Publisher's DOI: 10.1016/j.ijhcs.2017.03.003

Task Workflow Design and its impact on performance and volunteers' subjective preference in Virtual Citizen Science

James Sprinks[a], Jessica Wardlaw[b], Robert Houghton[c], Steven Bamford[d], Jeremy Morley[ef]
[a,b & e]*Nottingham Geospatial Institute, University of Nottingham, Triumph Road, Nottingham, UK, NG7 2TU*
[c]*Human Factors Research Group, University of Nottingham, University Park, Nottingham, UK, NG7 2RD*
[d]*School of Physics & Astronomy, University of Nottingham, University Park, Nottingham, UK, NG7 2RD*
[f]*Ordnance Survey, Explorer House, Adanac Drive, Southampton, SO16 0AS*
[a]*james.sprinks@nottingham.ac.uk,* [b]*jessica.wardlaw@nottingham.ac.uk,* [c]*robert.houghton@nottingham.ac.uk,*
[d]*steven.bamford@nottingham.ac.uk,* [e]*jeremy.morley@os.uk*

ABSTRACT

Virtual citizen science platforms allow non-scientists to take part in scientific research across a range of disciplines. What they ask of volunteers varies considerably in terms of task type, variety, user judgement required and user freedom, which has received little direct investigation. A study was performed with the *Planet Four: Craters* project to investigate the effect of task workflow design on both volunteer experience and the scientific results they produce. Participants' feedback through questionnaire responses indicated a preference for interfaces providing greater autonomy and variety, with free-text responses suggesting that autonomy was the more important. This did not translate into improved performance however, with the most autonomous interface not resulting in significantly better performance in data volume, agreement or accuracy compared to other less autonomous interfaces. The interface with the least number of task types, variety and autonomy resulted in the greatest data coverage. Agreement, both between participants and with the expert equivalent, was significantly improved when the interface most directly afforded tasks that captured the required underlying data (i.e. crater position or diameter). The implications for the designers of virtual citizen science platforms is that they have a balancing act to perform, weighing up the importance of user satisfaction, the data needs of the science case and the resources that can be committed both in terms of time and data reduction.

Keywords: Citizen Science, engagement, task workflow, interface design

1. Introduction

Citizen science, also known as "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 crowdsourcing techniques. Extant citizen science projects require the participant to either act as a sensor and collect data, typically 'in the wild' with an array of mobile technologies, or analyse previously collected data through internet-based Virtual Citizen Science (VCS) platforms (Reed et al., 2012). Launched in 2009, the Zooniverse (www.zooniverse.org) is home to some of the internet's most popular VCS projects, which contribute to a wide range of research, with volunteers asked to, for example, classify different types of galaxies from photographs taken by telescopes (www.galaxyzoo.org), transcribe historical ships logs and weather readings (www.oldweather.org), or mark craters found on images of planetary surfaces (www.moonzoo.org).

As a relatively new form of activity, online citizen science research has tended to be driven by concerns around the core science rather than being considered as something that can be designed to suit its user population (with some exceptions, e.g., Prestopnik and Crowston, 2012). This is perhaps ironic given the importance of the 'citizen' to the endeavour, especially as the effectiveness of a citizen science venture is related to its ability to attract and retain engaged users, both to analyse the large amount of data required, and to ensure the quality of the data collected (Prather et al., 2013). Current VCS platforms tend to require the user to carry out tasks in a very repetitious manner, the design of which are arguably driven more by the 'science case' (analogous to a 'business case' in industry) rather than any consideration of the experience of the citizen scientist (Cox et al., 2015). In the study reported here we make a first step in considering how VCS platforms can be designed to better meet the needs of the citizen scientists by exploring whether the influence of manipulating task flow predicted with similar

systems would affect the rate and number of features indicated, as well as user ratings on difficulty and usability issues. We also investigate how these factors affect the (volunteered) data's volume and accuracy by comparing it with expert judgements.

Some studies have considered motivation amongst citizen science volunteers (Reed et al., 2013; Eveleigh et al., 2014), but not considered the form of work activity itself in any depth. This may be considered remiss since forty years of research have identified a relationship between motivation, satisfaction and work design (Hackman and Oldham, 1975; Oldham and Hackman, 2010) and in recent times has been directly applied to online crowdwork (Kittur et al., 2013). Factors such as task variety, complexity and autonomy were identified as important influences on motivation and productivity, all of which can be influenced by VCS design.

We begin with a review of relevant literature on the interplay between motivation, performance and task design in the areas of Citizen Science, work design and HCI. We then introduce Planet Four: Craters – a Zooniverse citizen science project that consists of three separate interfaces that vary in task workflow design (TWD) for the marking of craters on the surface of Mars, and present a laboratory study that directly compares participants' performance and experience across the three interfaces. Finally the impact of TWD on these results, and the implications for VCS platforms and other online mechanisms, are discussed.

2. Background

2.1. Citizen Science as a Distinct Form of Enquiry

Although VCS is a relatively new form of work, nascent research considers Citizen Science practices in their own right, beyond the scientific problems they address (Jordan et al., 2015). These studies have investigated aspects including, but not limited to: VCS typology and functionality (Prestopnik and Crowston, 2012; Reed et al., 2012); gamification (Deterding et al., 2011; Curtis, 2014; Eveleigh et al., 2013; Iacovides et al., 2013); volunteers' extrinsic motivation (Raddick et al., 2009; Reed et al., 2013; Mankowski et al., 2011; Mao et al., 2013); and volunteer behaviour (Ponciano et al., 2014; Crowston and Fagnot, 2008; Rotman et al., 2012; Nov et al., 2011). These studies, however, are predominantly concerned with the initial attraction of volunteers to a VCS platform and visceral aspects of their design, without consideration of their experience and performance in executing tasks i.e. the work that they do once they arrive, which are not easy to control. Although some recent research has considered the effect of task and judgement on volunteer performance (Hutt et al., 2013), and how they should be designed dependent on volunteer commitment (Eveleigh et al., 2014), no study to date has directly experimented with the manipulation of TWD elements to investigate their effect on volunteer behaviour, experience and scientific output. This represents an as yet missed opportunity, as TWD can be practically affected at the design stage of a project, and so it would be beneficial to understand its potential influence on the performance of citizen science.

Factors including volunteer engagement (Lahav et al., 1995), data volume (Lintott et al., 2011) and data accuracy (Hennon et al., 2014) are key to ensuring that citizen science endeavours process the large amount of data available to the standard required in order to add value to existing datasets, and as such are used to measure the success of a project. Several decades of human factors and work design research has revealed a connection between TWD factors and similar performance measures, and so the broader research question of this study is: can the lessons learnt regarding the effect of TWD on similar systems be applied to the citizen science case? If they can, whether completely or in part, it would suggest that TWD could be tailored at the design stage to improve the performance of a citizen science project. This could be achieved through an approach that practically is easier to implement compared to

the considerations of existing citizen science research, regarding the extrinsic motivation provided by the science theme addressed.

2.2. Relevant insights from perceptual psychology and the design of work

VCS platforms involve processes, mechanisms and methodologies that have historically been used in other similar systems, and as such there is a wealth of research regarding their design and implementation. For example, VCS platforms, in general, ask participants to carry out a task from a discrete set of different task types (Pelli and Farell, 2010): detection (is a stimulus present/identifiable?), discrimination (the difference between two stimuli) and matching (adjusting an attribute of two stimuli until they are equal). Such tasks force the observer to make corresponding judgements (Farell and Pelli, 1999), including yes/no (is something present or not), forced choice (pick the closest match the stimuli is to a selection of pre-defined examples) and rating scales (assess the magnitude of a certain attribute of the stimuli based on a given scale). Research on these different task types in the context of image analysis shows that they affect the performance and experience of the human actor. In one of the few studies that directly considers the citizen science case, Hutt et al., (2013) compared three approaches that generate 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.

