The Meteoritical Society

Meteoritics & Planetary Science 54, Nr 6, 1325–1346 (2019)
doi: 10.1111/maps.13277

Planet Four: Craters—Optimizing task workflow to improve volunteer engagement and crater counting performance

J. SPRINKS[1*], R. HOUGHTON[2], S. BAMFORD[3], and J. G. MORLEY[1,4]

[1]Nottingham Geospatial Institute, University of Nottingham, Nottingham NG7 2RD, UK
[2]Human Factors Research Group, University of Nottingham, Nottingham NG7 2RD, UK
[3]School of Physics & Astronomy, University of Nottingham, Nottingham NG7 2RD, UK
[4]Present Address: Ordnance Survey, Southampton, UK
[*]Corresponding author. E-mail: james.sprinks@ntu.ac.uk

(Received 01 September 2016; revision accepted 11 February 2019)

Abstract–Virtual citizen science platforms allow nonscientists to take part in scientific research across a range of disciplines, including planetary science. What is required of the volunteer can vary considerably in terms of task type, variety, judgment required, and autonomy—even when the overall goal is unchanged. Through analysis of our live Zooniverse *Planet Four: Craters* citizen science platform, the effects of task workflow design factors including volunteer autonomy, task variety, task type, and judgment required on volunteer engagement and crater marking performance were investigated. Website analytics showed volunteers using the Full interface (most autonomy and variety) were more likely to return to the platform, although the amount of time spent per visit was unaffected by the interface used. However, analysis of performance suggested that how this time was used did differ. The interface involving the least complex task resulted in the greatest amount of data and rate of collection, although this also coincided with a greater number of false positives when compared with the expert. Performance in terms of agreement, both between participants and with the expert judgment, was significantly improved when using the Stepped interface for crater position and the Ramped (Mark) when measuring diameter—interfaces that both directly measured the metric with a specific, delineated task. The implications for planetary scientists considering the citizen science route is that there is a balancing act to perform, weighing the importance of volunteer engagement with scientists' data needs and the resources that can be committed to data validation.

INTRODUCTION

Citizen science, or "public participation in scientific research" (Hand 2010), can be described as research conducted, in whole or in part, by amateur or nonprofessional participants often through crowd-sourcing techniques. It increasingly utilizes virtual citizen science (VCS) platforms (Reed et al. 2012) that gather scientific analysis from remotely sensed imagery, both of the Earth and other solar system bodies, through a website interface. Due to the abundance of data, planetary science is a prime candidate for, and adaptor of, citizen science and more specifically VCS platforms. One of the prime planetary science VCS use-cases is that of crater marking, predominantly used as a technique for age estimation (McGill 1977). Relying on both crater identification and measurements of diameter, it is a highly repetitive process and despite a small number of studies suggesting the contrary (Kirchoff et al. 2011; Hiesinger et al. 2012) has been deemed suitable to be undertaken by an untrained audience. As such, a number of VCS platforms have been developed allowing volunteers to mark craters on solar system bodies (moonzoo.org, cosmoquest.org, nasaclickerworkers.com), the outcomes of which have been comparable to the expert equivalent (Robbins et al. 2014).

Despite this apparent success, problems have been identified concerning the use of nonexperts to perform crater marking tasks. Data can be contaminated in a

number of ways, such as false-positive identifications, missing identifications (Tar and Thacker 2015), and issues regarding the measurement of smaller craters (Robbins et al. 2014), all of which contribute to reduce the usefulness of citizen science data sets to the scientific community. In order to combat this problem, a number of retrospective techniques have been proposed to "clean up" the data set, including data filtering (Bugiolacchi et al. 2016), modeling error rates (Tar et al. 2016), and expert comparison (Marshall et al. 2015). It could be argued that such approaches are ironic, creating extra work for the scientist involved within a system predominantly designed to reduce their workload.

Current VCS platforms tend to require the volunteer to carry out tasks in a very repetitious manner, their design arguably driven more by the needs of the scientific problem rather than consideration given to the experience of the citizen scientist (Cox et al. 2015). This could be considered incongruous as the effectiveness of a citizen science platform can be related to its ability to attract and retain volunteers, both in order to analyze the quantity of data required and to ensure its quality (Prather et al. 2013). This study makes a first step in considering how crater marking VCS platforms can be designed to better meet the needs of the volunteer, by exploring whether manipulating task design and presentation can affect their engagement and ultimately the data produced, in order to reduce the retrospective workload required correcting for errors.

First, the background section reviews the relevant literature on the interplay between engagement, performance, and task design in the domains of citizen science, work design, and human–computer interaction. Planet Four: Craters is then introduced—a Zooniverse crater marking citizen science platform consisting of three differing interfaces that vary in task workflow design (TWD) for annotating the surface of Mars. Finally, the results of a live study are presented that directly compares volunteers' engagement and performance across the three interfaces. The impact of TWD on volunteer engagement and crater marking performance, and future implications for planetary science VCS platforms are discussed.

BACKGROUND

Virtual citizen science platforms, including crater marking examples, involve mechanisms and methodologies that have historically been used within similar systems, and as such there exists a wealth of research regarding their design and implementation. For example, volunteers are generally asked to carry out a task from a discrete set of different task types recognizable from typologies of visual tasks (Pelli and Farell 2010): detection (is stimulus present/identifiable?), discrimination (is the stimulus a crater?), and matching (adjusting an attribute of two stimuli until they are equal —for instance, drawing a line from the center to the rim of a crater). Such tasks subsequently force the volunteer to make a corresponding judgment (Farell and Pelli 1999) that can include: yes/no (is a crater present or not), forced choice (what type of crater is it?), and rating scales (assessing the magnitude of an attribute based on a given scale—for example, deciding on the correct size bin a crater belongs to). Research regarding such task types in the realm of image analysis has shown an effect on the performance and experience of the human actor. For example, Hutt et al. (2013) compared the generation of image annotations. Three forms of response were contrasted: classifications, scoring, and ranking, against a ground truth estimate derived from expert annotation. Ranking was found to be the most accurate data versus expert annotation, and also the most reliable in terms of inter-participant agreement, with classification type tasks showing the lowest level of agreement. It was also found that participants produced data comparable with that of experts in terms of overall quality.

There is also the question of how tasks are configured and presented to the volunteer via the platform interface. Current VCS platforms often require volunteers to do the same task(s) repeatedly over a large number of images, in an almost "data entry" like manner for no financial recompense. Arguably, this scenario is analogous to that of the Fordist production line and the fractionation of tasks. Researching this phenomenon, Hackman and Oldham (1975) introduced the "Job Diagnostic Survey" in order to derive an understanding of this type of work and how it could be redesigned to improve engagement and productivity. Design elements including task variety, complexity, and autonomy were identified as important factors, all of which can be influenced by designers of the work. Extending these findings, subsequent research has found a positive correlation between engagement and autonomy (Dubinsky and Skinner 1984; Chung-Yan 2010), task variety (Dubinsky and Skinner 1984; Ghani and Deshpande 1994), and task complexity (Gerhart 1987; Chung-Yan 2010). Although such research predominantly concerns pay for work over extended periods of time, which might not be true of planetary citizen science volunteers (Eveleigh et al. 2014), it acts as the inspiration for this study informing design directions that could be applied to planetary science and other VCS cases.

With these ideas in mind we now introduce the concept of task workflow design as the core construct of

this work. Workflow can be defined as a series of tasks that comprise an overall process that need to be completed in order to take the work from initiation to completion. Its design can involve considerations such as the type of tasks involved, their interaction, and the sequence in which they need to be completed (i.e., sequential or parallel). These considerations can be directly related to the factors described by Hackman and Oldham (1975), and as such could influence engagement and performance. Originally a concept is associated with the manufacturing and business industries (Schmidt 1998; Huang 2002); the notion has been extended to forms of crowd-sourced work due to the analogy that can be made between them. Principally this research has considered TWD in an overarching manner, investigating how complex processes can be deconstructed into tasks that are achievable by untrained participants (Kulkarni et al. 2011, 2012) and how their deconstruction influences performance and engagement (Cheng et al. 2015); other research has considered how certain TWD elements (Dow et al. 2012; Allahbakhsh et al. 2013) and the way tasks are ordered (Cai et al. 2016) can affect overall performance. Existing research regarding the TWD of virtual citizen science platforms has tended toward a retrospective approach, studying the design of existing platforms and their performance in terms of volunteer engagement and data collection (Hutt et al. 2013; Eveleigh et al. 2014; Tinati et al. 2015) and making recommendations and design conclusions based on the findings. In this paper, we introduce the Zooniverse site Planet Four: Craters as the research context within which we can directly manipulate TWD to explore how task variety, complexity, and autonomy affect the user engagement and crater marking performance when using VCS platforms.

PLANET FOUR: CRATERS

Developed in 2013, Planet Four: Craters was created to address two separate goals, i.e., (1) to contribute to scientific efforts to date the surface of Mars and (2) to directly experiment with interface design by controlling for its effects with a single science case. Participants' primary task was to mark the position and size of craters found on remotely sensed imagery of the planet. This section will briefly describe the different tools and interfaces that have been developed for participants to mark craters, and the types of task and judgment they involve.

Crater Marking Tools

Crater Present tool: This is a simple "on/off" button, with which the participant indicates if any craters are present on the image shown (the circle turning red to indicate "yes," see Fig. 1). In essence, this tool facilitates a detection task through making a forced choice (yes/no) judgment.

Crater Position tool: This tool allows users to mark the center of each crater in the image with a simple click of the mouse. It involves both a detection task (is a crater present?) with a matching task (aligning the position mark with the center of the crater) through making a matching judgment.

Crater tool: This tool allows users to mark a circle around the edge of each crater, by clicking in the center and dragging the cursor to the edge. The user can resize the circle to "fine tune" its final position. This also involves a detection task and two matching tasks for each crater (the center and edge) by means of matching judgments.

Interface Design

The three different classification interfaces were distinct in their presentation of some or all of the tools outlined, in order to vary the task type, judgment, task variety, and autonomy. Figure 1 shows the three variations of the interface—Full, Ramped, and Stepped.

Full: The full interface presents all of the tools described to the participant, and allows the participant to use them in any order or way they deem appropriate. Participants even had to decide how many of the tools to use for each image; for instance, if an image contains a large number of craters, the participant may deliberately choose to just press the "craters present" button and move on, without physically marking any of them with the other tools provided.

Stepped: The stepped interface makes all of the tools available to the user but in a very controlled, predefined order. The participant uses each tool and performs each task in turn on each image, and moves on to the next once they have indicated they have finished (through pressing a "next task" button). The tools increase in complexity over each step in terms of the number of tasks and judgments they require.

