potential exoplanets from the Kepler mission, images of lunar surface features from the Lunar Reconnaissance Orbiter, ancient Greek manuscripts from the Egyptian city Oxyrhynchus, and even weather observations made by the Royal Navy during World War I.

In this paper, we focus on responses from participants in two Zooniverse projects: Galaxy Zoo and Moon Zoo. In Galaxy Zoo (www.galaxyzoo.org), the first project created by Zooniverse, citizen scientists classify images of galaxies from the Sloan Digital Sky Survey and from the Hubble Space Telescope. Galaxies are categorized based on traits such as whether or not a galaxy has a spiral morphology, the number of arms in a spiral galaxy, the prominence of central bulge, etc. Since its launch in July 2007, the Galaxy Zoo program has produced significant results, with over 50 million galaxy classifications by over 150 000 people and has led to over 20 peer-reviewed scientific research publications (e.g., Bamford *et al*. 2009; Cardamone *et al*. 2009; Lintott *et al*. 2008; see www.zooinverse.org/publications for a list of publications). Moon Zoo (www.moonzoo.org), launched in 2010, asks citizen scientists to identify lunar surface features in images taken by NASA's Lunar Reconnaissance Orbiter (Joy *et al*. 2011). Craters are identified and tagged by size and shape, and other surface features are similarly identified. To date, over 2×10^6 images have been classified. Figure 1 shows the user interfaces for both Galaxy Zoo and Moon Zoo during the time of our study.

What sorts of people volunteer their time as citizen scientists? Raddick *et al*. (2013) used user forum responses and interviews to characterize the population of citizen scientists and to investigate their motivations for participating in Galaxy Zoo. They found that citizen scientists tend to be middle-aged (with mean age of 43.0 years and a standard deviation of 14.6 years), male (81.3%), college educated (60.4% for those who went through a US-type postsecondary education system and 53.2% for those who went through a UK-type system), and predominantly reside in either the US (36.0%) or the UK (29.8%). What motivates these people to become citizen scientists? Raddick *et al*. (2010) identified over a dozen categories for motivation to participate in Galaxy Zoo. The three most popular reasons, which make up over half of the responses, were (in no particular order) (1) citizen scientists are interested in astronomy, (2) citizen scientists want to contribute to original scientific research, and (3) citizen scientists are awed by the vast scale of the universe.

In order for the scientific community to leverage the contributions of citizen scientists to their fullest extent, we must enable and encourage citizen scientists to participate in data analysis projects over long time intervals, with maximal efficiency, and for analysis tasks of greater complexity. Advancements in the design of citizen science projects and methods to engage citizen scientists must be informed by research into citizen scientists' motivations and self-identities, as well as their knowledge and abilities. Prior research has investigated citizen scientists' motivations and self-identities (e.g., Raddick *et al*. 2010; Raddick *et al*. 2013). This paper reports on our initial investigation into the conceptual knowledge and reasoning abilities of citizen scientists.

The data reported here comes from members of the public serving as citizen scientists who voluntarily provided responses to questions presented through the Galaxy Zoo and Moon Zoo projects. As part of this research program, we developed and administered two new assessments instruments. We designed the first, the Zooniverse Astronomy Concept Survey (ZACS), to be administered to citizen scientists working on Galaxy Zoo. We designed the second, the Lunar Cratering Concept Inventory (LCCI), to be administered to citizen scientists working on Moon Zoo. We collected and analyzed citizen scientists' responses to the ZACS and the LCCI in

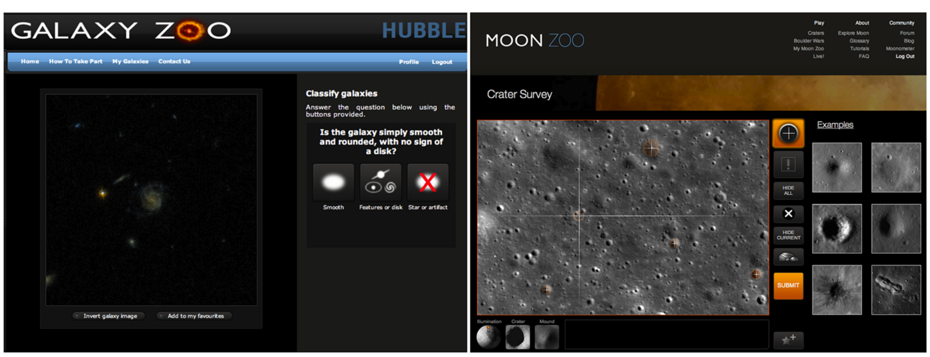


Figure 1. Screenshots of the Galaxy Zoo and Moon Zoo interfaces

order to see if there is a relation between a citizen scientist's level of involvement in either Galaxy Zoo or Moon Zoo projects and his or her understanding of the scientific concepts and scientific reasoning related to the astrophysics of galaxies (as measured by the ZACS) or lunar cratering (as measured by the LCCI). We quantified a citizen scientist's level of involvement by keeping track of the total number of "tasks" he or she had performed at the time he or she responded to an item, where a task is defined, in the case of Galaxy Zoo, as the classification (or sorting) of the morphology of a single galaxy, or, in the case of Moon Zoo, as the tagging (or identification) of features in a single image of the Moon's surface. Because there has been little prior research focused on measuring the knowledge and abilities of citizen scientists, this research also explores the effectiveness of different methodologies for administering online assessments of knowledge and abilities to citizen scientists.

This paper is organized as follows. In Section 2, we describe the process by which we developed the ZACS and the LCCI. In Section 3, we describe the initial implementation of the two assessments and a revised implementation devised to improve participants' response rate. Section 4 contains our analysis of the data. In Section 5, we analyze the responses from the small subset of citizen scientists who completed all the questions of the ZACS in order to perform an initial evaluation of the reliability of the ZACS and its items' discriminations. Section 6 summarizes the results and describes our future plans.