Beyond the task types and judgements required of citizen scientists, there is also the question of how the user interface presents them. Current VCS systems often require participants to do the same task(s) repetitively over a seemingly never-ending number of images, in an almost ‘data entry’ like manner, for no financial reward. This scenario is analogous to that found in the 1960s concerning the mechanisms of industrial work, including the fractionation and atomisation of tasks, the most well known being found on car production lines. In response to this, Hackman & Oldham (1975) developed the ‘Job Diagnostic Survey’ in order to better understand jobs and how they could be re-designed to improve motivation and productivity. Factors such as task variety, complexity and autonomy were identified as key to this process, all of which can be influenced in VCS design. Building on these findings, further research has found a positive correlation between motivation and task complexity (Gerhart, 1987; Chung-Yan, 2010), task autonomy (Dubinsky and Skinner, 1984; Chung-Yan, 2010) and variety (Ghani and Deshpande, 1994; Dubinsky and Skinner, 1984). Although the main body of this research concerns work over an extended period of time, which may or may not be true of volunteers regarding a citizen science platform (Eveleigh et al., 2014), the ideas act as the inspiration for this work as a form of design choice that could be applied to the VCS case.

2.3. Task Workflow Design

The concept of task workflow design is the core construct of this study. 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 & Oldham, and as such could influence motivation and performance. Whilst originally a concept associated with the manufacturing and business industries (Huang, 2002; Schmidt, 1998), the notion has been extended to forms of crowd sourced work due to the analogy that can be made between them. Predominantly 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. As previously mentioned, existing research regarding the TWD of virtual citizen science platforms has tended towards a retrospective approach, studying the design of existing platforms and their performance in terms of volunteer engagement and data collection (Tinati et al., 2015; Hutt et al., 2013; Eveleigh et al., 2014) and making recommendations and design claims based on the findings.

With the literature regarding task workflow design and perceptual psychology in mind, this paper sets out to explore TWD and its factors in the context of citizen science. In the next section we introduce a system overview of the Zooniverse site *Planet Four: Craters*, followed by the methodology of its use to directly manipulate TWD and explore how task type, variety, and autonomy can affect volunteer preference, experience, and performance. Inspired by the related Human Factors work summarised in the previous section, testable hypotheses were formulated for a mix of qualitative and quantitative dependant measures relating to volunteer behaviour, experience and performance:

H1: Volunteers using an interface with greater autonomy produce a greater volume of more accurate data.
H2: Volunteers prefer using an interface involving a greater variety of task types.
H3: Volunteers performing a task workflow with fewer task types produce greater data volume.

3. System Overview of Planet Four: Craters

Developed in 2013, Planet Four: Craters was created to address two separate goals: 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. An established method for ageing the surface of a planet is through analysis of the size and density of craters on its surface from meteorite impacts; the theory goes that smaller meteorites collide with a planet much more frequently than larger ones, and older surfaces have more craters because they have been exposed for longer.

3.1. Crater Marking Tools

Before explaining the design of the three interfaces, and how they present the crater marking task to the participant, this section will briefly describe the different tools that have been developed for participants to mark craters, and the types of task and judgement they involve.

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'). In essence, this tool facilitates a detection task through making a forced choice (yes/no) judgement.

Crater Position tool: This tool allows users to mark the centre of each crater in the image by positioning the cursor pointer and performing 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 centre of the crater) through making a matching judgement.

Crater tool: This tool allows users to mark a circle around the edge of each crater, by positioning the cursor in the crater centre, then clicking 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 centre and edge) by means of matching judgements.

3.2. 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, judgement, task variety and autonomy. Figure 1 shows the three variations of the interface: full, Batched and Sequenced, while figure 2 describes the order in which tasks are presented to the participant for each interface.

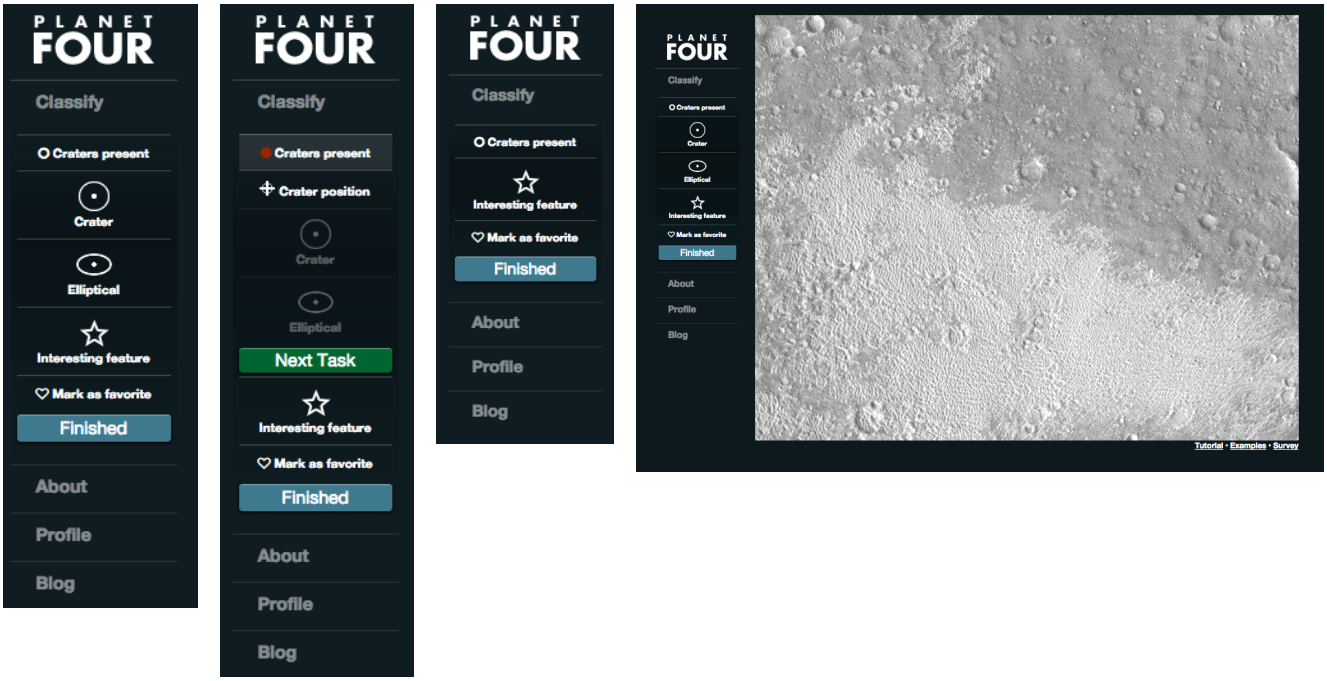

Fig 1. Planet Four: Craters interface designs. Left to right - Full interface where all tools are available, Sequenced interface where tools are used in turn activated by the 'Next Task' button and Batched interface where only one tool is used for each image. The tool interface appeared to the left of the image being analysed, as shown in the full screenshot (far right).

Sequenced: The Sequenced 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, 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 judgements they require.

Batched: The Batched interface is the simplest of the three, with the participant only using one tool per image. After completing a set number of images (15 in the case of our laboratory study) the tool changes i.e. the participant presses/depresses the 'craters present' button for each image in turn, then marks the centre 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 participant does not return to the same image twice, but does each new task on an entirely new and unseen batch of imagery.

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 have 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.

SEQUENCED: All tasks completed in a predefined order (increasing in complexity) for each image, that cannot be diverged from.

BATCHED: Participants have access to one tool and complete one task for a set number of images (*N*), then use another tool and complete another task for the next set of different images etc.