Ramped: The ramped interface is the simplest of the three, with the participant only using one tool per image. After completing a set number of images (10 in the case of Planet Four: Craters), the tool changes, i.e., the participant presses/depresses the "craters present" button for each image in turn, then marks the center of the craters with the "crater position" tool on each image, before finally marking a circle around the edge of each crater on each image. Each tool change represents a step up in complexity. The images presented for each tool are not the same as those seen previously, i.e., the 10 images analyzed using the

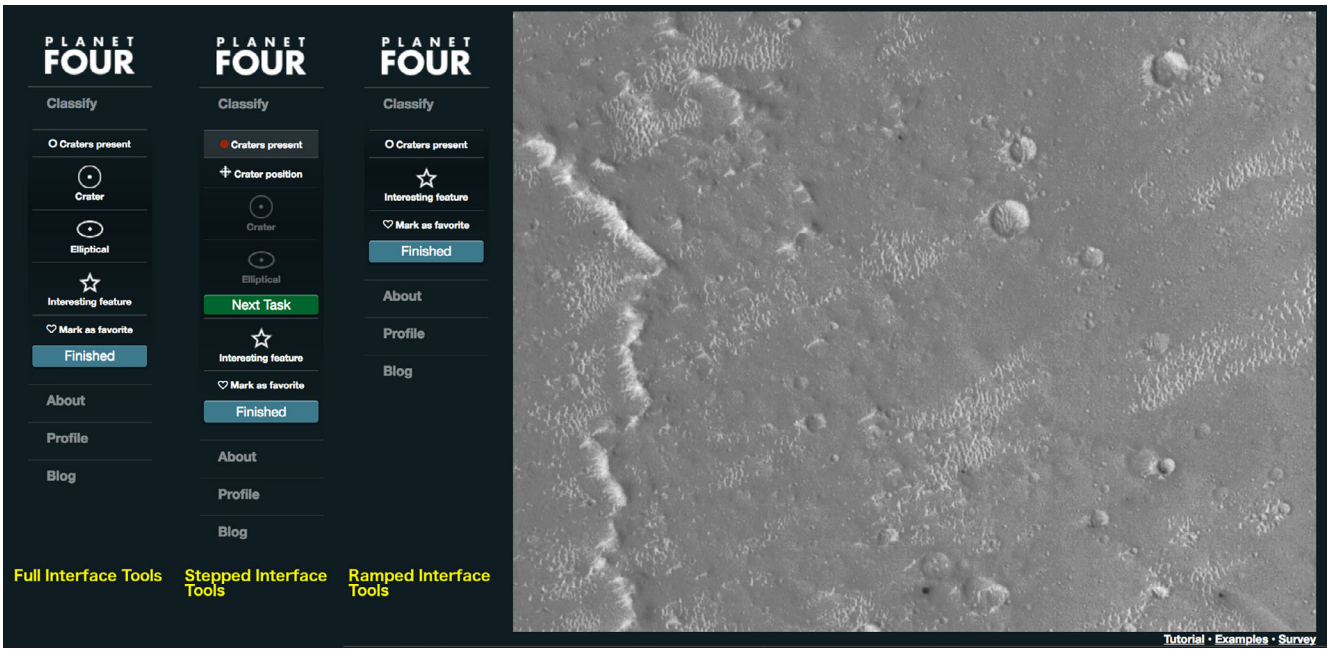


Fig. 1. Planet Four: Craters interface designs. Left to right—Full interface where all tools are available, Stepped interface where tools are used in turn activated by the "Next Task" button, and Ramped interface where only one tool is used for each image. The tool interface appeared to the left of the image being analyzed, as shown in the full screenshot (far right). (Color figure can be viewed at wileyonlinelibrary.com.)

"craters present" tool are not the same as the 10 analyzed using the "crater position" tool and so on. This is an important distinction, as although the tools are presented in the same order as with the Stepped interface, a volunteer would need to classify over 20 images to unlock the crater marking tool—a number that many volunteers did not reach.

METHODS

This study aimed to investigate the effect of manipulating the TWD on volunteer engagement and performance when carrying out the task of marking craters on the Planet Four: Craters citizen science platform. Inspired by the related Human Factors work summarized in the background section, testable hypotheses were formulated for a mix of dependant measures relating to volunteer behavior, engagement, and performance:

- H1: Volunteers performing a less "involved" task and judgment (i.e., Position marking–Ramped interface) produce greater data coverage.
- H2: The type of task performed and judgment made by the volunteer influences volunteer agreement (in terms of the data produced).
- H3: Volunteers using an interface involving greater autonomy, task variety, and complexity (Full) are more likely to return to the platform.
- H4: Volunteers using an interface involving greater task variety (Full and Stepped) will spend less time classifying per platform visit.

Experimental Design

The study took place in a "real world" online scenario where participation was unrestricted both temporally and geographically. Therefore, a between-subjects design has been used, where the task workflow design factors, autonomy, variety, task type, and volunteer judgment were manipulated. Three separate classification interfaces that varied in relation to these factors were employed, created in collaboration with the Zooniverse development team to run online through the Planet Four: Craters landing page. The University of Nottingham Engineering faculty's ethical committee approved the study aims, methods, and procedures, including in terms of guaranteeing the anonymization of data and prior informed consent of the participants. Additionally, contributors to the site agreed to the Zooniverse user agreement and privacy policy (zooniverse.org/privacy) that states: "Data from these projects are used to study online community design and theory, interface design, and other topics."

The four TWD factors (independent variables 1–4) were experimentally manipulated through the design of the three Planet Four: Craters interfaces as described in

Table 1. Task workflow design configuration of each interface.

Interface	Autonomy	Variety	Tasks	Task type(s)	Judgments
Stepped	Set order (Less autonomy)	All tasks per image (Most variety)	Do in order: Is crater present Mark position Mark size	Detection Detection & Matching Detection & 2x Matching	Yes/no Matching Matching
Ramped	Single task (Least autonomy)	Single task per image (Least variety)	Either: Is crater present Mark position or Mark size	Detection Detection & Matching Detection & 2x Matching	Yes/no Matching Matching
Full	Any order (Most autonomy)	All tasks per image (Most variety)	Pick from: Is crater present Mark position Mark size	Detection Detection & Matching Detection & 2x Matching	Yes/no Matching Matching

Table 1, and the impact of this manipulation was measured through participants' behavior regarding platform engagement (Dependent Variable 1, or DV1), and by measuring performance through participant-expert marking comparison (DV2), the number of markings made (DV3), and the time spent classifying each image (DV4).

As can be seen in Table 1, over time participants will perform each task type and make each user judgment use each of the interfaces. However, it is still possible to separate out these factors (in order to consider H1 and H2) through data collected using the Ramped interface, as each task/judgment is performed on a different set of images. Therefore, we can consider this marking data separately, as explained in the Results and Analysis section following. It is also worth noting that there are three experimental conditions represented by the interfaces described, but four main constructs under consideration. This is explained by the interplay that exists between different task workflow design factors (Dodd and Ganster 1996), meaning that realistically one cannot be manipulated without influencing another. For instance, if an interface is designed to maximize autonomy (the Full interface), this also means there must be greater variety so that the participant has the freedom to choose the type of task to complete. Likewise, if a detection task type is required to be completed, this in turn forces the participant to make a "yes/no" judgment—can they detect the crater or not?

Materials

For the study, participants analyzed two images taken by the Context camera on NASA's Mars Reconnaissance Orbiter (G05_020119_1895_XN_09N198W and G23_027332_1907_XN_10N202W). They were chosen because they contain a variety of landscapes common to the Martian surface; scientists at the University of Bristol also provided data from their existing analysis of the first image that was used for ground-truthing so that comparisons could be made between citizen scientist results and those measured by planetary science experts—in order to gauge performance. Before being uploaded to the platform, the images were "sliced" into a number of smaller images that can be more easily handled. Although not as large as high-resolution, push broom type imagery (HiRISE data sets for instance can be gigabytes in size), context camera imagery can often be many megabytes in size (34.6 and 93.8 MB, respectively, for the two images used in this study), making it time-consuming to render to a web browser. A total of 200 smaller image "slices" were created, measuring 840 × 648 pixels with an included overlap of 100 pixels to ensure features on the edges were adequately displayed.

Participants

The Planet Four: Craters site (www.craters.planetfour.org) went live on March 26, 2015. From this time until June 26, 2015, 606 registered Zooniverse volunteers (those who have provided a username and password) visited the site and classified at least one image, contributing a total of 13,136 classifications. A further 13,242 classifications (~50.2%) were made by unregistered volunteers. Due to the nature of studying an online community in an unrestricted environment, this work has an intrinsic, unestimated uncertainty as the same (unregistered) volunteer could, unknown to the investigator, use the website multiple times on different machines, on different browsers, or with cookies turned off.

Although it is not possible to know exactly how many unregistered volunteers contributed, the Zooniverse system assigns a unique ID to each

unregistered user that remains persistent across each visit—and although some users might have their session expired or use other devices, this is assumed to be in the minority (for instance, existing reports regarding disabling cookies [Winnicki 2016; Priebe 2009] put the rate at between 1% and 5%). Additionally, as the interfaces were presented to each participant at random for each visit, it is assumed that any occurrences of session expiry or disabled cookies should be equally likely for each interface design. If a volunteer was to classify craters while unregistered, then decided to register, and continued to classify, the previous (unregistered) ID would be linked to a username and all classifications (before and after registration) would be assigned to the same participant. In addition to the registered volunteers, an extra 974 unregistered session IDs were recorded, resulting in an estimated total of 1580 volunteers that took part in the study. This is in agreement with previous analysis carried out in other Zooniverse sites (Swanson et al. 2015), suggesting that unregistered volunteers can make up as much as two-thirds of the total.

Procedure

When visiting the Planet Four: Craters homepage and selecting the "classify" link, participants were taken to one of the three classifying interfaces (Full, Stepped, or Ramped—Fig. 1) on which the image slices were presented and the tools for marking craters made available. The interface displayed was selected at random, in a nonweighted way so that the probability of seeing each was equal. Once presented, the classification interface remained the same (Full, Stepped, or Ramped) for the duration of the visit and for each classification (although the tools available changed every 10 images with the Ramped interface, as explained in the Interface Design section), until the participant closed down the "classify" page. If the participant subsequently chose to return to the platform to make further classifications, the process was repeated, with a random interface being presented independent of whether the participant had used it previously or not.

If the participant was an unregistered volunteer, or registered but had not used the particular interface before, they were first guided through a tutorial learning how to use the interface and associated tools, marking craters on a separate example image. The marking and behavioral data collected regarding the example image was separated from the data analysis discussed later in this work. The order in which image slices were displayed to the participant was randomized, again in an unweighted way such that each slice was seen a similar number of times during the study. This was to account for bias caused by learning the system, tasks, and each interface. The order in which image slices were displayed to each participant was also randomized, to prevent bias being caused by image content (images with little or no craters appearing in the same interface each time, etc.), learning effects, and fatigue which could, if unaccounted for, influence the data collected.

RESULTS AND ANALYSIS

Dependent variable measures were recorded through participant behavior, both through crater marking performance and platform engagement. The following section presents the results and analysis for each method in terms of their relation to the independent variables regarding TWD.

Website Analytics

Through the study of website analytics associated with the Planet Four: Craters site, measures of participant engagement have been derived in terms of the amount of time they spent on the site, how much analysis they carried out, and how often they returned. When considering the average amount of time spent on the platform by participants, a one-way analysis of variance (ANOVA) (Field 2009) showed no significant difference ($F(2,670) = 0.551$, $p = 0.577$), with the Full, Stepped, and Ramped interfaces having mean times of 20 m 40 s ± 3 m 14 s, 20 m 48 s ± 3 m 50 s, and 17 m 11 s ± 1 m 41 s, respectively (Fig. 2). However, when considering the number of images classified per visit, a difference can be seen. For each interface, the number of images classified followed a lognormal, long-tailed distribution. Therefore, the nonparametric Kruskal–Wallis test (Field 2009) was performed and showed that there was a statistically significant difference between each interface ($\chi^2(2) = 81.75$, $p = 0.001$). A Bonferroni post hoc test revealed that the number of images classified when participants used the Ramped interface (25 ± 2 standard error) was significantly greater than when using the Full and Stepped interface (18 ± 2, $p = 0.001$ and 10 ± 1, $p = 0.001$, respectively). The number of images classified using the Full interface was also significantly greater than when using the stepped ($p = 0.001$).

Finally, although the interface that each volunteer was presented with was random for each visit (i.e., they did not by design have the ability to favor and use the same interface every time they visited the platform), it is possible to evaluate each interface in terms of the rate that volunteers returned to the platform for a second visit. Table 2 shows the number of participants that

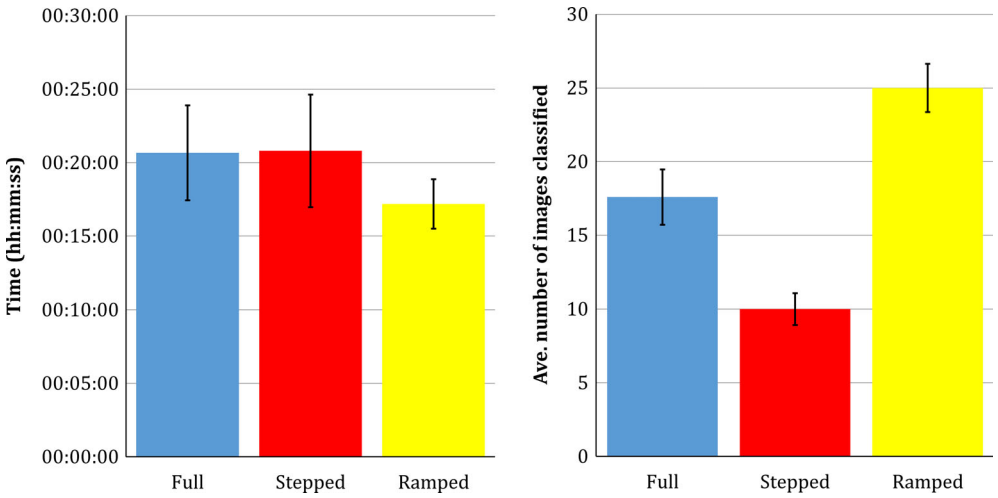


Fig. 2. Time spent on site and number of images classified (with standard error shown) per interface. (Color figure can be viewed at wileyonlinelibrary.com.)

were presented with each interface on their first visit, how many returned to the platform on a second separate occasion, and how many did not return beyond their first visit.