2. DEVELOPMENT OF THE ZACS AND LCCI

When developing the ZACS and the LCCI, we first had to decide which astronomical topics these instruments should cover. For the ZACS, our goal was to create an instrument composed of multiple-choice questions that would effectively probe the understandings of Galaxy Zoo participants on fundamental astronomical concepts related to galaxies. For the LCCI, our goal was to create an instrument composed of multiple-choice questions that would effectively probe the understandings of Moon Zoo participants on fundamental concepts related to lunar surface features and cratering. We ultimately decided that the ZACS and the LCCI should include topics beyond just galaxy physics and lunar cratering, respectively; in particular, we included items on associated astrophysical topics that our research group deemed someone should understand if they are to have more than a superficial understanding of galaxies and lunar craters. For example, we included items on the ZACS about stellar physics because, in order to properly understand the composition of galaxies, one must understand how and why different types of galaxies vary in terms of their stellar populations. The version of the ZACS that we ultimately administered to Galaxy Zoo's citizen scientists was composed of 40 multiple-choice items on light and spectroscopy; nuclear fusion; the relationship between a star's color, temperature, mass, and lifetime; sizes and scales; the star formation history of galaxies; the different morphologies of galaxies; galaxy collisions and the morphology-density relation for galaxies; the evidence for dark matter in galaxies; and the effects of expansion of the universe on galaxies. The version of the LCCI that we ultimately administered to Moon Zoo's citizen scientists was composed of 22 multiple-choice items on features that result from impacts and ejected material; the relative dating of impact features; the relationships between the size of a crater, the shape of a crater, and the energy of the impactor; the origin of lunar craters; the cratering rate as a function of time; and the differences between the maria and the highlands. Both the ZACS and the LCCI contain items that explicitly match to the sorting and identification tasks associated with their respective Zooniverse investigations. However, some of the items on the ZACS and the LCCI extend to concepts beyond what one would need to know in order to successfully perform the sorting and classification tasks required of participants in Galaxy Zoo and Moon Zoo.

The multiple-choice items included in the ZACS and the LCCI were explicitly designed to be well-matched to the conceptual knowledge and reasoning abilities of students who complete a college-level general education introductory astronomy courses on these topics and to elicit and assess a wide range of known conceptual and reasoning difficulties. Note that the majority of students who take introductory general education astronomy courses at the college level are freshmen non-science majors (Rudolph *et al*. 2010), and that the majority of the citizen scientists involved in Zooniverse have college degrees (Raddick *et al*. 2013).

The first drafts of the ZACS and the LCCI were composed of items we adapted from the Center for Astronomy Education's (CAE) existing bank of summative assessment questions (used in prior research efforts, e.g., Hudgins *et al*. 2006; Prather *et al*. 2004; Prather, Rudolph, and Brissenden 2009; Wallace, Prather, and Duncan 2011) and formative assessment Think-Pair-Share questions. We then recruited members of CAE, the Zooniverse team, and galaxy astronomers and planetary scientists, whose specialties overlap with the survey concepts to critique the items for their scientific accuracy and whether or not the items had the right content focus and appropriate level of conceptual rigor for the target audience. We went through several iterations of this critiquing process before project members were in agreement about the final versions of the items and the overall face validity of the ZACS and the LCCI (see Note-1).

3. IMPLEMENTATION OF SURVEYS

3.1. First Implementation of the ZACS and LCCI

Administering the ZACS and LCCI online to citizen scientist participating in the Galaxy Zoo and Moon Zoo programs presented a number of unique challenges that we normally do not encounter when conducting research with students in formal classroom settings. Ideally, we would like to have conducted a traditional data gathering protocol that would have had participants provide responses to all items of the ZACS or LCCI starting on their first day as citizen scientists with the Galaxy Zoo or Moon Zoo projects (pre-test), and then again gathered their responses at some time later, once participants had gained specified levels of experience with performing the tasks associated with their Zoo investigation (post-test). Through such an implementation we would have gathered matched pre/post data for all items of the ZACS and LCCI from participants at all levels of experience with Galaxy Zoo and Moon Zoo projects.

However, for our initial research we were determined to use an implementation strategy that would minimize the risk of a citizen scientist ceasing to volunteer his or her efforts. As a result, we could not require participating citizen scientists to answer items of the ZACS and/or LCCI at any time during any visit to the Zooniverse. Since our participating citizen scientists log onto Zooniverse projects infrequently and for unpredictable amounts of time, and since they were allowed to opt-out of answering individual items, it was difficult to get complete sets of responses from participants, even though the period of our research was conducted over several months. Additionally, at the onset of this research, the experience level of participating citizen scientists ranged from first-time users to experienced citizen scientists who had analyzed several thousand galaxies or tagged several thousand Moon surface images. These issues, coupled together, make conducting a controlled pre/post data collection, as described above, impossible.

In an effort to gather the most descriptive set of question responses possible given the constraints described above, we first adopted the following novel implementation strategy: Every day, one new item from the corresponding ZACS or LCCI was made available to each citizen scientist. The item was randomly selected from the list of items that each citizen scientist had yet to answer. With this implementation, users are never asked to spend more than a couple minutes per visit answering questions. We hoped that all items of the ZACS and LCCI would eventually be answered through consecutive visits by citizen scientists (both relatively new and experienced) who consistently participated in their respective Zooniverse projects. In order to unobtrusively incorporate the survey into the Zooniverse project websites, we elected to provide an optional link in the upper left corner of the user interface. Once the link is clicked, the random question is administered to the participating citizen scientist. In Galaxy Zoo, the link read “What do you know about the Universe?” and it appeared beneath a line of text that read “Take a Quiz.” In Moon Zoo, the link read “Would you like to test your knowledge of the Moon? Take a Quiz.” These links were the only prompts citizen scientists received to take either the ZACS or the LCCI.