FULL: Participants have access to all the tools and can complete all crater marking tasks for each image in any chosen order.

Fig 2. Flow diagram of tools available to the user for each interface

4. Methodology

4.1. Experimental Design

This study aimed to investigate the effect of manipulating TWD on volunteer preference, experience and performance, when carrying out the crater marking task on the Planet Four: Craters project. Using a within-subjects design, the TWD factors autonomy, variety, task type and volunteer judgement were manipulated. Three separate classification interfaces that varied in relation to these factors were employed, in conjunction with a questionnaire including NASA Task Load Index (TLX) type statements to assess volunteer opinion and perceived workload. The NASA-TLX framework has been used in previous research to measure workload regarding a number of on-screen and HCI type tasks (Harrower and Sheesley, 2005). In essence, it is a standardised framework where participants rate their perception of a task's workload by indicating the contribution of six factors – temporal demand, mental demand, physical demand, effort, frustration and performance. Additionally a text box allowed participants to explain their answers in order to add context to the findings.

The four TWD factors (independent variables 1 to 4) were experimentally manipulated through the design of the three Planet Four: Craters interfaces as described in table 1, and the impact of this manipulation measured through participants' self-reports giving their opinions on preference, engagement and experience (dependent variable 1), and by measuring performance through participant-expert marking comparison (DV2), the number of markings made (DV3) and the time spent classifying each image (DV4).

Table 1. Task Workflow Design configuration of each Planet Four: Craters interface

Interface	Autonomy	Variety	Tasks	Task Type(s)	Judgements
Sequenced	Set order (Least autonomy)	All tasks (Most variety)	Do in order:		
			Is crater present	Detection	Yes/no
			Mark position	Detection & Matching	Matching
			Mark size	Detection & 2xMatching	Matching
Batched	Set order (Least autonomy)	Single task (Least variety)	Either:		
			Is crater present	Detection	Yes/no
			Mark position or	Detection & Matching	Matching
			Mark size	Detection & 2xMatching	Matching
Full	Any order (Most autonomy)	All tasks (Most variety)	Pick from:		
			Is crater present	Detection	Yes/no
			Mark position	Detection & Matching	Matching
			Mark size	Detection & 2xMatching	Matching

The task type and judgement classifications for each interface have been adapted from Pelli & Farell's (2010; 1999) work regarding psychophysical methods, as follows:

Task Types:

- **Detection:** The goal of a detection task is to determine the existence of a stimulus, i.e. it is detectable by the observer above the background 'noise'. In this case, detecting a crater on the Martian surface.
- **Matching:** The observer has to adjust a stimulus along one or more physical dimensions until it matches another stimulus in terms of some perceptual attribute. In this case, matching an annotation (either a point for the crater centre or a circle for the crater outline) with the visible centre or edge of the crater in the image.

Associated Participant Judgements:

- **Yes/No:** Usually used for detection tasks, the observer has to judge whether the stimulus was present, or classify the percept "did you see it?" – The observer only has two response options, yes or no. For instance, is a crater present?
- **Matching:** Two stimuli are presented, and the observer has to adjust and ultimately judge when one exactly matches the other. The level to which this can be a 'perfect match' can vary, for instance it might be dependent on the amount of contrast between the crater and its surroundings, which in turn may be related to its age/erosion.

As can be seen in table 1 there are 3 experimental conditions represented by the interfaces described, but 4 main constructs under investigation. The reason for this lies in the interplay between task workflow design factors (Dodd and Ganster, 1996), meaning that in practical terms one cannot be manipulated without altering another. For instance, if an interface is designed to restrict the variety of task available (for instance the batched interface), this also means that autonomy must also be restricted, as the

participant does not have 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' judgement – i.e. can they detect the stimulus or not?

4.2. Materials

For the study, participants analysed an image taken by the Context camera on NASA's Mars Reconnaissance Orbiter[1]. It was chosen because it contains a variety of landscapes common to the Martian surface; scientists at the University of Bristol also provided data from their existing analysis of this image, that we used in place of ground-truthing so that comparisons could be made between citizen scientist results and those measured by planetary science experts. Before being uploaded to the platform, the image was 'sliced' into a number of smaller images that can be more easily handled. Original NASA imagery is often gigabytes in size, making it time-consuming to render to a web browser. A total of 78 smaller image 'slices' were created, measuring 840 x 648 pixels with an included overlap of 100 pixels to ensure features on the edges were adequately displayed.

4.3. Questionnaire Design

To obtain participant views and opinions, each participant completed a questionnaire after using each interface. The questionnaire contained Likert-type statements and 'free-text' responses concerning the *design & usability*, *tasks & tools* and *imagery*. Prior to these sections, participants completed a general demographics section regarding their background and experience.

The *design & usability* section of the questionnaire was made up of three statements regarding the general design, usability and appearance of the site as a whole. This was intended to determine the opinions of participants regarding general website navigation, logic and general ease of use separate from the more specific scientific task required. A free-text box was provided after the statements to allow participants to add any extra thoughts not covered by the statements and to add context to their responses.

The *tasks & tools* section consisted of 9 statements more directly concerned with the specific tools used to mark craters (as described in section 3.1.1) and the tasks required of the participant. This section intended to determine the participants' opinions regarding the task workflow design of the platform, and the suitability of the tools provided. Again a free-text box was provided allowing participants to add detail and context. Of the 9 statements, 6 of them were variations of the NASA-TLX design, revised to be more specific to the crater-marking task. Although NASA-TLX is a calibrated assessment tool and therefore should not normally be amended, the reason for this variation is explained through how the questionnaire was derived. This study has been developed in agreement with a number of other parties, including members of the Zooniverse development team and Planet Four science team. As such the questionnaire design required agreement across those involved, and it was decided due to likely future deployments involving an existing citizen science community, that the statements should be related as closely as possible to the specific citizen science platform to avoid confusion or misunderstanding.

The *imagery* section contained 4 statements concerned with both the quality and content of the images displayed, intended to determine the degree to which these factors assisted or hindered the task of identifying marking craters on the surface of Mars. As with the other two sections, a free-text box was provided to allow participants to contribute additional thoughts.

4.4. Participants

[1] Image ID G05_020119_1895_XN_09N198W, available from the NASA Planetary Data System: http://pds-imaging.jpl.nasa.gov/search/

30 participants (19 male, 11 female) were recruited through email lists, social media posts and subsequent 'word of mouth'. In terms of age, they were between 22 and 60 years old (mean = 28, median = 26). Eight of the participants had previously taken part in the original *Planet Four* project, while 10 participants had never heard of the site. The remaining 12 were aware of Planet Four but had never taken part. Regarding experience with the Zooniverse, only four participants had visited other sites and they were predominantly of a similar space theme (Galaxy Zoo (4), Planet Hunters (1), and Ancient Lives (1)). Participants were gifted a £10 (around $15) Amazon voucher for their participation in the study. All participants have been educated to a university-degree level, however none have had any formal training directly relating to planetary science. As such, this is representative of the education and experience regarding existing citizen science volunteer communities (Raddick et al., 2013).

4.5. Procedure

All study participants came to the same room (individually) and carried out the experiment on the same laptop, to keep factors such as lighting conditions and screen setup constant and ensure that they did not influence the image analysis task. Before using each interface, each participant completed an online tutorial to learn how to use the tools, marking craters on a separate example image. Participants then used each of the interfaces in turn to mark craters on a set number of the image slices (15 on each interface); to mitigate bias caused by learning of the system, the order in which the interfaces were presented was manipulated so that the same number of participants tested the interfaces in the same order. The order in which image slices were displayed to each participant was also randomised, to prevent bias being caused by image content (images with few or no craters appearing in the same interface each time etc.). After using each interface, participants completed the questionnaire to share their views as previously described. There was no time limit to complete the task, participants were allowed as long as needed to complete the requisite number of image slices. Participants spent an average time of 9m19secs ± 1m39secs using each interface, and 1m47secs ± 32secs on the tutorial image.