Participants who were presented with the Full interface on their first visit were more likely to return to the site, with 24% returning for a second visit and 76% not returning to the site. In comparison, only 15% and 12% of participants returned to the site when presented with either the Stepped or Ramped interface on their first visit, respectively.

Table 2. Number of returners (volunteers who came back to the platform for a second visit) per first interface used.

First interface	Number of volunteers	Number of returners	Number of non-returners
Full	516	125 (24%)	391 (76%)
Registered	201	50 (25%)	151 (75%)
Unregistered	315	75 (24%)	240 (76%)
Stepped	524	76 (15%)	448 (85%)
Registered	199	27 (14%)	172 (86%)
Unregistered	325	49 (15%)	276 (85%)
Ramped	540	65 (12%)	475 (88%)
Registered	206	27 (13%)	179 (87%)
Unregistered	334	38 (11%)	296 (89%)

Crater Marking Behavior

Participant crater marking behavior has been compared across each interface in terms of percentage of participants who marked craters per image, number of markings made per image, and the time spent on each image. As the Ramped interface requires participants to use each tool individually over a number of images before moving on to the next, the heading "Ramped (position)" represents data where participants only mark the center of craters and "Ramped (mark)" represents results where participants mark the shape. Large values of standard deviation similarly exist, and are again explained by image variation (with some images being comparatively featureless, and others having many to mark).

When considering the percentage of participants that marked at least one crater per image, this again followed a long-tailed, lognormal distribution for each interface due to the influence of images with no craters present. As such, a nonparametric Friedman test (Field 2009) was performed and showed a significant difference between each interface ($\chi^2(3) = 64.56$, $p = 0.001$). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that participants were more likely to identify that an image featured at least one crater to mark using the Full interface ($44.47 \pm 1.64\%$) compared to the Ramped (mark) and Stepped ($36.25 \pm 1.88\%$, $Z = 7.068$, $p = 0.001$ and $39.98 \pm 1.49\%$, $Z = 3.679$, $p = 0.001$, respectively). Participants were also more likely to mark at least one crater when using the Ramped (position) interface compared again to the Ramped (mark) and stepped ($44.63 \pm 1.78\%$, $Z = 6.429$, $p = 0.001$ and $Z = 3.04$, $p = 0.014$, respectively).

In terms of the number of crater markings per image, a repeated measures ANOVA (Field 2009) showed a statistically significant difference between each

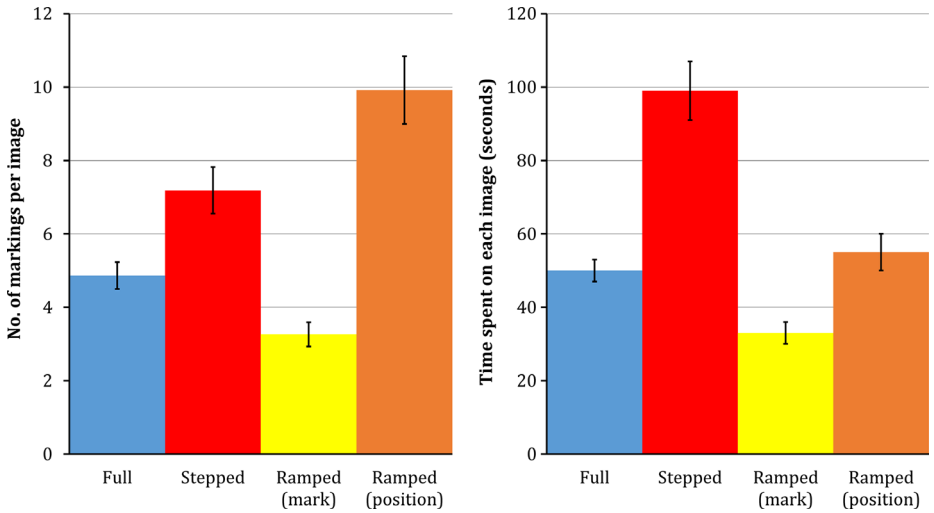


Fig. 3. Crater marking results (with standard error shown). (Color figure can be viewed at wileyonlinelibrary.com.)

interface ($F(1.895, 377.146) = 44.944$, $p = 0.001$). Post hoc tests using the Bonferroni correction revealed that the Ramped (position) interface resulted in a greater number of markings (9.92 ± 0.92) compared to the Stepped (7.19 ± 0.62, $p = 0.001$), Full (4.86 ± 0.37, $p = 0.001$), and Ramped (mark) interfaces (3.26 ± 0.33, $p = 0.001$). Similarly, the Stepped interface also resulted in a greater number of markings compared to both the Full and Ramped (mark) interfaces ($p = 0.001$). Finally, when considering the amount of time participants spent on each image, a statistically significant difference again exists across interfaces ($F(1.924, 382.843) = 32.390$, $p = 0.001$). Participants spent more time per image using the Stepped interface compared to the Ramped interfaces (99 ± 6.36 s versus 33 ± 4.03 s and 55 ± 4.81 s, $p = 0.001$) and more time compared to the Full (50 ± 2.76 s, $p = 0.001$). Participants significantly spent the least amount of time per image using the Ramped (mark) interface compared to the other three. Figure 3 shows the average number of crater markings and the average time spent on each image using each interface.

Participant Agreement

In order to assess participant agreement in terms of crater marking, multiple markings of the same crater (or same perceived crater) by different volunteers have to be assessed. To achieve this, an estimation of which markings relate to the same circular feature has to be made, which is by no means a trivial task. A combination of the lack of formal experience of the participants, along with the limited training provided can produce large variations in crater diameter and positional estimations. To assist with this task, ArcMap GIS software was used, specifically its *Grouping Analysis* function (Esri 2014). The function takes a nearest neighbor clustering approach, performed using x and y position values with diameter. The maximum linking length (i.e., the furthest apart two markings can be considered to be a part of the same cluster) was set as the diameter field to ensure markings belonging to the same cluster at least have some overlap. After performing the function on each interface data set, the clustering results were reviewed by the author in order to check for any obvious omissions or incorrectly amalgamated results. While this type of validation approach obviously introduces a subjective aspect, no clustering algorithm is perfect and makes assumptions based on the constraints it is given (Halkidi et al. 2001), and often inspection by eye is the best solution.

Table 3 shows the crater marking results for each interface, in terms of the number of crater clusters identified (craters marked by more than one participant), the average standard deviation of the cluster center position, and the average standard deviation of the cluster crater diameter in terms of screen pixels (i.e., the level of agreement between participants). By only considering craters marked by at least two participants, any issues regarding malicious annotations or user "mis-clicks" (for instance, double-clicking by mistake) that can exist with crowd-sourced data are considerably less likely to occur, as it is highly improbable that two participants will make this type of error in a similar position on the same image. However, other systematic errors, made due to common misconceptions, image, or crater variability, will be included making it possible to evaluate interface and task design in terms of mitigating their influence (as we have a known, expert solution for comparison).

Table 3. The crater marking participant agreement for each interface (with interquartile range shown). The comparison sample refers to the 57 craters marked using each of the interfaces—and therefore can be directly compared.

	Full interface	Stepped interface	Ramped (Mark)	Ramped (Position)
No. of crater clusters	379	212	83	496
S.D. of position (in pixels)				
All clusters	1.77 (2.30–1.33)	1.97 (2.16–1.23)	1.99 (2.41–1.97)	1.47 (1.85–1.11)
Comparison sample	2.17 (2.85–1.87)	1.62 (2.09–1.20)	1.92 (2.63–1.26)	1.97 (2.42–1.59)
S.D. of diameter (in pixels)				
All clusters	5.83 (8.12–4.29)	3.19 (4.96–1.64)	4.24 (6.41–2.38)	No data
Comparison sample	6.69 (8.51–5.57)	3.41 (4.51–2.22)	4.00 (6.24–2.18)	No data

Additionally, in the absence of an absolute ground truth, many existing citizen science projects use this type of participant agreement as a measure of probability that the feature has been correctly identified (Swanson et al. 2015); its strength increases with the number of participants that have contributed to the marking cluster (this relationship is further explored later in the Results section).

Various approaches have been taken, from setting the minimum number of annotations per cluster based on a comparison with the size-frequency distribution of expert markings (Robbins et al. 2014) to performing a prefilter, removing any markings from volunteers without the requisite experience (Bugiolacchi et al. 2016). Such processes often have the benefit of hindsight, with any filtering and weighting being conducted after the data have been collected and reviewed. As the focus of this paper was to consider task flow and interface factors at the design stage, all of the volunteer data will be included (save those removed for potential malicious reasons as previously mentioned). This allows each interface design to be compared across the data at differing filter levels (2-annotation clusters up to a 20+), as well as the amount of markings being "thrown away" is considered. This is potentially a very important measure considering that volunteers' time and effort is being given for free (Sprinks et al. 2015).

By far the most crater clusters were identified by participants using the Ramped (Position) interface, and by far the least using the Ramped (Mark) interface (~a fivefold difference). These results tally well with the average number of markings per image data described in the previous section. The order regarding the Stepped and Full interface however is reversed, suggesting that although on average participants using the Stepped interface made more markings per image, the fact that significantly fewer images were classified per participant on average has resulted in fewer clusters being identified overall.

Of the total number of crater clusters marked, 57 were identified across all four interfaces, and so can be compared like-for-like in terms of participant agreement (labeled as comparison sample in Table 3). At first glance, it could be construed that by not including all craters that have been identified, this sample could be a misrepresentation. For instance, in Table 3, the Ramped (Position) interface goes from being the interface with the greatest agreement in terms of position to having one of the least. However, this demonstrates the importance of creating it. By comparing only the same craters identified on the same image, it is possible to separate out any effect on agreement caused by crater and image variability. Crater position on screen, size, contrast, lighting, and degradation can vary and all can have an influence on agreement and marking accuracy (Robbins et al. 2014) —which could hide or amplify any effect caused by interface design. While such factors are an important consideration when deciding whether to take the citizen science approach, the focus of this paper is regarding interface and task design—a factor more readily effectible at the design stage of a project. At this point, the performance and agreement of participants are only compared between interfaces, and not with the expert (this is done later in the paper in the Participant–Expert Comparison section). This is in order to evaluate the influence of interface and task design factors at an earlier stage of the process (for instance at the filtering stage where clusters with fewer contributions are removed), before expert comparison is sought or in the case that it is not available.