The details of this initial implementation of the ZACS and the LCCI is depicted schematically in Figure 2. Each day that a participating citizen scientist visited Galaxy Zoo, he or she had the option to voluntarily answer one item. The number of tasks performed by the user at the time he or she answered the question was saved with each response. In the example shown here, we illustrate how two different citizen scientists (User A and User B) respond to the ZACS or the LCCI over a period of four days. User A logs into either Galaxy Zoo or Moon Zoo each of the four days and responds to a different item each day. The number of tasks completed by User A is recorded at the time he/she responds to an item. User B only responds to an item on Day 2 and Day 4; the total number of tasks he/she had completed are also recorded each time he/she responds to an item.

The ZACS was incorporated into Galaxy Zoo in December 2010 and the LCCI was incorporated into Moon Zoo in May 2010. During the initial implementation of the ZACS, which lasted until August 2011, we collected 11 699 responses to items on the ZACS from 6216 citizen scientists. During the initial implementation of the LCCI, which lasted until June 2011, we collected 5877 responses to items on the LCCI from 4424 citizen scientists. After the end of this initial implementation phase, we found that only 29 out of the 6216 participants from Galaxy Zoo answered all of the items on the ZACS. Of the nearly 4424 Moon Zoo participants who answered

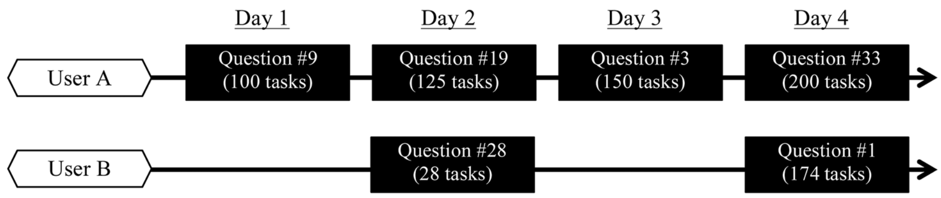


Figure 2. A representation of the initial implementation of the LCCI and ZACS

LCCI items, only 14 answered all of the items. Figure 3 is a histogram of the number of participants versus the total number of questions they have answered for the ZACS or the LCCI. Figure 3 shows that over 78% of the participating citizen scientists answered only one item. This may be due, in part, to the fact that many people log onto Galaxy Zoo or Moon Zoo once, try it out for a short period of time and then never come back. This response rate from participants is typical of many web-based collaborations, surveys, and research efforts (e.g., Adler *et al*. 2008; Anthony, Smith, and Williamson 2009; Ortega, Gonzalez-Barahona, and Robles 2008). With this initial implementation strategy, citizen scientists were able to easily ignore the ZACS and LCCI item link when they visited the Zooniverse site. Through this implementation, we achieved a participant response rate that was well below our desired level.

3.2. Revised Implementation of the ZACS

In order to improve the response rate to items on the ZACS, we decided to try a revised implementation strategy for participating Galaxy Zoo citizen scientists. We revised the implementation as follows: Rather than providing only one new random item with each visit to Galaxy Zoo, we now permitted participating citizen scientists to answer multiple items per visit. The link prompting citizen scientists to take the ZACS was identical to the link we used in the initial implementation described above. Twenty minutes after answering any item, a link to a new item was made available to the participant. The revised implementation strategy ran from August 2011 until September 2012. A schematic illustrating this revised implementation strategy is shown in Figure 4. In the example we illustrate how two citizen scientists (User A and User B) respond to the ZACS during a single session in which they are logged into Galaxy Zoo. User A logs in and answers Item 33 three minutes after it becomes available. Twenty minutes later, Item 5 becomes available and User A answers it four minutes later. After another twenty minutes, Item 12 becomes available to User A. In contrast, when User B logs in, Item 8 is available, but he/she does not answer it for half an hour. Twenty minutes after User B answers Item 8, Item 19 becomes available.

During the revised implementation, we received a total of 20 165 responses to items on the ZACS from 5541 citizen scientists. Figure 5 provides a histogram of the number of participants versus the total number of questions they have answered from the ZACS for both the initial and the revised implementation. As shown in Figure 5, in spite of our revised implementation strategy, the majority of citizen scientists still only answered a single item (57%). A two-tailed χ^2 test reveals there is no statistically significant difference between the two

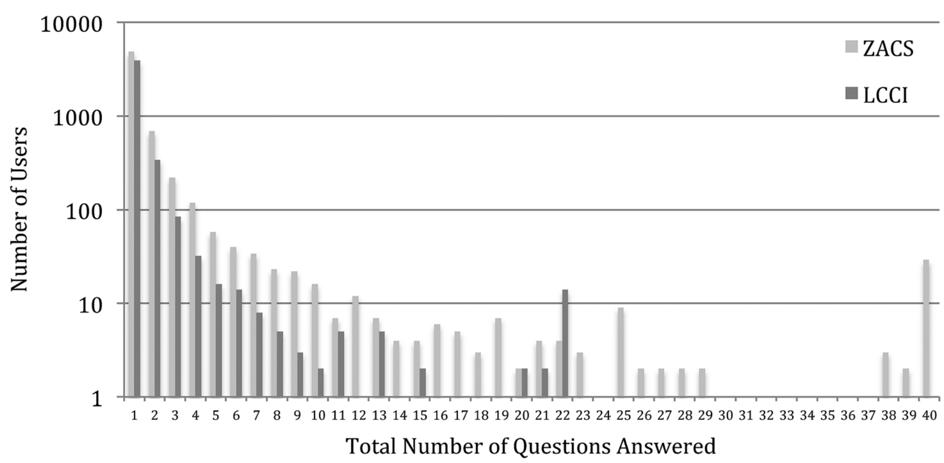


Figure 3. Histogram of the number of participants versus the total number of questions they have answered for the ZACS or the LCCI