5. Results and Analysis

Dependent variable measures were recorded both through participant self-reporting and, more behaviourally, through crater marking performance. Regarding participant crater markings, they have not only been evaluated in terms of their abundance but also their agreement, both with the expert equivalent and with markings of the same crater made by other participants. In the absence of an absolute ground truth, current citizen science projects predominately use two ways of validating the data collected. When available, 'gold standard' data created by the expert scientific community is used to compare with the volunteer data (Swanson et al., 2015). However, due to the abundance of data that requires analysis (hence the need for a citizen science solution in the first instance) there is often only a small sample of expert data available for comparison, and therefore in its absence participant agreement is used as a measure of certainty (Freitag et al., 2016). By considering these two measures separately in this study's analysis, it is ensured that any findings regarding task workflow design are applicable to both approaches used by the wider citizen science community. The following section presents the results and analysis for each method in terms of their relation to the independent variables regarding TWD.

5.1. Questionnaire Results

Participant Likert responses to a number of statements included in the *design & usability, tasks & tools* and *imagery* sections of the questionnaire showed no statistically significant difference between each interface. This is perhaps to be expected since many of the design features, tasks performed and tools used along with the image format are constant throughout the experiment. When considering the visual

appeal of the site and image quality, participant responses tended to be positive, and again showed no significant difference between each interface. Similarly, over 70% of all scores participants gave to the NASA-TLX statements relating to perceived task workload fell between 1 and 4 (low demand), with no statistically significant difference existing between each interface.

However, differences between the interfaces emerged in participants' scores for how quickly they learnt to use them. A repeated measures ANOVA with a Greenhouse-Geisser correction shows a statistically significant difference in participant scores between interfaces (F (1.297, 37.621) = 6.232, p = .011), and post hoc tests using the least significant difference correction revealed that participants felt they learnt the Full interface more quickly than the Sequenced (mean score of 7.90 ± 1.54 vs. 6.87 ± 2.06, p = .001). There was also a statistically significant difference in the scores participants gave for how easy they found each interface to use (F (1.817, 52.684) = 4.957, p = .013); the Full interface was rated easier to use than the Sequenced (mean score of 6.93 ± 1.68 vs. 6.07 ± 2.18, p = .011) and slightly easier than the Batched (6.70 ± 1.80), although the difference is not statistically significant (p = .118). Participants also scored the Batched interface easier to use than the Sequenced (p = .039), and participant responses suggested that the Sequenced interface is the least easy to use and access of all three.

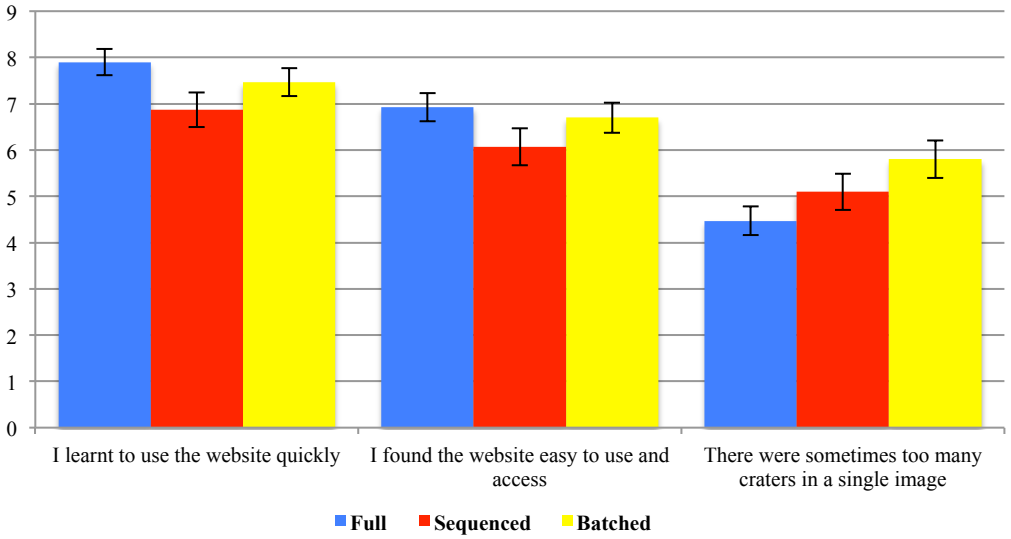

Fig 3. Response differences per interface (with standard error shown)

Concerning the crater count for each image, participants' responses differed across the interfaces (F(1.811, 52.507) = 5.184, p = .011). Participants felt that there were sometimes too many craters in an image when marking them with the Batched interface compared to the Full (mean score of 5.80 ± 2.20 vs. 4.47 ± 2.05, p = .004) and slightly more than the Sequenced (5.10 ± 2.16), although the difference is not statistically significant (p = .053). Figure 3 shows the average level of agreement between participants to these statements for each interface, along with standard error.

In addition to their responses to the Likert-type statements, several participants also provided 'free response' replies that add context to their scores. The responses for all three parts of the questionnaire were collated for each interface and subsequently coded under four categories based on their content: *usability, accuracy, tool issues* and *imagery. Usability* comments were concerned with the general usability and mechanics of each interface; *accuracy* comments focussed on how accurately craters could be marked; *tool issue* comments specifically concerned the tools provided to mark craters; and *imagery* comments discussed the remotely sensed imagery displayed. Table 2 shows a breakdown of how many comments participants made for each interface within these four categories. As can be seen,

the frequency of comments across each of the interfaces is similar apart from the Sequenced regarding *usability* comments, where the figure is much higher compared to the others.

Table 2. Number of responses by category and interface

Topic	Full Interface	Sequenced Interface	Batched Interface
Usability	3	9	4
Accuracy	6	5	7
Tool issues	7	6	6
Imagery	3	2	2
Total:	19	22	19

Comments in the *usability* category predominantly concerned the autonomy allowed by each interface, in terms of the tools available and the order in which they could or could not be used (participant P19):

> *"I don't like to be forced to use a certain task order, and I couldn't go back or switch tools…"*

Comments in the *accuracy* category shared a theme regarding issues with marking smaller craters, and the difficulties that arise (P4):

> *"Small craters were difficult to mark, and it was hard to decide if they were craters at all…"*

A number of the comments in the *tool issues* category again mentioned the marking of smaller craters, this time directly attributing the problem to tool use and design (P19):

> *"The dots of the tools were too big for small craters - annoying. This was in all three conditions."*

Other *tool issues* comments indicated a 'zoom' tool would help users to navigate around the image, presumably to aid the marking of smaller, less clear craters (P11):

> *"A zoom function would be nice…"*

Finally, comments on the *imagery* related to its content, i.e. the number of craters found in an image (S28):

> *"Some images contain many small craters which were hard to mark, it was hard to distinguish what was a crater and what was another land feature."*

More general comments pertained to its quality (P7):

> *"Too bad the quality is low…"*

Other participant comments that did not fall into the categories described were either general praise of the overall appearance of the website or requests for more information about the overall scientific goal, and were common across each of the interfaces used.

5.2. Crater Marking Results

Participant crater marking behaviour has been compared across each interface in terms of percentage of participants who marked craters per image, number of markings per image and time spent on each image. As explained in the experiment design section, the Batched interface requires participants to use one tool across a number of images, and then another tool etc. Batched (Position) therefore represents results where participants only mark the centre of craters and Batched (Mark) represents results where participants mark the shape. The comparatively large values of standard deviation can be explained by image variation, with some images containing no craters and others having several dozen (a common occurrence for VCS platforms involving planetary data).