A Friedman test showed a significant difference between each interface ($\chi^2(3) = 15.042$, $p = 0.002$) when considering agreement in crater position. Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that crater position markings made using the Stepped interface varied significantly less (therefore greater agreement) than those made using the Full (standard deviation

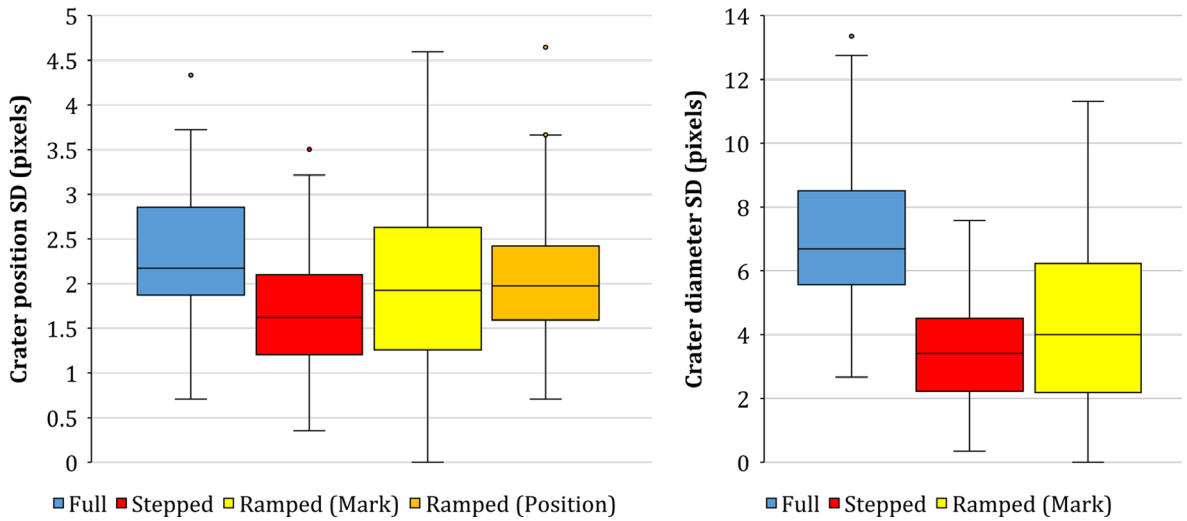


Fig. 4. Comparison of marking participant agreement per interface (median and interquartile range shown). (Color figure can be viewed at wileyonlinelibrary.com.)

median of 1.62 pixels [interquartile range 2.09–1.20] versus 2.17 [2.85–1.87], $p = 0.001$). This is also true when compared to markings made using the Ramped (Mark) and Ramped (Position) interfaces, although the difference is not significant (1.92 [2.63–1.26] and 1.97 pixels [2.42–1.59], respectively). Regarding participant marking agreement in terms of crater diameter, again a Friedman test revealed a significant difference between each interface ($\chi^2(2) = 44.947$, $p = 0.001$). Post hoc analysis using the Wilcoxon signed-rank test with a Bonferroni correction applied showed that the diameter of markings made using the Full interface was significantly less in agreement than those made using the Stepped and Ramped (Mark) interfaces (6.69 pixels [interquartile range 8.51–5.57] versus 3.41 [4.51–2.22], $p = 0.001$ and 4.00 [6.24–2.18], $p = 0.001$, respectively). Figure 4 shows participant agreement in terms of position and size across each interface.

Breaking down participant agreement in terms of crater identification further, Fig. 5 shows the number of crater clusters marked versus the minimum number of participants that marked them. As can be seen, the pattern of a greater number of crater clusters identified using the Ramped (Position) and Full interfaces compared to the Stepped and Ramped (Mark) continues independently of how many participants have contributed to the cluster.

Participants using the Ramped (Position) and Full position also showed better agreement on an individual crater level, with the maximum number of participants marking any crater being 33 and 51, respectively, compared to 11 when using the Ramped (Mark) and Stepped.

Participant–Expert Comparison

Crater clusters identified by participants using each interface have been compared to those identified by experts from the University of Bristol. Through their research in estimating the seismic activity of Mars's Cerberus Fossae region (Taylor et al. 2013), a crater survey discovered a total of 365 craters across the sample area presented to Planet Four: Craters volunteers. Table 4 compares this expert benchmark with the crater marking clusters made using each interface in terms of crater identification.

Breaking down the crater clusters identified using the Full interface, 268 of clusters were confirmed as a crater by the expert giving a precision of 71%, with 111 misidentified as false positives (a false discovery rate of ~29%). Regarding the expert crater markings, 97 (false-negative rate ~27%) were missed and not marked as a cluster by participants. Although participants using the Stepped interface identified fewer crater clusters, a greater proportion was confirmed as a crater by the expert (166, precision ~78%), with 46 (false discovery rate ~22%) misidentified as false positives. Due to the reduced number of clusters identified, more of the markings made by the expert have been missed (199, false-negative rate ~55%). Continuing this trend, the fewest number of crater clusters identified by participants using the Ramped (Mark) interface has resulted in the greatest percentage being confirmed as a crater by the expert (73, precision ~88%), with 10 false positives (false discovery rate ~12%). Likewise, the reduced number of clusters identified overall has resulted in the most expert crater markings being missed

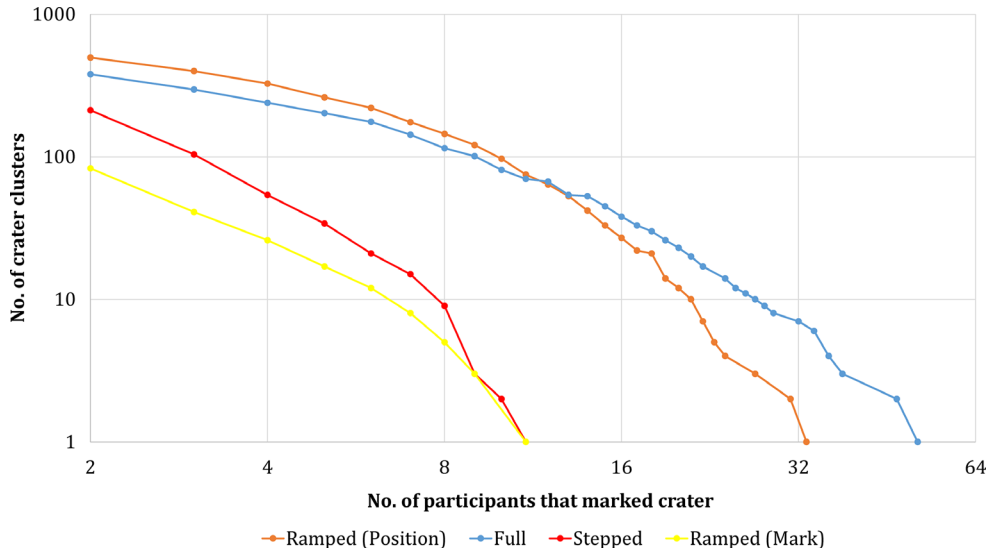


Fig. 5. Participant crater identification agreement. (Color figure can be viewed at wileyonlinelibrary.com.)

Table 4. Crater identification compared to expert.

Interface	No. of craters marked	True positives	False positives	False negatives	Precision (%)	Sensitivity (%)
Full	379	268	111	97	71	73
Stepped	212	166	46	199	78	45
Ramped (Mark)	83	73	10	292	88	20
Ramped (Position)	496	303	193	62	61	83

by participants (292, false-negative rate 80%). Finally, although the Ramped (Position) interface resulted in the greatest total number of clusters that were confirmed by the expert (303), proportionally more clusters were misidentified as craters (193, ~39% of the total). The greater number of clusters made, however, has resulted in fewer of the expert markings being missed by participants (62, false-negative rate ~17%).

Continuing the comparison with the expert markings, out of the 365 identified, 54 were subsequently correctly identified as a crater cluster by participants using all four of the different interfaces, and therefore can be directly compared in terms of their variation from the expert equivalent. Figure 6 shows a "slice" of the study image, with the average markings made using each interface along with those made by the expert. When considering the average difference between participant crater central position and expert crater position, this has been analyzed as a ratio of the size of the crater. A single pixel difference when considering a 200 pixel diameter crater is clearly less significant than when considering a 10 pixel diameter crater and so a percentage difference has been compared. A Friedman test showed a significant difference between each interface ($\chi^2(3) = 11.196$, $p = 0.011$). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that the position of markings made using the Ramped (Mark) interface was significantly farther away from the expert equivalent than those made using the Full and Stepped (average median ratio difference of 0.18; interquartile range [0.27–0.11] versus 0.13 [0.23–0.07], $p = 0.007$ and versus 0.11 [0.20–0.07], $p = 0.003$, respectively). Markings made using the Ramped (Position) interface (0.16 [0.20–0.11]) were also further away position-wise from the expert than the Stepped ($p = 0.05$).

Regarding the average difference between participant crater diameter and the expert equivalent, this has also been analyzed as a ratio of the size of crater. A Friedman test showed that there was a significant difference between each interface ($\chi^2(2) = 5.778$, $p = 0.05$). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction

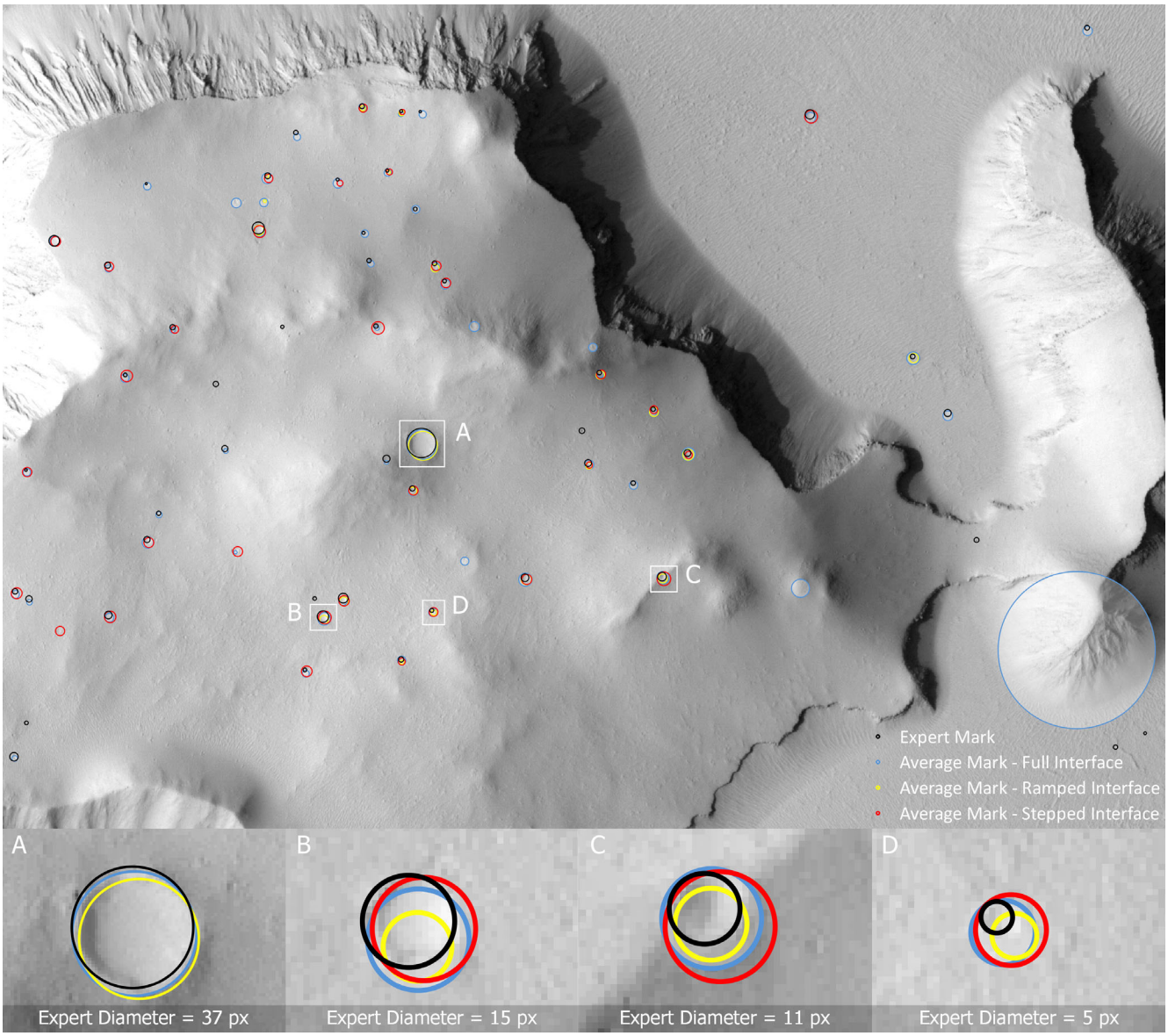


Fig. 6. Sample image of expert and participant average crater markings for each interface. The top, large image shows the expert markings in black, and the average participant marking using each interface (blue = Full, yellow = Ramped, and red = Stepped) calculated using the clustering algorithm described in the Participant Agreement section. The bottom of the image shows four example craters (labeled A to D), in order of size as denoted by the expert. Captioned below each image is the diameter of the expert marking in pixels (px). (Color figure can be viewed at wileyonlinelibrary.com.)

applied revealed that the diameter of markings made using the Ramped (Mark) interface were significantly more in agreement with the expert than those made using the Full and Stepped interface (median ratio difference of 0.23 [interquartile range 0.50–0.10] versus 0.50 [0.83–0.14], $p = 0.001$ and 0.34 [0.73–0.12], $p = 0.03$, respectively). Figure 7 shows the average difference in position and diameter between the expert markings and those made by participants using each interface.