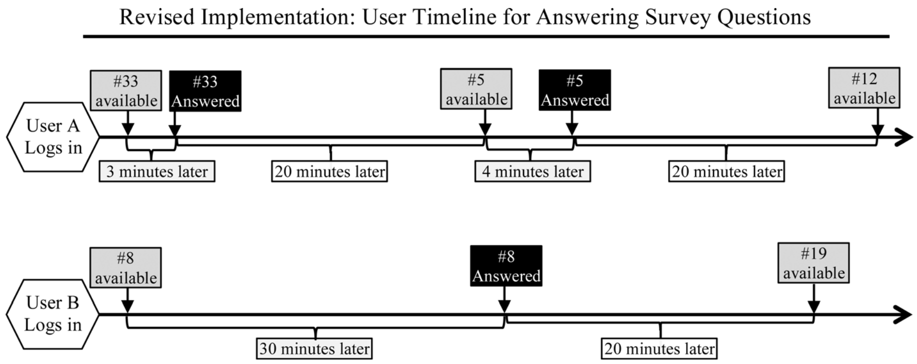


Figure 4. A representation of the revised implementation of the ZACS

distributions shown in Figure 5 ($p = 0.9951$). We therefore combine the data from both implementations in our analysis in Section 4.

4. ANALYSIS AND RESULTS

As a first step in our analysis, we calculated the difficulty of each item on the ZACS and LCCI, where "difficulty" is conventionally defined as the percentage of correct responses (Lord and Novick 1968). This is an important metric to examine because we want the items to span a wide range of conceptual knowledge and reasoning abilities from our target population. If the items are all too easy, then the items will all have high difficulties and we will not be able to differentiate between citizen scientists with different knowledge and reasoning abilities. Likewise, if the items are all too hard, then the items will all have low difficulties and only the citizen scientists with the highest knowledge and reasoning levels will score well, making it difficult to gain information about citizen scientists with lower abilities. Table 1 shows the difficulties of each item on the ZACS and the LCCI, as well as the total number of responses each item received. Figure 6 shows the difficulty of the items on the ZACS and the LCCI, ordered from hardest to easiest. Table 1 and Figure 6 show that both the ZACS and the LCCI have items that span a wide range of difficulties, which indicates that the surveys are composed of items that allow us to differentiate between citizen scientists with different levels of conceptual knowledge and reasoning abilities.

At this point, one typically examines other statistics that quantify the performance of an assessment instrument and its items, such as Cronbach's α and point biserial correlations. However, this is where the limitations of both implementation methodologies manifest themselves. In order to calculate point biserials and Cronbach's α, one must have respondents' total scores for all items of the entire assessment. Since only a small fraction of our respondents answered every item on the ZACS (1% of the combined total from both implementations) and the

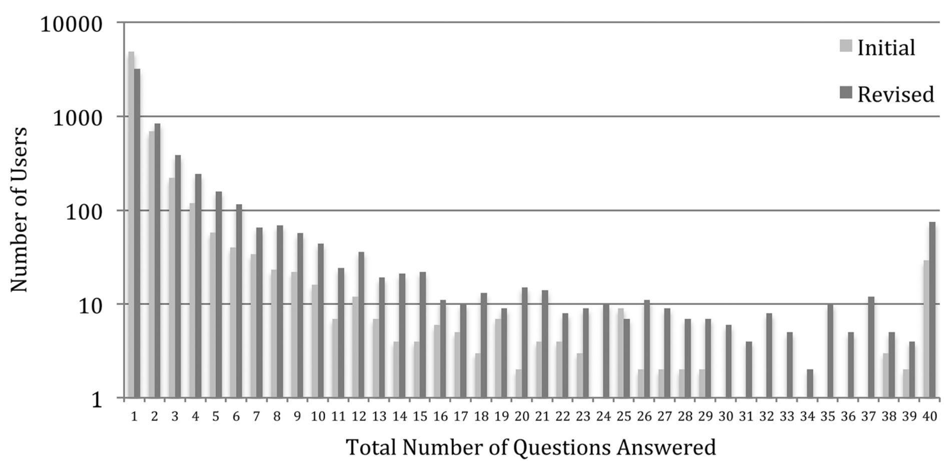


Figure 5. Histogram of the number of participants versus the total number of questions they have answered for the ZACS. Light bars represent data from the first implementation ("initial") and dark bars represent data from the revised implementation ("revised")

Table 1. The difficulties of each item on the ZACS and the LCCI, as well as the total number of responses *N* to each item

ZACS			LCCI		
Item	Difficulty	N	Item	Difficulty	N
1	66%	857	1	23%	239
2	79%	799	2	75%	281
3	75%	839	3	22%	259
4	25%	801	4	47%	248
5	52%	788	5	36%	228
6	36%	804	6	42%	238
7	71%	848	7	29%	268
8	56%	825	8	60%	298
9	74%	807	9	22%	289
10	52%	797	10	56%	264
11	54%	785	11	13%	247
12	49%	813	12	17%	296
13	50%	707	13	70%	290
14	79%	769	14	65%	276
15	25%	771	15	5%	276
16	51%	790	16	65%	266
17	42%	750	17	50%	213
18	56%	803	18	59%	240
19	15%	795	19	53%	273
20	34%	773	20	40%	302
21	47%	763	21	58%	269
22	61%	836	22	65%	317
23	51%	773			
24	60%	848			
25	39%	791			
26	37%	804			
27	52%	813			
28	18%	729			
29	98%	861			
30	69%	831			
31	88%	810			
32	52%	815			
33	23%	756			
34	8%	796			
35	12%	780			
36	46%	752			
37	33%	826			
38	37%	788			
39	69%	760			
40	85%	811			

LCCI (0.3%), we did not have a sufficiently representative data set to calculate these standard psychometric statistics.