A Friedman test revealed a statistically significant difference in the percentage of participants that marked at least one crater per image between the interfaces (X^2 (3) = 6.1, p = .05). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that a greater percentage of participants marked at least one crater per image with the Batched (Position) interface compared to the Sequenced (63.24 ± 33.76% vs. 53.67 ± 33.21% of users marking craters per image) although the difference is not statistically significant (p = .098). The percentage of participants that marked craters per image for the Full and Batched (Mark) are 58.24 ± 35.45% and 57.94 ± 35.02% respectively.

A repeated measures ANOVA test with a Greenhouse-Geisser correction also showed a statistically significant difference in the number of crater markings per image (F (2.656, 201.83) = 7.416, p = .0005). Post hoc tests using the Bonferroni correction revealed that the Batched (Position) interface resulted in a greater number of markings (3.61 ± 4.67) compared to the Full (2.46 ± 2.93, p = .001), Sequenced (2.55 ± 4.17, p = .003) and Batched (Mark) (2.24 ± 2.85, p = .001) interfaces. Finally when considering the amount of time spent on each image, a statistically significant difference again exists across interfaces (F (1.570, 119.290) = 12.755, p = .0005). Participants spent more time using the Sequenced interface compared to the Batched interfaces (55 ± 64 seconds vs. 28 ± 16 seconds, p = .001) and more time compared to the Full (37 ± 27 seconds, p = .01).

In summary, the Batched (Position) interface resulted in less null images returned (images with no craters marked) than other interfaces and a significantly greater number of craters markings per image. Participants using the Sequenced interface spent significantly more time classifying each image.

5.3. Participant Agreement

To assess the agreement between participants regarding the crater markings made, crater-marking clusters have been identified, defined as the combination of markings made by 2 or more participants of the same crater.

Although participants using the Full and Sequenced interfaces identified a similar number of crater-marking clusters (182 and 185 respectively), more crater-marking clusters were marked on the Batched (Position) interface (298, ~61% greater). Conversely, the Batched (Mark) interface has resulted in slightly fewer (163) and overall the results tally well with the average number of markings per image data described in the previous section.

Of the total number of crater-marking clusters marked, 86 were identified across all four interfaces, and so can be compared like-for-like in terms of participant agreement. They covered a range of sizes from a few to several hundred pixels (10s to 100s of metres) in diameter. A repeated measures ANOVA test with a Greenhouse-Geisser correction showed a significant difference between each interface (F (2.616, 222.36) = 4.863, p = .004) when considering agreement in crater position. Post hoc tests using the Bonferroni correction revealed that crater position markings made using the Sequenced interface varied significantly less (greater agreement) than those made using the Full and Batched (Mark) (standard deviation of 2.00 ± 0.90 pixels vs. 2.53 ± 1.15 pixels, p = .001 and vs. 2.57 ± 1.51

pixels, $p = .002$ respectively). Likewise, position markings made using the Batched (Position) interface showed significantly greater agreement than those made using the Full (2.20 ± 0.91 pixels vs. 2.53 ± 1.15 pixels, $p = .033$) and greater agreement than those made using the Batched (Mark) though the difference is not significant ($p = .065$). These results are illustrated by Figure 4, which shows the mean standard deviation of crater positions and diameters.

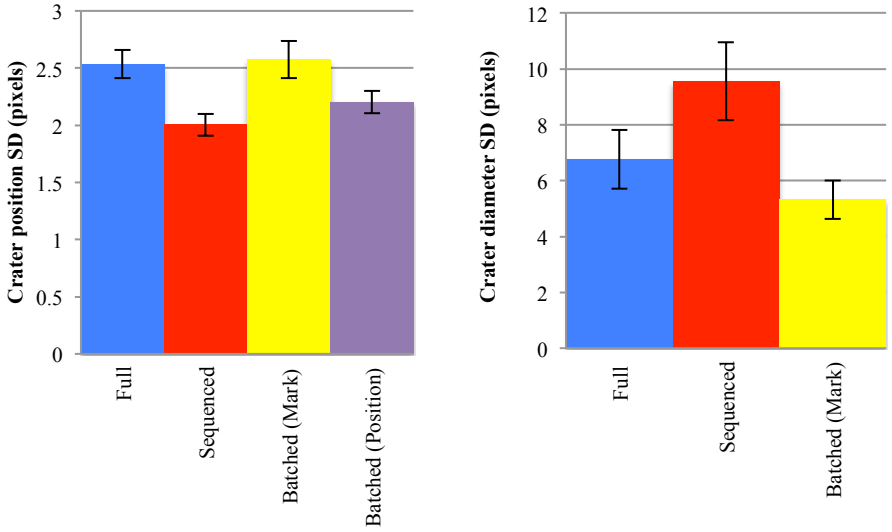

Fig 4. Comparison of marking agreement per interface

The agreement in crater diameter markings was, again, significantly different between each interface ($F(1.580, 142.17) = 19.199$, $p = .0005$). The diameter of markings made using the Sequenced interface (standard deviation of 9.55 ± 13.26 pixels) showed significantly less agreement than those made using both the Full (6.77 ± 10.10 pixels, $p = .001$) and Batched (Mark) (5.32 ± 6.57 pixels, $p = .001$). The diameter of markings made using the Batched (Mark) interface also showed significantly more agreement than those using the Full ($p = .007$).

In summary, the Batched (Position) interface resulted in the most crater clusters being identified (continuing the trend found in the previous section). Participants using the Batched (Position) and Sequenced interfaces had greater agreement regarding crater position compared to the Full and Batched (Mark) interfaces, with the Sequenced interface having the greatest agreement of all. In terms of crater diameter, the Batched (Mark) interface resulted in the greatest participant agreement, whilst the Sequenced interface resulted in the least.

5.4. Participant – Expert Comparison

An advantage of developing a purpose-built science case and having a dedicated science team is the access to previous research and expertise this affords. Experts from the University of Bristol identified 365 separate craters on the CTX image used in this study, 280 of which have been identified as a crater-marking cluster (clusters of participant markings identifying the same crater) on at least one of the interfaces. Table 3 compares the expert data with the crater marking clusters made using each interface regarding crater identification.

Table 3. Crater identification compared to expert, including measures of precision (fraction of participant marked craters that were confirmed by the experts) and recall (fraction of craters identified by experts also returned by participants).

Interface	No. of Craters Marked	No. of True Positives	No. of False Positives	No. of False Negatives	Precision	Recall
Full	182	168	14	197	92.3%	46.0%
Sequenced	185	170	15	195	91.9%	46.6%
Batched (Mark)	163	148	15	217	90.8%	40.5%
Batched (Position)	297	247	50	118	83.2%	67.7%

Using the expert data in the absence of ground truth, the performance of participants using the Full and Sequenced interfaces was very similar, with a comparable number of identified clusters also confirmed by the expert (precision ~92%). Participants also identified a similar number of crater marking clusters erroneously i.e. false positives (~8% of clusters) and identified a correspondingly similar number of craters from the full expert catalogue (recall rate ~46%). The fewer crater-marking clusters participants identified using the Batched (Mark) interface was matched by a lower number confirmed by experts (precision ~90%) and a greater number of craters identified by experts but missed by participants (recall rate ~40%). Finally, although a greater total number of clusters marked on the Batched (Position) interface were confirmed by experts, more crater-marking clusters were misidentified as a proportion of the total (precision ~83%). Participants, however, missed fewer of the experts' markings when a greater number of clusters were made (recall ~68%).

Out of the 365 expert crater markings made, 84 were subsequently correctly identified as a crater-marking cluster on all four of the interfaces, and were thus used to directly compare participants' markings to the experts' equivalent.