Breaking down crater identification further, the clusters identified using each interface are compared with the expert data in terms of their size and frequency. Figure 8 shows the cumulative crater frequency plots for each interface compared to the expert as a function of crater diameter, and the relative deviation from the expert per size bin.

In terms of general crater identification, the Full interface distribution follows most closely to that of the expert, with the higher frequencies seen perhaps

expected when considering the high number of false positives marked as shown in Table 4. Considering the curve of the Stepped and Ramped (Mark) interfaces, the frequencies at smaller crater diameters are significantly less than that of the expert, while at larger diameters, they are greater. This pattern is shown when considering the relative deviation of each size bin from the expert (right side of Fig. 8), with the four crater bins smaller than 10 pixels in diameter making up to 24% less of the total crater count and the three larger than 10 pixels making up to 35% more (however, the large error due to the small count must be borne in mind) when compared to the expert. Additionally, the relative deviation shown in Fig. 8 also reveals that this pattern regarding smaller and larger crater diameters is also true for the Full interface—a pattern lost in the cumulative frequency plot due to the overall large number of craters marked.

It is also worth noting that the deviation from the expert, both negatively in terms of smaller crater diameters and positively in terms of larger diameters, is greater than can be explained by the numbers of false positives and false negatives shown in Table 4 for each interface. This suggests that a number of the true positive marking clusters, although confirmed as a crater by the expert, have been marked at a larger diameter than the expert and therefore have fallen into an incorrect size bin. This issue can be visualized in Fig. 6 that shows an image slice with markings, where the black expert markings are consistently smaller than those made by volunteers on each of the interfaces. Perhaps this highlights a limitation of the imagery and the annotation tools supplied that interface or task design cannot overcome, with volunteers able to identify smaller craters but unable to mark them accurately. This is an issue that has been highlighted in previous work, where even the accuracy or experts can degrade when considering craters <10 pixels in size (Robbins et al. 2014). If craters of this size are key to the science being addressed, a solution could be to provide "zoom" tools or present the imagery in a "zoomed in" state so that volunteers can more easily mark them.

Finally, Fig. 9 considers the number of participants that contributed to a cluster versus the proportion that are true positive, compared to the expert. Taking the expert data as a ground truth, it can be seen as the number of participant contributions required before a cluster can be seen to definitely represent a crater.

Clusters identified by participants using the Ramped (Mark) interface required the least amount of markings, with all clusters made up of five or more participant contributions also recognized as a crater by the expert. This was followed closely by those made on the Stepped interface, where six or more participant markings were needed. The figure is somewhat higher for clusters made on the Full interface, with those made up of nine or more participant contributions all recognized by the expert as a crater. The Ramped (Position) interface required the highest number of participant markings, as only clusters made up of 13 or more resulted in a 100% expert agreement, over twice as many as was required using either the Stepped or Ramped (Mark) interfaces. This apparent difference in performance in terms of participant–expert agreement also holds true when considering clusters made up of fewer contributions. For example, of those clusters made up of only two participant markings, those made using the Stepped and Ramped (Mark) interfaces coincide with an expert equivalent more often than those made using the Full and Ramped (Position) interfaces (0.66 and 0.83 versus 0.41 and 0.21, respectively).

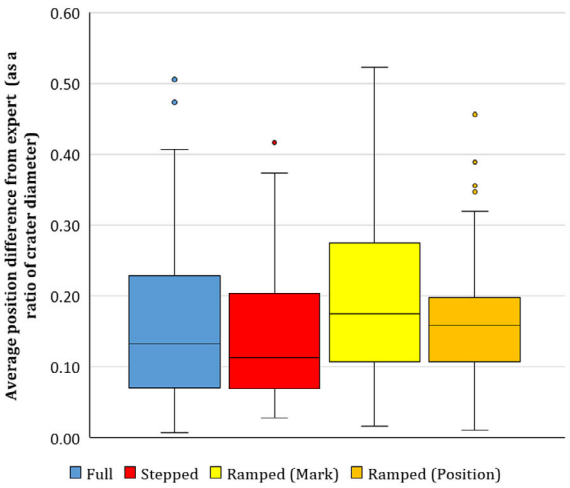
Quality versus Quantity: Modeling Crater Marking Rates

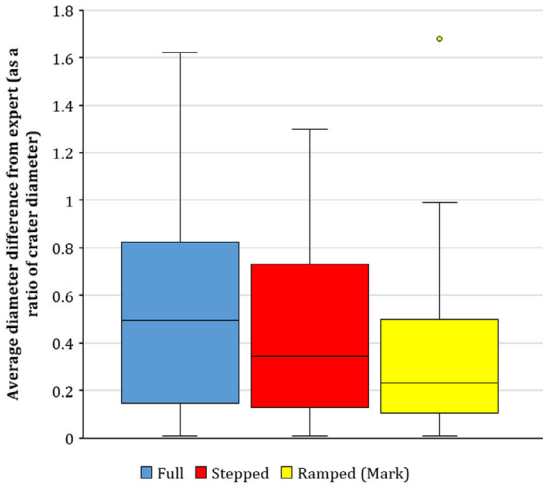
Through synthesizing the crater marking results previously discussed, it is possible to model the crater marking rates of the project over time for different performance criteria. While performance in terms of crater position and diameter accuracy has already been discussed, Fig. 10 shows the crater identification rate (number of craters marked) over time for three different minimum precision levels for each interface. By considering precision, it is not only possible to consider the amount of usable marking data collected over time on each interface but also the amount of data effectively "thrown away"—equating to wasted volunteer effort. In order to calculate the total number of craters identified on a given day since launch for each of the precision rates described, the following formula was used:

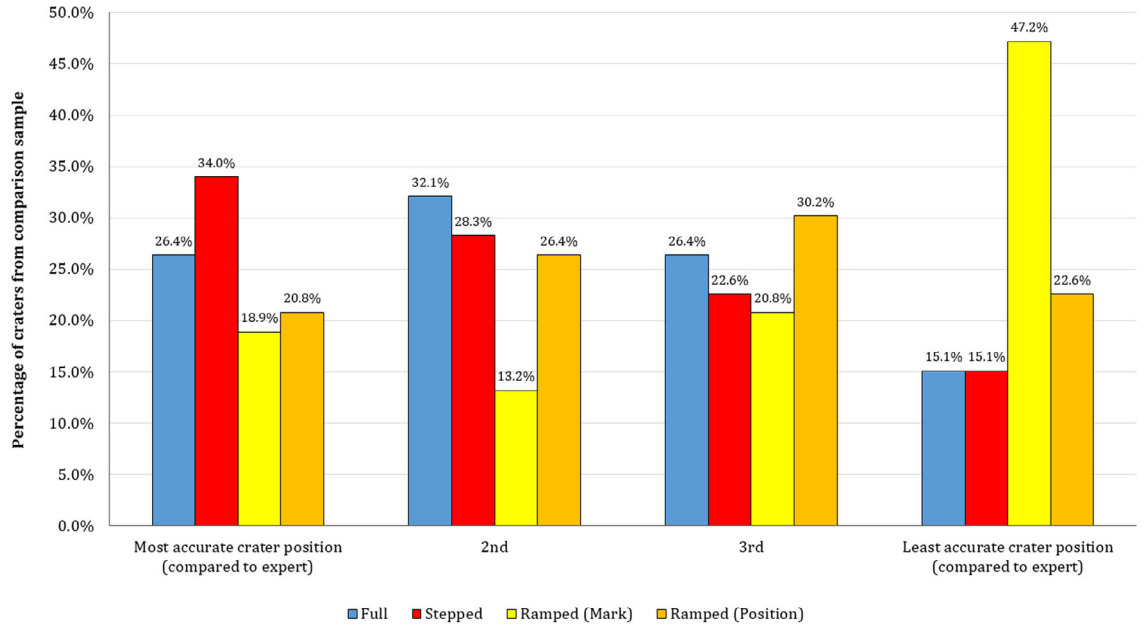
$$\text{Craters identified} = \frac{(V_{\text{n}} + V_{\text{n}}R)IM}{M_{\text{c}}}$$

where V_{n} is the number of new volunteers (calculated from website visitor behavior, Table 2), R is the return rate of new visitors (calculated from website visitor behavior, Table 2), I is the number of images classified by each volunteer per visit (see Fig. 2), M is the number of markings made per image (see Fig. 3), and M_{c} is the number of markings required per cluster to achieve the precision (see Fig. 5).

For all three precision levels, and with each interface, the number of craters identified plateaus over time. This is due to participant visit behavior following a long-tail distribution, with the number of new volunteers falling heavily each day further from launch. This is in common with the majority of existing online


Average position difference from expert (as a ratio of crater diameter)
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Full
Stepped
Ramped (Mark)
Ramped (Position)


Average diameter difference from expert (as a ratio of crater diameter)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Full
Stepped
Ramped (Mark)


Percentage of craters from comparison sample
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
26.4%
34.0%
18.9%
20.8%
32.1%
28.3%
13.2%
26.4%
26.4%
22.6%
20.8%
30.2%
15.1%
15.1%
47.2%
22.6%
Most accurate crater position (compared to expert)
2nd
3rd
Least accurate crater position (compared to expert)
Full
Stepped
Ramped (Mark)
Ramped (Position)

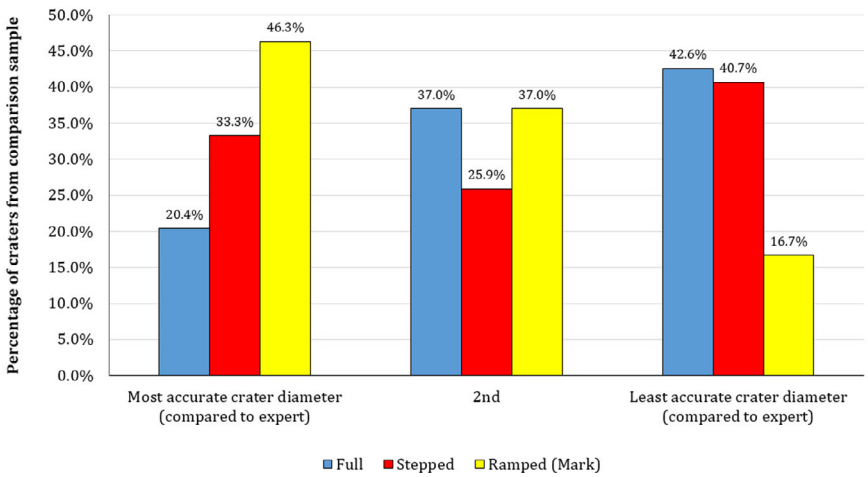

Percentage of craters from comparison sample
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
20.4%
33.3%
46.3%
37.0%
25.9%
37.0%
42.6%
40.7%
16.7%
Most accurate crater diameter (compared to expert)
2nd
Least accurate crater diameter (compared to expert)
Full
Stepped
Ramped (Mark)

Fig. 7. Top: Position and diameter difference of participant markings compared to expert (median and interquartile range shown). The large variations of average diameter difference from the expert are due to variations in crater size, which perhaps hides any differences between each interface. Middle: As the interfaces have been compared on a crater by crater basis, it is possible to show the percentage of craters for which each interface is most thorough to least accurate when compared with the expert. This graph shows the percentage of craters versus the interfaces' accuracy ranking in terms of crater position. For instance, the Ramped (Mark) interface is the most accurate for only 18.9% of craters, but least accurate for 47.2% when compared to the expert position. Bottom: Similarly, this graph shows the percentage of craters versus the interfaces' accuracy ranking in terms of crater diameter. For instance, the Ramped (Mark) interface is the most accurate for 46.3% of craters and least accurate for 16.7% when compared to the expert diameter. (Color figure can be viewed at wileyonlinelibrary.com.)