As the next step in our analysis procedure, we grouped participant's responses into three bins based on the total number of tasks completed by the citizen scientists at the time they answered an item from the ZACS or the LCCI. The first bin, which we call Level 1, includes responses given when 0 to 200 tasks had been completed. This level was chosen since 200 tasks can easily be completed in a small number of sessions, and many users

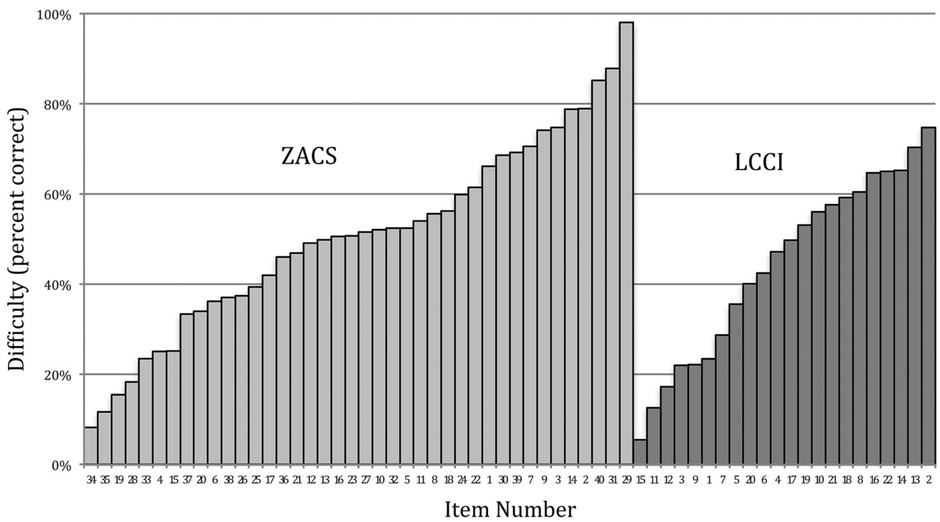


Figure 6. The difficulty of the items on the ZACS and the LCCI, ordered from hardest to easiest

who sign up for Zooniverse do not participate beyond this level. The second bin, called Level 2, includes responses given when 201 to 1000 tasks had been completed, and represents citizen scientists who elected to stay with Galaxy Zoo or Moon Zoo longer than the majority of participants. The third bin, called Level 3, includes responses given when more than 1000 tasks had been completed, and represent the small portion of the citizen scientists who are the most dedicated participants. Table 2 summarizes, for each of these three bins, the number of Galaxy Zoo and Moon Zoo users who fall into that bin and the number of their responses to items from the ZACS and the LCCI. Note that while it is possible for a citizen scientist to appear in multiple bins, in practice only a very small percentage of citizen scientists actually changed bins (9% of Galaxy Zoo participants and 2% of Moon Zoo participants).

Figure 7 shows the percentage of correct responses, across all items on the ZACS, of the Level 1, Level 2, and Level 3 populations of citizen scientists working in Galaxy Zoo. Figure 8 shows the percentage of correct responses, across all items on the LCCI, of the Level 1, Level 2, and Level 3 populations of citizen scientists working in Moon Zoo. We find, for both the ZACS and the LCCI, that citizen scientists with the greatest levels of participation (Level 3) have a higher percentage of correct responses than less experienced users (Levels 1 and 2), and that Level 2 citizen scientists have a higher percentage of correct responses than the least experienced citizen scientists (Level 1). We find this result quite interesting as it suggests that participating in Zooniverse programs leads to increased knowledge and reasoning abilities for the citizen scientist. However, we cannot rule out the possibility that the participants who become more deeply involved with citizen science *already* had higher levels of knowledge and abilities, and that the participants who become less involved had lower levels of knowledge and abilities prior to participating in Zooniverse. Since we do not have data on citizen scientists before they began performing tasks for Zooniverse, we cannot tell if the measured differences in knowledge and abilities are the direct result of participation in Zooniverse projects. We also cannot use traditional tests of statistical significance (e.g., Mann-Whitney, χ^2, *t*-tests, etc.) to determine if the differences between the percentages of correct responses for the three levels of citizen scientists are statistically significant; this is due to the fact that each participating citizen scientists may have answered different items and different numbers of items, within a particular level and between the different levels. Resolving the issues raised above requires a new implementation methodology, different from both of the strategies described in this paper, which at a minimum

Table 2. The number of citizen scientists who fall into the Level 1, Level 2, and Level 3 bins and the number of responses they provided to items from the ZACS and LCCI

		ZACS		LCCI	
Bin	Tasks completed	# of Answers	# of Users	# of Answers	# of Users
Level 1	0–200	9681	6201	4906	4155
Level 2	201–1000	3989	715	582	196
Level 3	1001+	2848	213	389	74

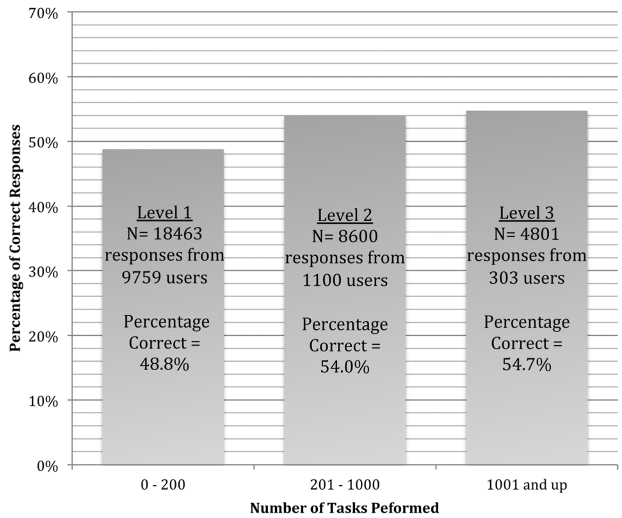


Figure 7. The percentage of correct responses on ZACS items for each of the three bins of citizen scientists (Levels 1-3)

actually requires the participating citizen scientists to provide answers to all the questions of the ZACS and/or LCCI during a small number of visits to their Zooniverse project.