A repeated measures ANOVA test with a Greenhouse-Geisser correction showed a significant difference (F (2.619, 217.379) = 2.075, p = .05) in the average difference between participants' and experts' crater positions between each interface. Post hoc tests using the Bonferroni correction revealed that the markings made using the Batched (Position) interface were significantly closer to the expert equivalent than those made using the Full and Batched (Mark) interfaces (average difference of 3.99 ± 1.33 pixels vs. 4.44 ± 1.55, p = .036 and vs. 4.41 ± 1.75, p = .040 respectively). Markings made using the Sequenced interface (4.21 ± 1.13) were also closer in position to experts' than the Full and Batched (Mark) interfaces, though the difference was not significant.

A repeated measures ANOVA again showed a significant difference in the average difference between participant crater diameter and the expert equivalent between each interface (F (1.499, 130.415) = 5.439, p = .011). Post hoc tests with the Bonferroni correction revealed that the diameter of crater markings made using the Sequenced interface were more significantly different to the expert equivalent compared to those made using both the Full and Batched (Mark) interfaces (average difference of 8.80 ± 5.89 pixels vs. 6.49 ± 5.93, p = .001 and vs. 6.59 ± 10.55, p = .028 respectively). Figure 5 shows the average difference in position and diameter between the expert markings and those made by participants using each interface. Although the differences in both crater position and diameter across each interface may seem small (a few pixels at most) when considering both inter-participant agreement and expert comparison, they could be important due to the resolution of the imagery involved. The Context camera imagery used typically has a resolution of ~6m per pixel, and so even a sub-pixel difference can be meaningful when considering the scientific application of aging the surface, where craters as small as 10s of metres in diameter are included.

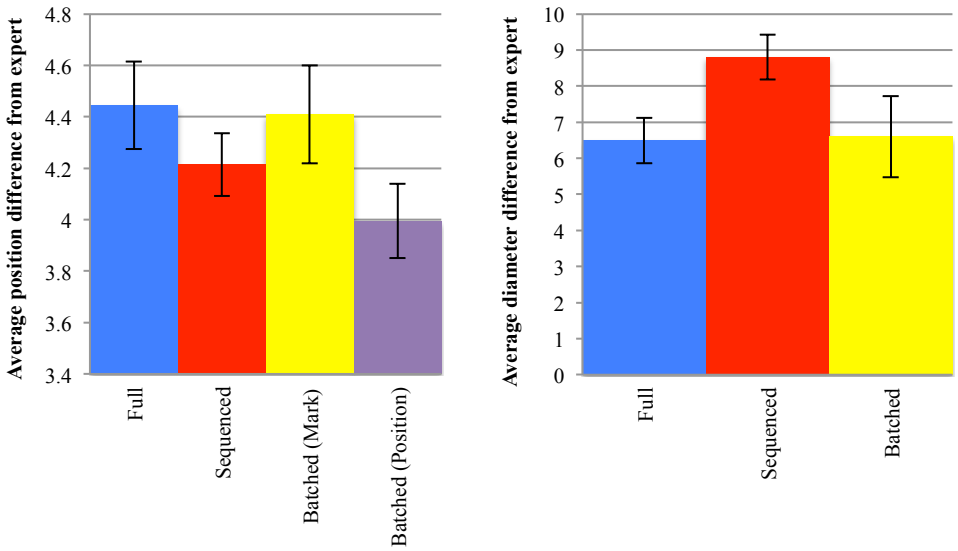

Fig 5. Position and diameter difference of participant markings compared to expert

Finally, Figure 6 shows the number of participants that contributed to a cluster versus the percentage that were true positives when compared with the expert data i.e. the number of participant markings required before a cluster can definitely be considered to represent a crater. Clusters identified by participants using the Batched (Mark) interface required the least amount of markings, with all of the clusters of four or more participant contributions also recognised by an expert equivalent. This was followed closely by those made on the Full interface where five or more participant markings were needed. Of the clusters identified with the Sequenced interface, those made up of seven or more participant contributions were also recognised by experts as a crater. The Batched (Position) interface required the most (eight or more) participant markings for clusters to be in 100% agreement with experts, twice as many markings than required for the Batched (Mark) interface.

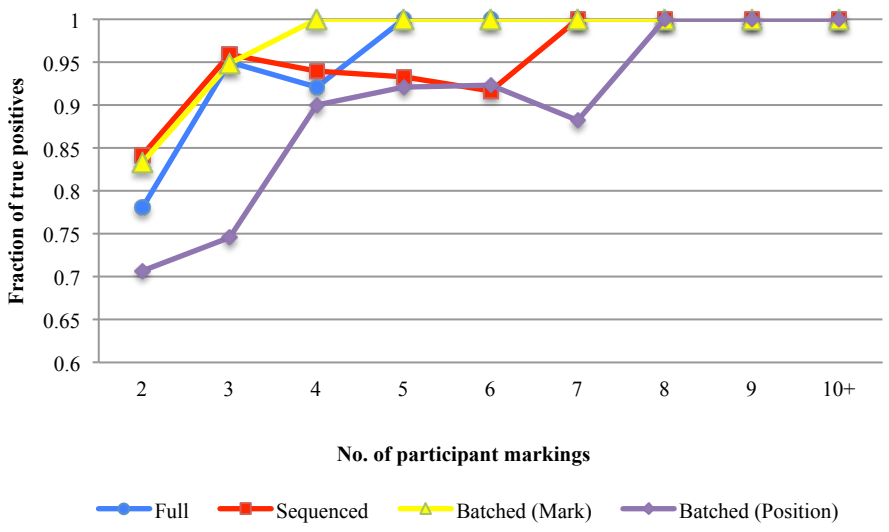

Fig 6. Cluster participant contribution verses fraction of true positives

It should be noted that while the percentage of true positives rises steadily with the number of participant markings per cluster made on the Batched (Mark) interface, there is a slight dip of ~2%

before rising to 100% of true positives as the number of participant markings increases when considering clusters on the other three interfaces. This suggests such crater misidentifications are not a mistake, but that there are artefacts or features on the image besides craters that appear to be craters, at least in the eyes of the majority of participants.

In summary, the Batched (Position) interface resulted in the highest proportion of expert-marked craters being identified by participants, but also resulted in a higher rate of false-positive markings. The Batched (Position) interface has also resulted in the closest agreement with the expert equivalent regarding crater position, whilst participants using the Sequenced interface had closer positional agreement compared to both the Full and Batched (Mark) interfaces. Regarding crater diameter, participants using the Sequenced interface had significantly the least agreement with the expert equivalent measurement. Finally, regarding the number of participant contributions required for a crater-marking cluster to be a certain true-positive (compared to the expert equivalent), the Batched (Marked) interface required the fewest contributions, followed in ascending order by the Full, Sequenced and finally Batched (Position) interface.

5.5. Summary and Hypotheses

In summary, the hypotheses related to task type and volunteer judgement were supported by the analyses, whereas the hypotheses related to autonomy and variety were only supported in part:

- Hypothesis H1 (greater autonomy = greater data volume & accuracy) is not supported by the findings regarding measures of performance, with the Full interface (affording the greatest autonomy) not resulting in any significant improvement.
- Hypothesis H2 (preference for greater variety) is supported by the self-reported findings of the questionnaire that the number of craters was significantly more of an issue when using the single-task Batched interface, however participants' free responses suggested autonomy was more important.
- Hypothesis H3 (fewer task types = greater data volume) is supported by the finding that more crater clusters were marked using the Batched (Position) interface.

In the following section this disparity is unpacked through discussing both the self-reported and behavioural findings of the work.