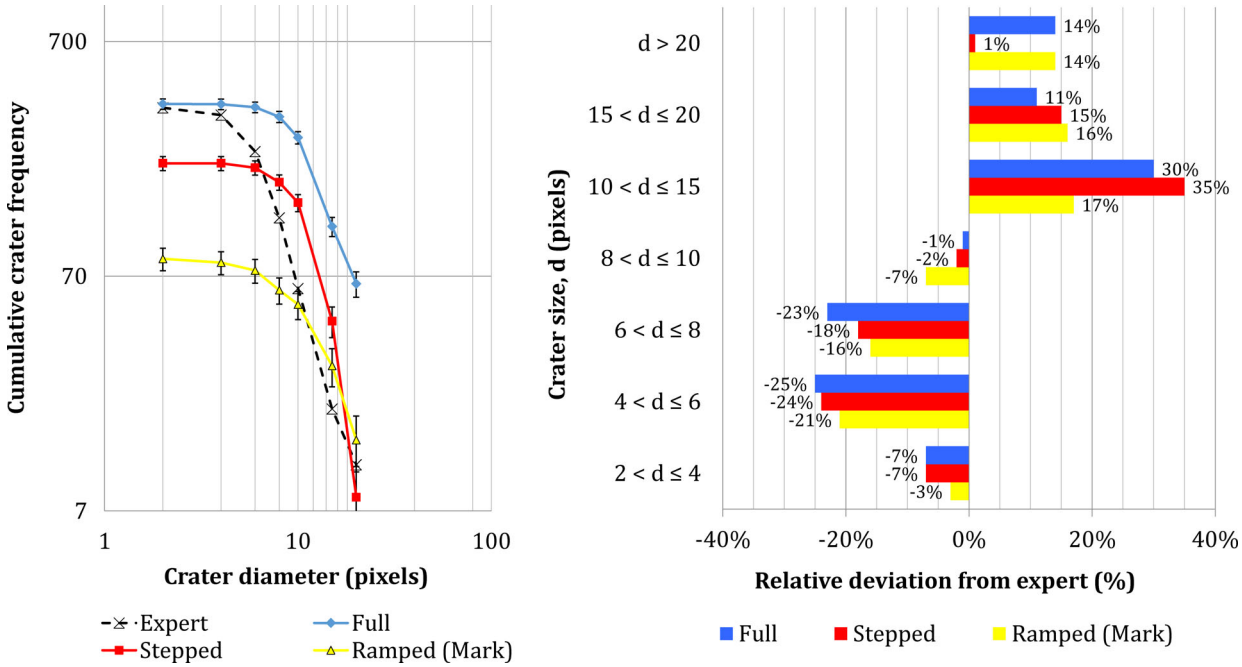


Fig. 8. Cumulative crater frequency plots for each interface compared to the expert (left), and the relative deviation from the expert for each diameter size bin (right). (Color figure can be viewed at wileyonlinelibrary.com.)

citizen science projects where most volunteers only visit once and perform a few tasks (Nov et al. 2011).

If a 100% precision rate is required (i.e., all of the crater clusters identified by participants have been confirmed as craters by the expert), then over 30 days the Stepped interface results in the greatest number of craters being identified, due to the lower number of participant markings required per cluster (six) combined with the high average number of markings made per image (7.19). Although for the first 5 days the Full interface results in the fewest number of craters identified, over time the higher participant return rate becomes a factor, and after 30 days more craters will be identified using this interface than both the Ramped versions.

Considering the 75% precision rate, in this condition, the Ramped (Mark) interface results in the greatest number of crater identifications over 30 days since the launch of the project. This is due to the reduced number of markings required per cluster (two, down from five for 100% precision) overcoming the relatively small average number of markings made per image (3.26). Conversely, the number of markings required per cluster using the Full interface remains relatively high (six), meaning that any positive effect due to the higher participant return rate is dampened resulting in the fewest number of crater identifications over the 30-day period.

The picture regarding the 50% precision condition is similar to that for 100% precision, with the Stepped interface returning the greatest number of crater identifications over 30 days since launch. This is again due to a reduction of the number of participant markings required per cluster (two, down from five required for 75% precision)—a reduction that is not seen with the Ramped (Mark) interface. The Full interface results in a low number of crater identifications to begin with, but again the higher participant return rate becomes a factor over time,

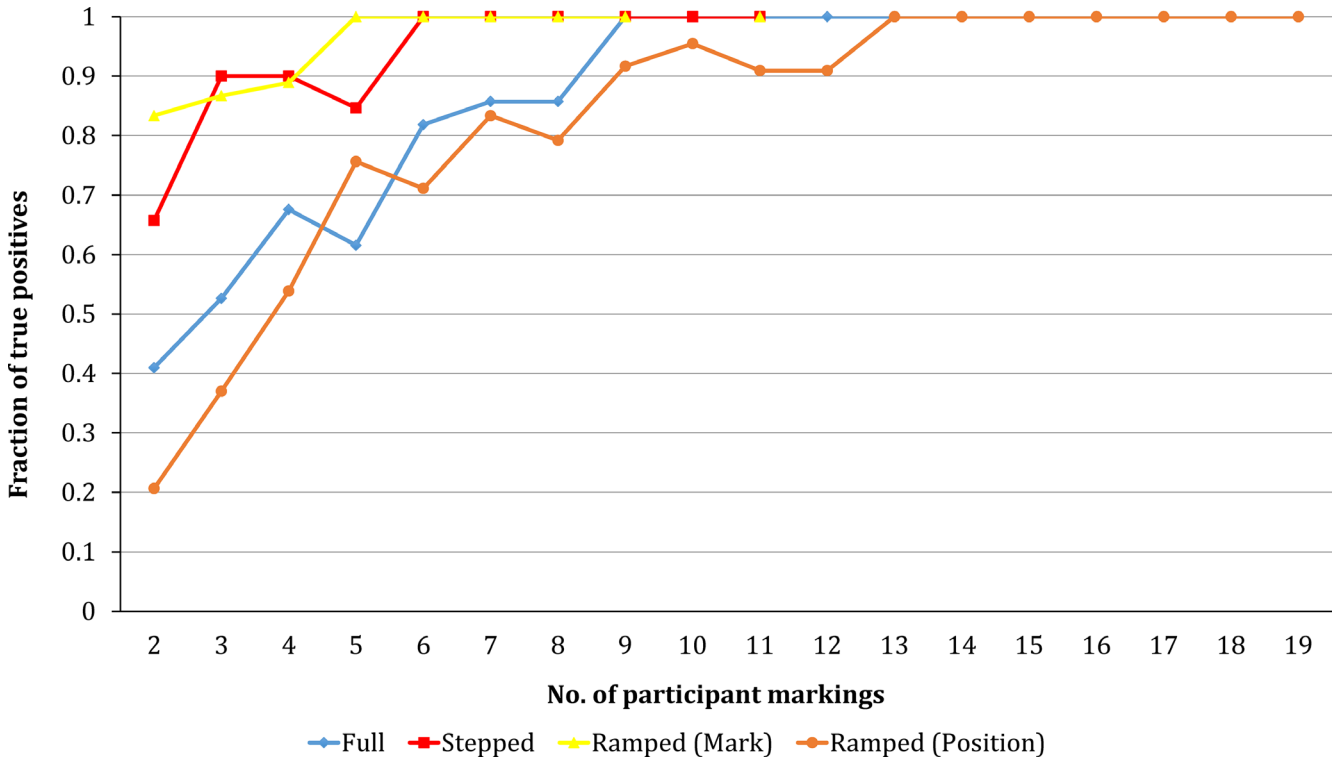


Fig. 9. Participant cluster contribution versus fraction of true positives when compared to expert. (Color figure can be viewed at wileyonlinelibrary.com.)

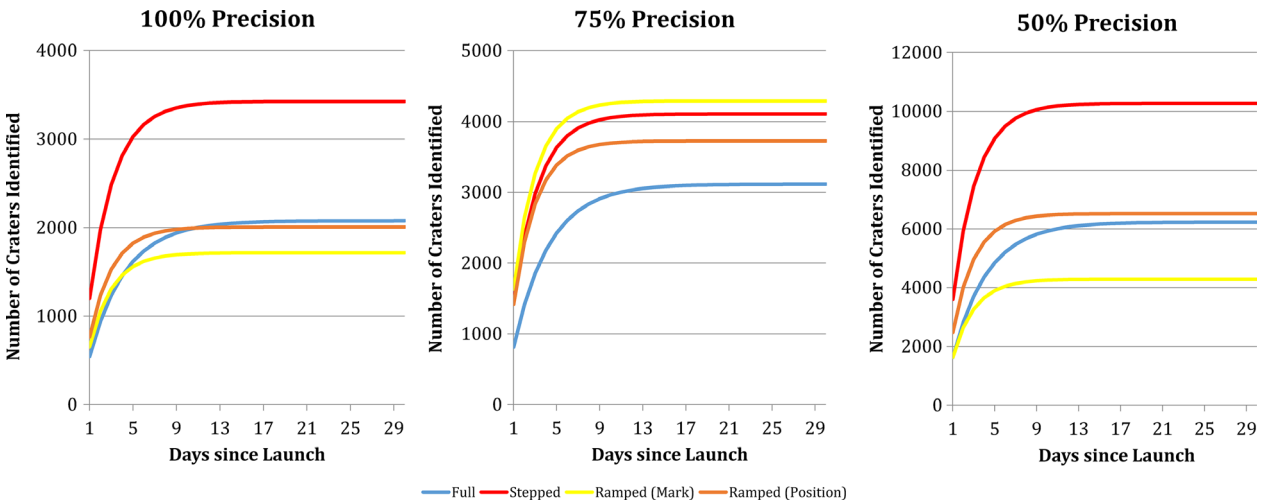


Fig. 10. Number of craters identified over time using each TWD interface at differing minimum precision rates. (Color figure can be viewed at wileyonlinelibrary.com.)

bringing the total of number of crater identifications almost in line with the Ramped (Position) interface over the full 30-day period.

For all three precision conditions, the Ramped (Position) interface represents a "middle ground" in terms of the number of crater identifications made over time. This is due to any positive effect regarding the high number of crater markings made per image, and therefore the high number of clusters identified, being canceled out by the low participant return rate, and high number of participant markings required per cluster (13 for 100% precision).

The model described in Fig. 10 also reveals an interplay between the number of unwanted false positives collected and number of true positives identified. This relationship is strongest when considering the Stepped interface. For instance, at 50% precision over 30 days, the number of craters identified would plateau at ~10,000 according to the model. This has the caveat however of also resulting in an equal amount of false positives to deal with compared with true positives, ~5,000 of each. Increasing the number of markings per cluster to achieve 100% precision (the far left graph of Fig. 10) would remove all false positives. However, while there are no false positives to filter out, the amount of true positives collected over the 30 days plateaus at ~3400 craters. This means that ~1600 fewer true positive craters have been discovered in order to remove the noise of false-positive crater identifications. Adversely, in the Ramped (Mark) interface, only ~400 fewer true positive craters are discovered in order to achieve 100% precision (no false positives) compared to 50% precision (one false positive for each true positive crater identified). However, even at the lower 50% precision, the total amount of true positive craters identified plateaus at ~2000 craters, ~3000 fewer than with the Stepped interface.

Finally, while Fig. 10 considers the number of craters marked over time at arbitrary cutoff precisions, it is also possible to look at the relationship between quantity and quality directly. Through analyzing the number of craters marked and their expert comparison at different cluster sizes (as described in Fig. 9), the relationship between precision and sensitivity can be described. Figure 11 shows this relationship for each of the interfaces.

As expected, with each interface there is a negative correlation between precision and sensitivity, meaning that to remove false positives (increasing precision) some true positive identifications will also be lost (reducing sensitivity). Figure 11 also reveals issues that are not apparent when considering precision alone, as in Fig. 10. For instance, although at 100% precision (all craters marked are true positive) the Stepped interface results in the most crater markings over 30 days (~3400), the sensitivity achieved is only ~6%. This means ~50,000 true positives have not been identified (false negatives). Alternatively, the Full interface results in ~2000 crater markings (1000 less than the Stepped) but at a sensitivity of ~28%, meaning ~5000 false negatives (10 times less than the Stepped interface).

The difference in maximum sensitivity that is achievable on each TWD interface is also highlighted. As alluded to previously, the Stepped interface performs poorly in this respect with a maximum sensitivity of 53% (achieved at 71% precision) when all clusters from ≥2 participant markings are included. The Ramped (Mark) interface threshold is even lower, with a maximum sensitivity of 20% (achieved at 87% precision). The remaining interfaces' maximum sensitivity rates are higher, at 73% (achieved at 71% precision) for the Full and 83% (achieved at 61% precision) for the Ramped (Position). Therefore, if sensitivity (marking the highest proportion of craters that are present) is a key objective, the Full or Ramped (Position) interfaces would be more appropriate despite the lower precision and fewer amount of craters identified over time (see Fig. 10).