As further evidence that we need a better way to administer the ZACS and the LCCI, consider the subset of items from the ZACS shown in Figure 9. All six of the items in Figure 9 require respondents to compare spiral and elliptical galaxies. In order to correctly answer all six items, one must have a fairly sophisticated mental model that they can use to coherently reason about the differences between the color, gas and dust abundance, and star formation activity of spiral and elliptical galaxies.

We start with a comparison of the first four items, which are similar in that they all ask participants about differences between spiral and elliptical galaxies. If a citizen scientist has a deep understanding of the differences between spiral and elliptical galaxies, then he or she should correctly answer all four items, but if his or her understanding is limited to only the visual aspects of galaxies (i.e., spiral galaxies have a spiral shape), then he or she should do better on Item 1 than on Items 2-4 (note that an understanding of the visual differences between galaxy types is all that is *required* for participating in Galaxy Zoo). This is because Items 2-4 are similar to Item 1, except that they frame the question in terms of characteristics of galaxies (e.g., the abundance of gas, dust, and blue stars) instead of the name of the galaxy type. Do we see different response rates for these four items?

Both Items 1 and 2 were correctly answered by a similar percentage of citizen scientists (59% and 52%, respectively). For Item 3, however, we see that asking participants to choose the galaxies with abundant blue stars causes a large drop in correct answers, to only 39%. Item 4 is the opposite of Item 3 in that it asks citizen scientist to identify galaxies with very *few* bright blue stars as opposed to *many* bright blue stars. We see that the

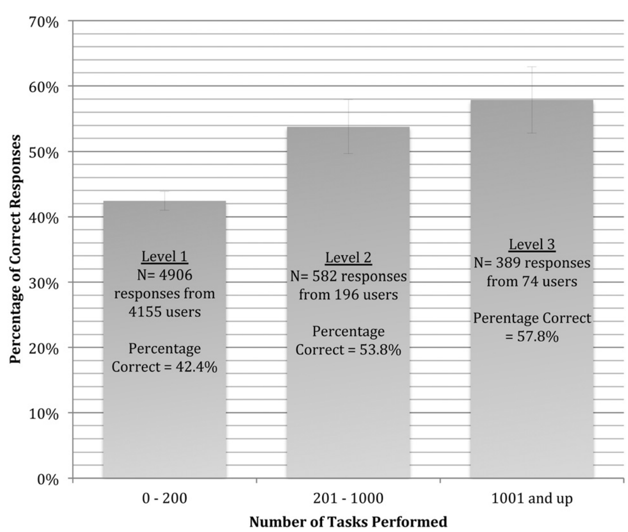


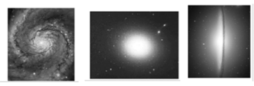
Figure 8. The percentage of correct responses on LCCI items for each of the three bins of citizen scientists (Levels 1-3)

1) How many of the three galaxies shown below could be a Spiral Galaxy?

a. (29%) One
b. (59%) Two
c. (09%) Three
d. (0%) None
e. (03%) There is insufficient information to determine this.

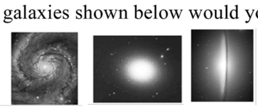
2) In how many of the three galaxies shown below would you expect to see regions of abundant gas and dust?

a. (20%) One
b. (52%) Two
c. (21%) Three
d. (0%) None
e. (7%) There is insufficient information to determine this.



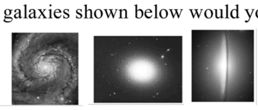
3) In how many of the three galaxies shown below would you expect to contain many bright blue stars?

a. (33%) One
b. (39%) Two
c. (11%) Three
d. (0%) None
e. (17%) There is insufficient information to determine this.



4) In how many of the three galaxies shown below would you expect to there to be very few bright blue stars?

a. (38%) One
b. (37%) Two
c. (6%) Three
d. (04%) None
e. (15%) There is insufficient information to determine this.



5) A galaxy that appears to have very few bright, blue stars, likely:

a. (2%) never had blue stars in the galaxy.
b. (35%) had more blue stars long ago that are no longer present.
c. (52%) has been around long enough for the blue stars to have evolved into red main sequence stars.
d. (11%) never contained enough gas to have blue stars develop.

6) Why are the arms of spiral galaxies blue?

a. (19%) The arms are usually moving toward us, so they are Doppler shifted to blue wavelengths.
b. (23%) The gas and dust in the arms filter out all but the blue light from stars in the arms.
c. (48%) Active star formation is occurring in the spiral arms.
d. (9%) Most of the stars of the disk are in the arms of the galaxy and their light makes it appear blue.

Figure 9. A subset of items from the ZACS that ask respondents to compare spiral and elliptical galaxies. The percentages show the percent of responses given for a particular answer choice. The correct answers are shown in **bold**

same percent of responses are correct for both Items 3 and 4 (39% and 38%, respectively). This suggests that many of the citizen scientists do not have a coherent understanding of the significance of blue stars in relation to star formation, stellar lifetimes and galaxy properties.

Items 5 and 6 further probe citizen scientists' conceptual and reasoning abilities regarding the properties of stars and galaxies. In Item 5 the majority of citizen scientists (52%) chose the incorrect response that blue main sequence stars evolve into red main sequence stars. Only 35% chose the correct answer, which is that the galaxy once had blue stars that are no longer present. In Item 6, 48% of citizen scientists correctly answered that spiral arms "appear" blue because they are sites of active star formation, while 52% were drawn to one of the three choices which model common incorrect ideas.