6. Discussion

This section combines the self-reported quantitative and qualitative findings of the questionnaire with the quantitative measures of participant agreement and performance to paint a broader picture of how TWD factors can affect a VCS platform's output and the preference of its volunteers. Participants' quotes are used to highlight their issues and concerns and are related back to their behaviour and performance.

6.1. The Effect of TWD on Participant Preference

Participant responses to the NASA-TLX type statements that concerned the workload of the crater-marking task showed no statistically significant difference across each interface. This was also the case with statements regarding the 'ease of use' of the tools and the imagery presented. This is perhaps unsurprising as the core task of identifying and marking a crater and the imagery presented were constant across the study. This view is supported by the 'free response' replies of the participants, with a similar number of comments made on these topics and raising the same concerns independent of

interface. For instance, comments on accuracy and the marking tools overwhelmingly concerned the difficulty of marking small craters; an issue raised across all study conditions.

Participant responses varied, however, regarding the number of craters in an image. Participants using the Batched interface showed a significantly greater agreement with the statement that there were too many craters to mark, suggesting an issue with the repetitiveness associated with the prescription of a single task. This is supported by a number of participant quotes, for instance:

P11: *"Sometimes I was wondering if there is a size threshold on craters. Some images contain many small craters that I couldn't be bothered to mark*."

P17: *"Different tasks made it less boring to interact with. Rather than having the same task throughout…"*

This demonstrates the impact of the task's variety on volunteers' intrinsic motivation to take part (Dubinsky and Skinner, 1984).

The greatest concern of participants, however, according to their questionnaire responses, was the autonomy they had for carrying out the task. Likert-type responses showed that participants found the Sequenced interface significantly less easy to use and access, and the reasons for this were explained by their 'free' responses:

P12: *"Least usable option: tools that were disabled shouldn't be visible. Sequence of work steps was not great too - having to revisit craters is not fun."*

P18: *"I don't like to get stuck in a linear set of actions when identifying craters…"*

P11: *"I do NOT like the "sequenced" format. It feels like I'm working with fragments. It also gives me the feeling (real) that I'm being condescended to (real? illusion?) - as if the tasks are being meted out in small dollops that my poor 'citizen non-scientist' brain might be able to manage…"*

P28: *"Being able to use the tools whenever - not in a specific order, was certainly far superior…"*

This suggested that participants were not in favour of being forced to complete the tasks in a set order, as they did with the Sequenced interface, especially when they were aware of other tasks that needed to be done. Furthermore, the number of comments on this issue suggests that autonomy could have a larger effect on volunteers' satisfaction than task variety for VCS projects, as was found by previous research (Chung-Yan, 2010), which could impact upon their intrinsic motivation to return to the project.

6.2. Considering Autonomy in Citizen Science Interface Design

Regarding hypothesis H1, linking an increase in autonomy with better task performance in terms of data volume and accuracy, the results do not support this position. Participants using the Full interface, allowing access to complete any task in any order, did not perform better in terms of the amount of craters identified, or in measurement agreement either between participants or with the expert equivalent. However, the results did provide evidence of a similar effect, with the interface allowing the least autonomy (Sequenced) reducing performance in terms of the time spent on each image. Participants spent significantly longer analysing each image, without producing a greater number of crater identifications. Although the effect of the extra time taken per image is minimised in the

laboratory setting of this study (as all participants classified the same number of imagery on each interface) it could have a negative effect in a 'live' environment. Since the time that volunteers can donate to a VCS platform is finite, the extra time taken per image could very well translate into less images being classified overall.

Considering participants' free-text responses to the questionnaire, responses suggested a preference for greater autonomy, or more directly described a frustration with being restricted. Firstly, this frustration is associated with being forced to return to the same crater to do a set order of tasks using the Sequenced interface, alluding to a feeling that their time is being wasted. Secondly, and perhaps of more concern, relates to a feeling of distrust – with volunteers indicating a feeling that they are only being given tasks one at a time in a controlled manner as the experts do not think they can be trusted to do more. This suggests that while there is no direct evidence that greater autonomy increases performance, it certainly plays a role in ensuring the work is satisfying to do, supporting Hackman & Oldham's (1975) inclusion of the construct in their JCT work.

6.3. The Effect of Task Variety

Although less frequent compared to those related to autonomy, participant questionnaire free-text responses support hypothesis H2 (preference for greater variety). A number of comments expressed a belief that having a number of different tasks made the platform more interesting to interact with, when compared to repeating the same task using the Batched interface. Additionally, several comments mentioned issues regarding the number of craters within a single image when using the Batched interface, losing motivation to mark large numbers of them. Although not directly mentioning task variety, it could be argued that giving the participant different tasks to perform on images with many craters could help mitigate this issue.

Although participant responses indicated a preference towards greater task variety, this did not necessarily translate into an improvement in crater-marking performance. For instance, analysis showed that participants using the Batched (Position) interface, where only one task is available (least task variety), produced the greatest number of crater classifications per image, and greater agreement in terms of the number of participants who identified and marked the same crater. Whether this improvement is related to participants being able to concentrate on one single task, or more attributed to the type of task and judgement required (hypothesis H3, described in 5.4) is unclear, highlighting the complexities that exist regarding the interplay of task workflow design factors. What is clear however, is that virtual citizen science practitioners should be aware that the preferred task workflow design configurations of their citizen scientist community will not automatically produce the best task performance for their science case, depending on the needs of the project.

6.4. The Types of Task and Judgements Required

As previously mentioned, analysis of crater marking behaviour has revealed that participants using the Batched (Position) interface produced the greatest data volume. They made significantly more crater markings per image, more marking clusters were produced by participants identifying the same crater, and a greater number of participants contributed to each marking cluster. This supports hypothesis H3 (Eveleigh et al., 2014), with the simpler task workflow involving fewer task types (1 detection, 1 matching – 1 mouse click) resulting in a greater data volume collected compared to the workflows involving more task types of the other interfaces. However, although this greater volume corresponds to fewer false negatives (craters not identified) when compared to the expert data, it has also resulted in a greater number, and percentage, of false positives (features/artefacts incorrectly identified as a crater). Connected to this, marking clusters made using the Batched (Position) interface also required more participant markings (at least 8) before the percentage of true positives reached 100% - so at least 8

participants had to mark a crater for it to be seen as a definite true-positive. An added limitation of the data garnered from the Batched (Position) interface is that it is less detailed, returning only the position of identified craters and not the size. Although the interface advances to include all marking tools after a number of images, as the methodology section described, the advantage of greater coverage is then lost, with the Batched (Mark) interface returning fewer clusters and true positives than the Sequenced and Full interfaces.

These findings would suggest that researchers involved need to consider if the advantage of greater data volume, and therefore fewer unidentified features, outweighs the disadvantages of having to deal with the increase in false positives and needing a greater number of volunteers to classify each image to ensure the markings are correct. The level of detail required (crater existence, position or size for example), the number of images that need to be analysed and the potential size of the volunteer community taking part could all influence this decision.

Beyond the relationship described by hypothesis H3, analysis of participant markings also suggested that the type of task presented and judgement required influences participant marking agreement (Hutt et al., 2013). This effect also exists when comparing participant markings with the expert equivalent. Considering the position and diameter of craters marked by participants when compared to those of experts, it could be suggested that volunteer performance is superior when considering data or measures that are directly tied to the task they have been asked to do. For instance, markings of crater position showed a significantly greater inter-participant agreement when made using the Sequenced and Batched (Position) interfaces and a significantly greater agreement with the expert when using the Batched (Position) interface – both interfaces where participants marked the central position of the crater with a specific tool. This agreement was reduced with the Full and Batched (Mark) interfaces where crater position is calculated from the size markings rather than directly measured. This supports previous research regarding the advantages of breaking down a larger task into smaller micro-tasks (Cheng et al., 2015), that directly targets the metric required. Likewise, measures of crater diameter showed significantly more agreement both between participants and with experts when using the Full and Batched (Mark) interfaces compared to the Sequenced – where marking the size is one continuous task rather than broken down into steps, marking a crater's position before adjusting its size.