Finally, if a balance of both sensitivity and precision is required, again the Full and Ramped (Position) interfaces perform the best. Volunteers using the Full interface could achieve a precision of ~72% while maintaining a sensitivity at the same rate. Considering the Ramped (Position) interface, the figure rises further, with volunteers able to achieve a precision of ~74% while maintaining a similar sensitivity level.

DISCUSSION

Summary and Hypotheses

- H1 (simpler task = greater data coverage) is supported by the finding that more crater clusters were identified and marked using the Ramped (Position) interface—as shown in Table 3.
- H2 (task type and judgment influences participant agreement) is supported by the finding that the Full and Ramped (Position) interfaces resulted in greater participant agreement in terms of identification (markings per cluster)—as shown in Fig. 5. Furthermore, this suggests that the type of task and judgment has a greater effect on agreement compared to variety and autonomy (which varies from greatest to least between the Full and Ramped interfaces). Considering a different measure, crater position and diameter markings made by the Stepped interface (where tasks are separated out and completed in a procedural manner) were significantly more in agreement than those made using the Full interface where tasks are combined (see Fig. 4)—revealing the complex influence of task type and judgment depending on how agreement is considered.
- H3 (task variety and autonomy = more return visits) is supported by the finding that participants who were presented with the Full (most variety and autonomy) interface on their first visit returned at a greater rate compared to the other interfaces (Ramped and Stepped)—as shown in Table 2.
- H4 (task variety = less time per visit) is not supported by the findings, with no statistically

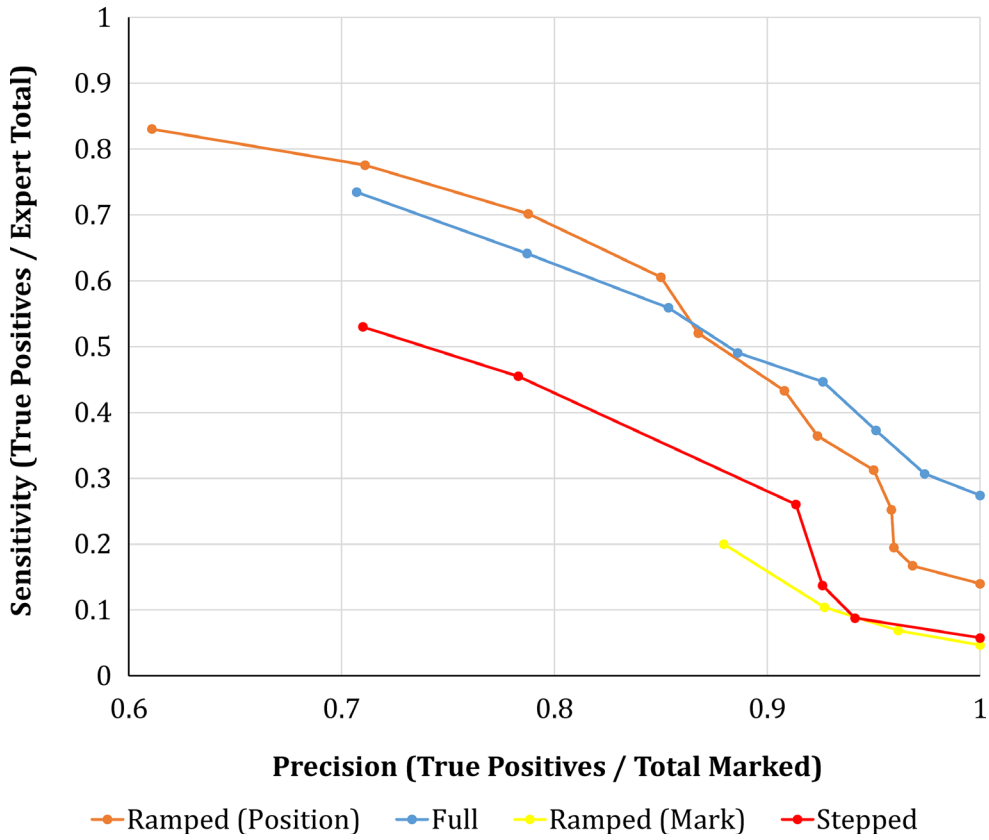


Fig. 11. Rate of precision versus rate of sensitivity for each TWD interface. (Color figure can be viewed at wileyonlinelibrary.com.)

significant difference in time per visit existing between each of the interfaces (see Fig. 2).

In summary, the hypotheses related to task type and judgments were supported by the analyses, whereas the hypotheses related to autonomy and variety were only supported in part. In the following section, this disparity is unpacked through discussing the behavioral findings of this study. Measures of website behavior, participant agreement, and performance compared to the expert are used to paint a broader picture of how task workflow design factors can affect a VCS platform's output and the engagement of its community.

Using Volunteers' Time Wisely

Website analyses regarding participant behavior indicated that varying the task workflow design factors of autonomy, variety, task type, and judgment did not influence how long participants spent on the platform per visit, around 20 minutes across each interface. This suggests that other influences are at play, perhaps external environmental factors. However, through the manipulation of TWD factors, how best to use this time can be controlled at the design stage of the platform.

For instance, one approach could be to utilize an interface involving a simpler task akin to the Ramped (Position) interface (one detection, one matching—one mouse click), resulting in more participants marking at least one crater per image, more crater clusters identified, and more images classified per visit (H1; Eveleigh et al. 2014). However, although this greater quantity of volunteer analysis results in fewer craters being missed (false negatives), it also increases the false-positive rate (crater markings that do not represent an actual crater) that will have to be dealt with. Additionally, the results collected from a simpler task are less detailed, returning only position and crater numbers rather than other metrics such as diameter (required for the age estimation of a surface). While such data might have a use in a citizen science context, for instance to filter out images with no craters before using participants' time on more in-depth analysis, it has little scientific use. Alternatively, a more prescribed interface could be used, forcing volunteers to complete a number of different tasks but in a set order (the Stepped interface). Although this ultimately results in volunteers taking more time per image, and therefore classifying fewer images during the 20-minute visit, arguably it results in a more thorough assessment. Volunteers' markings show both a greater agreement with

each other (H2) and with the expert judgment, although the total number of true positive clusters identified is lower.

The issue with both these approaches is that they either restrict the volunteers' autonomy or the variety of tasks available to undertake. Both these factors influence volunteer engagement (H3), a finding supported by participant behavior in terms of their return rate to the platform. It could be possible to find other ways to improve volunteer retention in such cases, for instance through gamification. The gamification approach has been used by VCS platforms that often involve simple, repetitive tasks in order to make them more enjoyable and therefore motivating and sustaining participation (Eveleigh et al. 2013). However, while this can help sustain the engagement of some volunteers, others can be put off by the competitive aspects, slowing their contribution and in some cases leaving the platform altogether (Iacovides et al. 2013). In a worst-case scenario, some of the most committed "hard-core" volunteers to a project could be lost, eschewing game-like aspects that belittle the importance of the science being addressed in favor of more "serious" interface designs (Bowser et al. 2013). Balancing gamification mechanics can also be a challenge, with consideration of the framing of tasks (i.e., the communication of their contribution) needed to ensure that accuracy and precision is not forfeited in favor of task completion totals (Mekler et al. 2013).

A different approach could be to use an interface that allows volunteers access to all the tools and lets them have the freedom to attempt tasks in any order, as with the Full interface. Although this represents a "middle ground" in terms of data quantity, volunteer agreement, and performance, it does provide a greater intrinsic motivation for volunteers to take part. Although the time spent per visit is not significantly longer, the maximum number of return visits by volunteers using such an interface design does increase.

The Right Tools for the Right Job

Analysis of crater markings in terms of inter-participant agreement supports H2, showing that the type of task presented and judgment required of the participant can affect marking agreement (Hutt et al. 2013). This effect has also been shown to extend in part when comparing markings to the expert equivalent. Expanding further, the results reinforce the importance of being "direct with your volunteers," in terms of providing them with tools that are purposely designed to complete the specific task and harvest the specific data required.

For instance, participant markings of crater position made using the Stepped interface showed a statistically significantly greater inter-participant agreement, an interface where a separate tool is provided to specifically mark the central position. This agreement was reduced with markings made using the Full interface, where position is calculated from the markings of size rather than directly measured. On a similar theme, measures of crater diameter showed significantly greater interparticipant agreement when using the Stepped and Ramped (Mark) interfaces compared to the Full—where marking the crater size is explicitly communicated to the participant as a separate task with a separate tool to use. Diameter measurements made using the Ramped (Mark) interface also showed significantly greater agreement when compared to the expert judgment than those made by participants using the Full interface.

Task Workflow Design and its Influence on Participant Behavior

Beyond the hypotheses addressed through this study as stipulated in the method, analyses of both the crater marking data and website engagement has given rise to other notable patterns in participant behavior. In terms of participant–expert comparison, participants using the Stepped and Ramped (Mark) interfaces showed greater agreement with the expert judgment than the other interfaces in terms of crater position and crater diameter, respectively. This result adds credence to previous research showing that volunteers spending longer on the task perform better (Prather et al. 2013), either by spending more time on each image (Stepped) or analyzing more images (Ramped [Mark] interface). However, the reduction in crater clusters marked does result in fewer true positives identified and more false negatives missed—meaning that the overall size-frequency distribution using each shows little agreement with the expert plot compared with the Full interface. Although participants using the Ramped (Mark) and Stepped interfaces identified fewer crater clusters and fewer participants contributed to each cluster, those that were marked required fewer participant contributions before the fraction of true positives reached 100% (all clusters of 5 and 6 participant markings compared to 9 and 13 when using the Full and Ramped (Position) interfaces, respectively). This ultimately means much fewer false positives to filter out during analysis.

These direct findings in conjunction with the derived relationships regarding crater identification over time and quantity versus quality presented in Figs. 10 and 11, suggest that it is possible to use TWD to influence participant behavior toward a specific need with regard to the data collected. For example, if the science case requires a sample of craters for further study, and therefore is less concerned about the number of false negatives, then 100% precision would be the goal (Fig. 11). The most suitable interface to use would

therefore depend on whether the size of the sample was of greater importance (Stepped, Fig. 10), or the accuracy of crater size within the sample (Ramped [Mark], Fig. 7). Alternatively, a science case could already include an existing data set, maybe compiled by an expert user-group, or identified as being gold standard (Freitag et al. 2016). Therefore, sensitivity could be deemed the more important performance indicator rather than precision (as false positives can be filtered out more easily), meaning the Full or Ramped (Position) interfaces would be the best design approach (Fig. 11). These two interfaces would also be the most suitable if a balance of precision (fewer false positives) and sensitivity (fewer false negatives) were required, as previously explained. It is worth noting that all such approaches could be improved through recognizing and addressing issues of image presentation (systematic across all the interface designs). As shown in Fig. 8, in agreement with previous research (Robbins et al. 2014), there is a limitation regarding smaller craters (<8 pixels in diameter), where volunteers either fail to mark them or mark them at a larger size. This could be solved through providing zoom-like tools, which would improve performance across each TWD approach.

The practical outcomes of these findings suggest that when considering utilizing a VCS platform the science team involved would have to balance the advantages of either greater precision, sensitivity, or data quantity with the disadvantages of having to "clean" out a greater number of false positives, or miss out on a number of potential markings (false negatives). A second balancing act would also be required when considering the volunteer community. One approach might be to build an interface that focuses on providing the user with as much task variety and freedom to complete them as possible, increasing volunteer engagement, and hence their intrinsic motivation to return to the site. Alternatively, an interface could be developed that restricts these factors, either by forcing volunteers to step through each task sequentially or by only making more involved tasks available when a set number of images have been completed. Although this could result in a certain amount of frustration on the part of the volunteer, and therefore a smaller community with fewer return visits, those that do remain might well return more accurate data. In considering such decisions, the level of detail required (i.e., crater existence, position, size, distribution, etc.), the number of images that need to be analyzed, and the potential size of the volunteer community taking part will all need to be considered.