What are we to make of these responses to the six items shown in Figure 9? The analysis provided above suggest that while many citizen scientists can identify a spiral or elliptical galaxy based on its shape, they do not necessarily understand the correlations that exist between galaxy types and the abundance of gas and dust, significance of bright blue stars, and presence of active star formation. However, as suggestive as these results are regarding the limitation in citizen scientists' understandings of galaxy astrophysics, we are unable to make any definitive claims due to the fact that the vast majority of participants in this study only answered a single item. In order to meaningfully compare responses across items and make inferences regarding participants understandings across different concept domains and lines of reasoning, we will need a compete data set from a representative population of citizen scientist for all the items of the ZACS and the LCCI.

5. CODA: RESULTS FROM 160 CITIZEN SCIENTISTS WHO COMPLETED THE ZACS

Although only 29 citizen scientists completed the ZACS during its initial implementation and only 75 completed it during the revised implementation, there were an additional 56 citizen scientists who completed the ZACS by answering some items during the initial implementation and some items during the revised implementation. This brings the total number of participants who completed the ZACS to 160. This number is large enough to use the responses of this population to calculate point biserials for each item on the ZACS and Cronbach's α for the ZACS.

The point biserials are shown in Table 3. We typically flag items whose point biserials are less than 0.30 or greater than 0.70 (Allen and Yen 1979; Bardar *et al*. 2007; Schlingman *et al*. 2012). As Table 3 shows, twenty-one of the forty items on the ZACS have point biserials smaller than 0.30. This result suggests that many items of the ZACS may poorly discriminate between respondents; however, this group of 160 citizen scientists may not be representative of the larger population of citizen scientists. We intend to reevaluated these results when a more comprehensive data set of complete responses is available.

We calculate Cronbach's α to be 0.78 for the ZACS. This is well above the conventionally accepted minimum value of 0.70 and suggests that the ZACS is a reliable instrument, at least from the perspective of classical test theory (George and Mallory 2009). However, this result should be reevaluated with data from a larger and more representative sample of citizen scientists.

We would like to categorize this special population of 160 citizen scientists into the Levels 1-3 bins discussed earlier and examine whether or not there are any differences between the average scores of the citizen scientists in each bin, but here we encounter a new problem: There is a wide range in the number of tasks these citizen scientists completed between the time they first answered an item from the ZACS and the time they answered the final item from the ZACS. At one extreme, one citizen scientist only completed 10 tasks between the time he or she began the ZACS and the time he or she completed the ZACS. At the other extreme, another citizen scientist completed approximately 20 000 tasks between the time he or she began the ZACS and the time he or she completed the ZACS. The average number of tasks completed by these 160 citizen scientists between answering the first and last item on the ZACS is 1300, with a standard deviation of 1981. This is significant because it means many of these citizen scientists move into a new bin (*i.e*., they go from Level 1 to Level 2, or from Level 2 to Level 3) during the time they take to complete the ZACS. There is therefore no clear way by which we can bin these citizen scientists.

The data from these 160 citizen scientists further emphasizes the difficulties associated with conducting research on the knowledge and abilities of citizen scientists. To get a significant and representative sample of completed

Table 3. Point biserials for the ZACS's items

Item	Point biserial	Item	Point biserial
1	0.34	21	0.33
2	0.11	22	0.25
3	0.28	23	0.37
4	0.38	24	0.25
5	0.17	25	0.25
6	0.45	26	0.38
7	0.46	27	0.30
8	0.36	28	0.16
9	0.43	29	0.21
10	0.25	30	0.41
11	0.37	31	0.08
12	0.43	32	0.00
13	0.33	33	0.04
14	0.30	34	0.06
15	0.07	35	0.19
16	0.10	36	0.32
17	0.30	37	0.31
18	0.22	38	−0.01
19	0.00	39	0.38
20	0.26	40	0.15

assessments requires a significant investment of time (it took over a year to get to 160), and even then we were unable to easily categorize the citizen scientists in this subset of our data, since their involvement in Galaxy Zoo changed greatly over that period of time.

6. CONCLUSIONS AND FUTURE WORK

This paper represents our initial investigation to measure the conceptual knowledge and reasoning abilities of citizen scientists. Over the past two years, we have designed and tested two new assessments, the Zooniverse Astronomical Concept Survey (ZACS) and the Lunar Cratering Concept Inventory (LCCI). In addition, we have tested the efficacy of two different online surveying implementation methods for gathering the responses of participating citizen scientists working with the Galaxy Zoo and Moon Zoo projects. In collaboration with content experts and education experts we established the face validity of the ZACS and the LCCI and then performed an analysis of responses to show that the items of the ZACS and LCCI represent a wide range of difficulties; these two pieces of evidence suggest that these assessments are well-suited for measuring the conceptual knowledge and reasoning abilities of citizen scientists on topics related to the astrophysics of galaxies and lunar cratering. Additionally, the point biserials and Cronbach's α we calculated for the ZACS, based on the subset of 160 citizen scientists who complete all items of the ZACS, show that the discriminatory power of many of the ZACS's items are acceptable and that the ZACS overall is a reliable instrument. By comparing the achievement on the ZACS and the LCCI of three different populations of citizen scientists (Levels 1-3), we found that those citizen scientists who complete more tasks in Galaxy Zoo or Moon Zoo appear to have greater knowledge and reasoning abilities than citizen scientists who complete fewer tasks. However, our initial examination of a subset of items from the ZACS also suggests that many citizen scientists do not possess a comprehensive or coherent understanding of the relationship between a galaxy's morphology (i.e., spiral or elliptical) and the properties and behavior of stars.

While the two different voluntary online implementation strategies used for administering the ZACS differed in terms of how many items a participating citizen scientist could answer in a day, each method was chosen in an attempt to minimize the distraction to citizen scientists working on Galaxy Zoo or Moon Zoo tasks. Unfortunately, the vast majority of participating citizen scientists responded to only a few items on the ZACS or the LCCI. Due to the limited number of items answered by each participant, we are unable to do any further analysis or make any further inferences about differences in the conceptual knowledge or reasoning abilities held by citizen scientists from the different populations of participants within the data set.

Perhaps the most important lesson to take away from this paper is the fact that how one assesses citizen scientists is a uniquely non-trivial problem. We have demonstrated two ways for administering assessments to citizen scientists that yield a lot of data, yet ultimately do not produce the type of data necessary to answer fundamental questions related to the knowledge and abilities of those citizen scientists. As citizen scientists begin to play more significant roles in the analyses of large and important data sets, scientists who are interested in measuring the educational effects of citizen science need to be cautious about using the implementation strategies we have tested here.

We are currently using a new implementation strategy for administering our assessments to Zooniverse's citizen scientists. Instead of providing one item at a time for citizen scientists to answer (or ignore), we formally request that they volunteer to participate in completing the assessment. If they agree, they are given sets of five or six items from an assessment at a time. They are still able to work through the entire assessment at a self directed pace, but will be given reminders if they are inactive for extended periods of time. We believe this will help to ensure that many more citizen scientists provide answers to all the items of our assessments. This new strategy has been in use with a new version of the ZACS since September 2012 and we already have over 500 people who have completed the ZACS. In future publications, armed with a more complete data set, we will conduct a more in-depth analysis of citizen scientists' responses and attempt to uncover whether participation in citizen science investigations leads to improved conceptual knowledge and reasoning abilities.

Notes

Note 1: For inquiries about the ZACS and the LCCI, please contact Edward Prather at eprather@email.arizona.edu.

References

Adler, B. T., de Alfaro, L., Pye, I., and Raman, V. 2008, “Measuring Author Contributions to the Wikipedia,” in *WikiSym’08: Proceedings of the 4th International Symposium on Wikis*, Porto, Portugal.

Allen, M. J., and Yen, W. M. 1979, *Introduction to Measurement Theory*, Long Grove, IL: Waveland Press.

Anthony, D., Smith, S., and Williamson, T. 2009, “Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia *Wikipedia*,” *Rationality and Society*, 21, 283.

Bamford, S. P., Nichol, R. C., Baldry, I. K., Land, K., Lintott, C. J., Schawinski, K., Slosar, A., Szalay, A. S., Thomas, D., Torki, M., Andreescu, D., Edmondson, E. M., Miller, C. J., Murray, P., Raddick, M. J., and Vandenberg, J. 2009, “Galaxy Zoo: the dependence of morphology and colour on environment,” *Monthly Notices of the Royal Astronomical Society*, 393, 1324.

Bardar, E. M., Prather, E. E., Brecher, K., and Slater, T. F. 2007, “Development and Validation of the Light and Spectroscopy Concept Inventory,” *Astronomy Education Review*, 5, 103.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., Lintott, C., Keel, W. C., Parejko, J., Nichol, R. C., Thomas, D., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A. and VandenBerg, J. 2009, “Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies,” *Monthly Notices of the Royal Astronomical Society*, 399, 1191.

George, D., and Mallery, P. 2009, *SPSS for Windows step by Step: A Simple Guide and Reference*, Boston, MA: Pearson Education, Inc.

Hudgins, D. W., Prather, E. E., Grayson, D. J., and Smits, D. P. 2006, “Effectiveness of Collaborative Ranking Tasks on Student Understanding of Key Astronomy Concepts,” *Astronomy Education Review*, 5, 1.

Joy, K., Crawford, I., Grindrod, P., Lintott, C., Bamford, S., Smith, A., Cook, A., and The Moon Zoo Team. 2011, “Moon Zoo: Citizen science in lunar exploration,” *Astronomy & Geophysics*, 52, 2.10.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. 2008, “Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey,” *Monthly Notices of the Royal Astronomical Society*, 389, 1179.

Lord, F. M., and Novick, M. R. 1968, *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. 2008, “On the Inequality of Contributions to Wikipedia,” in *Proceedings of the 41st Hawaii International Conference on System Sciences*, ed. R. H. Sprague, Jr., Los Alamitos, CA: IEEE Press, p. 304.

Prather, E. E., Rudolph, A. L., and Brissenden, G. 2009, “Teaching and Learning Astronomy in the 21st Century,” *Physics Today*, 62(10), 41.

Prather, E. E., Slater, T. F., Adams, J. P., Bailey, J. M., Jones, L. V., and Dostal, J. A. 2004, “Research on a Lecture-Tutorial Approach to Teaching Introductory Astronomy for Non-Science Majors,” *Astronomy Education Review*, 3, 122.

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, 010103.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., and Vandenberg, J. 2013, “Galaxy Zoo: Motivations of Citizen Scientists,” *Astronomy Education Review*, 12, 010106.

Rudolph, A. L., Prather, E. E., Brissenden, G., Consiglio, D., and Gonzaga, V. 2010, “A National Study Assessing the Teaching and Learning of Introductory Astronomy Part II: The Connection between Student Demographics and Learning,” *Astronomy Education Review*, 9, 010107.

Schlingman, W. M., Prather, E. E., Wallace, C. S., Rudolph, A. L., and Brissenden, G. 2012, "A Classical Test Theory Analysis of the Light and Spectroscopy Concept Inventory National Data Set," *Astronomy Education Review*, 11, 010107.

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011, "A Study of General Education Astronomy Students' Understandings of Cosmology. Part I. Development and Validation of Four Conceptual Cosmology Surveys," *Astronomy Education Review*, 10, 010106.

ÆR
010109-1–010109-14