CONCLUSION

Through the implementation of a "live" VCS platform study to test the effect of manipulating task workflow design factors, it was found that autonomy, variety, task type, and the volunteer judgment required had an effect on volunteers' behavior in terms of site usage and the data they produced.

Participant behavior in terms of interaction with the platform showed that although there is no significant difference in terms of visit duration between each interface (~20 minutes), participants who used the interface with the greatest variety and autonomy (Full) returned to the site more often. How participants' behaved during a visit did however vary, with participants using the Stepped interface (least autonomy) spending significantly more time on each image, and therefore analyzing less images in total.

When considering crater marking behavior, analysis has indicated that by manipulating task workflow design factors through different interface designs, performance can be influenced in differing ways depending on the type of measure considered. The interface involving a simpler task, less variety, and less autonomy (Ramped Position) resulted in more data being collected (Eveleigh et al. 2014) and at a faster rate in terms of image analysis time per volunteer (Jäkel and Wichmann 2006). The Full interface, featuring greater variety and autonomy, also resulted in more data being collected than those with greater restriction (Stepped and Ramped Mark). Although at first glance this supports H1 (greater autonomy = greater data volume), there is a caveat. Although a greater amount of data resulted in more true positive identifications when compared to the expert, it also increased the false discovery rate. This in turn resulted in more participants required to contribute (seeing the image and detecting the crater) before a crater cluster could definitely be considered a crater.

Performance in terms of crater measurement agreement, both between participants and with the expert judgment, was significantly improved when using the Stepped interface for crater position and the Ramped (Mark) interface for crater diameter. One conclusion to make from this finding is in support of the previous study, suggesting that performance is improved when participants completed a task that directly measured the required metric, and that the definition of the task is clearly separated from others (Hutt et al. 2013). A second reasoning has been revealed by the "live" nature of the study. It could be argued that this improvement in agreement could be a result of constraining the user in terms of their time and experience, being forced to spend more time on an image using the Stepped, or having to analyze a set number of images to "unlock" tasks with the Ramped (Prather et al. 2013). Although this approach is detrimental to user experience, and thus reduces the number of return visits, those volunteers that do remain perform better compared to the expert judgment.

Overall, the results of this study support the findings of previous related human factors research when considering volunteer engagement, with preference given to greater autonomy and variety suggesting that interfaces that incorporate these factors can provide an intrinsic motivation to take part. In terms of volunteer performance, again the influence of task workflow design factors differs depending on the performance measure concerned. The live nature of this study has additionally revealed the delicate balance between volunteer engagement and performance—reinforcing the importance that VCS developers and science teams consider the analysis required, the amount needed, and the prospective size of their volunteer community when considering a citizen science approach. While previous research has shown that citizen science overall is a valid method to use for crater counting and comparable to the expert equivalent (Robbins et al. 2014; Bugiolacchi et al. 2016), this study demonstrates the importance of considering how the task is designed and presented to the volunteer and its potential impact on the success of a project. It also reveals the advantages that can be realized through experimenting with different TWD configurations, both in order to manipulate volunteer behavior toward the data needs of the project, and to uncover limitations regarding the imagery presented.

Acknowledgments—This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/G037574/1, EP/G065802/1]; and the European Union's Seventh Framework Programme [grant number FP7/2007-2013], under iMars grant agreement no. 607379. It uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. Special thanks to Brian Carstensen, web developer and Michael Parrish, software developer based at the Adler Planetarium, Chicago for their support in developing the Planet Four interfaces. Special thanks also to Jenny Taylor, planetary seismologist at the University of Bristol, for developing the crater counting science case and identifying the required imagery.

Editorial Handling—Dr. A. J. Timothy Jull

REFERENCES

Allahbakhsh M., Benatallah B., Ignjatovic A., Motahari-Nezhad H. R., Bertino E., and Dustdar S. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17:76–81.

Bowser A., Hansen D., and Preece J. 2013. Gamifying citizen science: Lessons and future directions. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Presented at the CHI 13, Paris, France.

Bugiolacchi R., Bamford S., Tar P., Thacker N., Crawford I. A., Joy K. H., Grindrod P. M., and Lintott C. 2016. The Moon Zoo citizen science project: Preliminary results for the Apollo 17 landing site. *Icarus* 271:30–48.

Cai C. J., Iqbal S. T., and Teevan J. 2016. Chain reactions: the impact of order on microtask chains. Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems.

Cheng J., Teevan J., Iqbal S. T., and Bernstein M. S. 2015. Break it down: A comparison of macro- and microtasks. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15. New York: ACM. pp. 4061–4064.

Chung-Yan G. A. 2010. The nonlinear effects of job complexity and autonomy on job satisfaction, turnover, and psychological well-being. *Journal of Occupational Health Psychology* 15:237–251.

Cox J., Oh E. Y., Simmons B., Lintott C., Masters K., Greenhill A., Graham G., and Holmes K. 2015. Defining and measuring success in online citizen science: A case study of Zooniverse projects. *Computing in Science & Engineering* 17:28–41.

Dodd N. G. and Ganster D. C. 1996. The interactive effects of variety, autonomy, and feedback on attitudes and performance. *Journal of Organizational Behavior* 17:329–347.

Dow S., Kulkarni A., Klemmer S., and Hartmann B. 2012. Shepherding the crowd yields better work. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. New York: ACM. pp. 1013–1022.

Dubinsky A. J. and Skinner S. J. 1984. Impact of job characteristics on retail salespeople's reactions to their jobs. *Journal of Retailing* 60:35–62.

Esri. 2014. *Grouping Analysis - ArcGIS Pro*. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/grouping-analysis.htm

Eveleigh A., Jennett C., Lynn S., and Cox A. L. 2013. I want to be a captain! I want to be a captain!: Gamification in the old weather citizen science project. Proceedings of the First International Conference on Gameful Design, Research, and Applications, Gamification '13. New York: ACM. pp. 79–82.

Eveleigh A., Jennett C., Blandford A., Brohan P., and Cox A. L. 2014. Designing for dabblers and deterring drop-outs in citizen science. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14. New York: ACM. pp. 2985–2994.

Farell B. and Pelli D. G. 1999. Pscyhophysical methods, or how to measure a threshold and why. In *Vision research: A practical guide to laboratory methods*, edited by Carpenter R. and Robson J. New York: Oxford University Press.

Field A. 2009. *Discovering statistics using SPSS*. London: Sage Publications.

Freitag A., Meyer R., and Whiteman L. 2016. Strategies employed by citizen science programs to increase the credibility of their data. *Citizen Science: Theory and Practice* 1:2.

Gerhart B. 1987. How important are dispositional factors as determinants of job satisfaction? Implications for job design and other personnel programs. *Journal of Applied Psychology* 72:366–373.

Ghani J. A. and Deshpande S. P. 1994. Task characteristics and the experience of optimal flow in human–computer interaction. *Journal of Psychology* 128:381–391.

Hackman J. R. and Oldham G. R. 1975. Development of the job diagnostic survey. *Journal of Applied Psychology* 60:159–170.

Halkidi M., Batistakis Y., and Vazirgiannis M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* 17:107–145.

Hand E. 2010. Citizen science: people power. *Nature News* 466:685–687.

Hiesinger H., van der Bogert C. H., Pasckert J. H., Funcke L., Giacomini L., Ostrach L. R., and Robinson M. S. 2012. How old are young lunar craters? *Journal of Geophysical Research* 117:E00H10.

Huang C.-Y. 2002. Distributed manufacturing execution systems: A workflow perspective. *Journal of Intelligent Manufacturing* 13:485–497.

Hutt H., Everson R., Grant M., Love J., and Littlejohn G. 2013. How clumpy is my image? Evaluating crowdsourced annotation tasks. Presented at the 2013 13th UK Workshop on Computational Intelligence (UKCI), pp. 136–143. https://doi.org/10.1109/ukci.2013.6651298

Iacovides I., Jennett C., Cornish-Trestrail C., and Cox A. L. 2013. Do games attract or sustain engagement in citizen science?: A study of volunteer motivations. CHI '13 Extended Abstracts on Human Factors in Computing Systems. New York: ACM. pp. 1101–1106.

Jäkel F. and Wichmann F. A. 2006. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision* 6:13.

Kirchoff M., Sherman K., and Chapman C. 2011. Examining lunar impactor population evolution: Additional results from crater distributions on diverse terrains. Presented at the EPSC-DPS Joint Meeting 2011, p. 1587.

Kulkarni A., Can M., and Hartmann B. 2011. Turkomatic: Automatic, recursive task and workflow design for mechanical turk. Proceedings of the 11th AAAI Conference on Human Computation, AAAIWS'11-11. AAAI Press, pp. 91–96.

Kulkarni A., Can M., and Hartmann B. 2012. Collaboratively crowdsourcing workflows with turkomatic. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. New York: ACM. pp. 1003–1012.

Marshall P. J., Lintott C. J., and Fletcher L. N. 2015. Ideas for citizen science in astronomy. *Annual Review of Astronomy and Astrophysics* 53:247–278. https://doi.org/10.1146/annurev-astro-081913-035959.

McGill G. E. 1977. Craters as "fossils": The remote dating of planetary surface materials. *Geological Society of America Bulletin* 88:1102–1110.

Mekler E. D., Brühlmann F., Opwis K., and Tuch A. N. 2013. Disassembling gamification: The effects of points and meaning on user motivation and performance. CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13. New York: ACM. pp. 1137–1142.

Nov O., Arazy O., and Anderson D. 2011. Dusting for science: Motivation and participation of digital citizen science volunteers. Proceedings of the 2011 IConference, IConference '11. New York: ACM. pp. 68–74.

Pelli D. G. and Farell B. 2010. Psychophysical methods. In *Handbook of optics*, edited by Bass M. New York: McGraw-Hill, pp. 3.1–3.12.

Prather E. E., Cormier S., Wallace C. S., Lintott C., Raddick M. J., and Smith A. 2013. Measuring the conceptual understandings of citizen scientists participating in Zooniverse projects: A first approach. *Astronomy Education Review* 12(1):1–14.

Priebe J. 2009. A study of Internet users' cookie and javascript settings. *smorgasbork*. http://www.smorgasbork.com/2009/04/29/a-study-of-internet-users-cookie-and-javascript-settings/

Reed J., Rodriguez W., and Rickhoff A. 2012. A framework for defining and describing key design features of virtual citizen science projects. Proceedings of the 2012 IConference, IConference '12. New York: ACM. pp. 623–625.

Robbins S. J., Antonenko I., Kirchoff M. R., Chapman C. R., Fassett C. I., Herrick R. R., Singer K., Zanetti M., Lehan C., Huang D., and Gay P. L. 2014. The variability of crater identification among expert and community crater analysts. *Icarus* 234:109–131.

Schmidt M.-T. 1998. Building workflow business objects. In *Business object design and implementation II*, edited by Patel D. D., Sutherland D. J., and Miller J. London: Springer. pp. 64–76.

Sprinks J., Houghton R. J., Bamford S., and Morley J. G. 2015. Citizen scientists: The importance of being needed and not wasted. CHI Play '15 Workshop: The Annual Symposium on Computer-Human Interaction in Play. London.

Swanson A., Kosmala M., Lintott C., Simpson R., Smith A., and Packer C. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data* 2:150026.

Tar P. D. and Thacker N. A. 2015. Understanding and reducing crater counting errors. Presented at the Issues in Crater Studies and the Dating of Planetary Surfaces, p. 9027.

Tar P. D., Bugiolacchi R., Thacker N. A., and Gilmour J. D. 2016. Estimating false positive contamination in crater annotations from citizen science data. *Earth, Moon, and Planets* 119:47–63.

Taylor J., Teanby N. A., and Wookey J. 2013. Estimates of seismic activity in the Cerberus Fossae region of Mars. *Journal of Geophysical Research* 118:2570–2581.

Tinati R., Van Kleek M., Simperl E., Luczak-Rösch M., Simpson R., and Shadbolt N. 2015. Designing for citizen data analysis: A cross-sectional case study of a multi-domain citizen science platform. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15. New York: ACM. pp. 4069–4078.

Winnicki A. 2016. Just how many web users disable cookies or JavaScript? *Yell Blog*. https://blog.yell.com/2016/04/just-many-web-users-disable-cookies-javascript/