Although both VCS developers and science teams may be tempted to gather as many different measures as possible from a single volunteer contribution, there are caveats. The results of this study show that if it is important to capture more than one measure for the overall scientific goal of the project, performance is better both in terms of inter-participant and expert agreement when the tasks are clearly separated, and the participant is aware of the main goal of the task presented.

6.5. Implications for Virtual Citizen Science Design

Through observing the direct manipulation of task workflow design factors such as autonomy, variety, task type and judgement, this study has shown that they can influence the outcomes of a VCS project, in terms of data accuracy, volume, and volunteer preference and opinion. The implication of this finding is that different development strategies regarding the design of the volunteers' task workflow could be tailored to the specific requirements of a VCS project and its scientific goals. In addition to the influence of development design choices regarding task workflow design, crater-marking and self-reported analysis has also confirmed the existence of interplay between the factors themselves as found in previous research (Chung-Yan, 2010), collectively contributing to both participants' perceived motivation and the data produced.

For instance, although participants reported favouring autonomy and task variety, a small number referenced an interaction between the two. One less confident participant indicated a preference for task variety only when autonomy was restricted, i.e. they were led through each different task step by

step (Sequenced interface), and found having the freedom to choose from a variety of tasks (Full interface) led to a fear of performing poorly or "missing something out". When considering crater-marking behaviour, the interface involving the simplest task (Batched (Position)) resulted in the largest data volume collected by participants in agreement with Eveleigh et al. (2014). However, not only does such an interface restrict variety and autonomy to the detriment of participant preference, but there is also the caveat of reduced performance in terms of crater identification (with the greater number of false-positives marked causing a reduction in precision). Based on these findings, several potential scenarios exist towards which TWD could be tuned to achieve the desired outcome:

- **Data Volume**: If the scientific goal of a project is reliant on the amount of data produced, for instance if a certain spatial area has to be analysed, or a certain rate calculated, then a TWD configuration involving fewer task types (such as the Batched (Position) interface) could be used in order to ensure a higher recall rate. As previously mentioned however, this would result in a reduced precision.
- **Identification Accuracy**: If a higher rate of true positives is required, perhaps because dealing with false detections is particularly difficult with the associated science case, then the Full or Sequenced TWD configurations (providing greater task variety) could be used to ensure a higher precision rate. Based on the responses of participants, the Sequenced approach could be used to guide a relatively new volunteer community, whilst the Full approach could be used with a more experienced volunteer community – allowing them more autonomy to complete the tasks. Both approaches however would result in a lower recall rate.
- **Measurement Accuracy**: If the scientific goal is not only reliant on the accuracy of identification, but also on a certain accuracy of its measurement (be it size, position, or some other scaled response), then a TWD configuration involving tasks that directly measure this metric would be beneficial. For instance in the case of craters, the Batched (Mark) interface would provide the best accuracy in terms of measuring the diameter of the crater, whilst the Batched (Position) would be best for crater position – both configurations that clearly delineate the task required to the participant.

With each scenario and potential TWD solution, any benefit to one measure of performance could be offset by an effect on others. It could therefore be concluded that it is not only the consideration of individual TWD factors alone, but also their interplay and interactions, which determines the best VCS interface design in order to achieve a projects' scientific goals. Conversely, there is unlikely to be a 'one size fits all' task workflow design configuration that is the optimal framework for 100% of VCS projects. Developers will have to carefully consider the aims of the project, what is needed to achieve them (data volume, degree of accuracy, number of classifications etc.) and how TWD factors both individually and collectively can be considered to best support the process.

6.6. Limitations and Future Directions

In order to investigate the manipulation of task workflow design factors on both performance and volunteer preference, a laboratory setting was used. This allowed other factors, such as the imagery shown to the participants, the hardware used, and the environmental conditions that might influence performance and user experience to be kept constant. However, virtual citizen science projects are normally conducted online by volunteers 'in the wild', and as such platform developers do not have any control over their demographics, equipment, and environment or how long they spend on the platform. In order to mitigate any effects caused by these differences, participants of this study had a background representative of the average citizen scientist in terms of education and IT literacy, whilst the time

participants spent using each interface (approximately 10 minutes) is also comparable to the average visit time of existing Zooniverse projects (Sprinks et al., 2015). Another limitation of the study could be related to the fixed science case. In order to control for any extrinsic motivation caused by the science involved, a purposely-derived science case involving crater analysis was used. As such, it could be argued that any findings regarding the influence of task workflow design would only be applicable to crater marking activities. However, the types of visual inspection tasks and judgements involved have been widely used across a number of existing projects, independent of the discipline involved (Sprinks et al., 2015b). It therefore can be assumed that any findings and conclusions made regarding task workflow design can inform the wider citizen science community.

Perhaps the major limitation of this work is that in using a laboratory setting for the advantages previously explained, it is not possible to measure any impact on participants' platform engagement. Whilst qualitative responses have been considered regarding interface preference, future research should consider how this translates into volunteer behaviour when using the platform 'live' - in terms of how often they visit, for how long, and how much data they produce. Related to this, it would also be advantageous to consider the influence of task workflow design more longitudinally, investigating whether volunteer preference and performance changes over time, and how TWD can be adapted in response. Finally, in demonstrating the effect of task workflow design factors on performance and preference, this study has also highlighted the complexity regarding their interplay and influence. In order to better understand these complexities further work is needed to study how they are combined and configured in the overall task workflow design of a VCS platform.

7. Conclusion

Using a laboratory study to test the effect of manipulating Task Workflow Design factors with regards to a specific Virtual Citizen Science platform, we found that factors including autonomy, variety, task type and judgement type had an effect on both the volunteers' experience and their data output.

Through self-reported scores and associated 'free-text' responses adding context, participants indicated a preference for greater variety and autonomy. However, out of these two factors responses indicated autonomy as the most important, supporting the theory that the effect of variety is dependent on the timing of the change of task (Lasecki et al., 2014) – with participants preferring to control when it occurs. Crater marking data however does not support previous research (Chung-Yan, 2010) that found that greater autonomy also improves volunteer performance, with the interface of greatest autonomy (Full) not resulting in significantly better results in any of the performance measures used in this study.

The analysis of crater marking behaviour has also indicated that manipulating certain TWD factors affects performance, dependent on the type of performance measure considered. It was found that the interface that provided fewer task types, less variety and less autonomy (Batched (Position)) resulted in greater data volume (Eveleigh et al., 2014) collected at a faster rate (in terms of time per image). The caveat is, however, that whilst greater coverage resulted in a greater number of true positives when compared to the expert data, it also resulted in a greater number of false positives. This ultimately resulted in more participants required to contribute (i.e. seeing the image and detecting the crater) before a marking cluster can be definitely considered a crater. Performance in terms of agreement, both between participants and with the expert equivalent, was significantly improved when using the interfaces that included tasks that directly measured the required metric (Sequenced and Batched (Position) for crater position, Full and Batched (Mark) for crater diameter).

Overall, the results of this study support the findings of previous research when considering the opinions of the volunteer, with preference given to greater autonomy and variety suggesting that such interfaces can provide an intrinsic motivation to take part. However, when considering volunteer performance the picture is more complex, with different TWD factors affecting measures such as

coverage, participant agreement and expert agreement in different ways. VCS developers and science teams may well have to consider which lessons from existing TWD human factors and work design research are applicable to the citizen science case, by considering the type of data required, the amount of data that needs analysing and the prospective size of their volunteer community when considering the design of the tasks and how they are presented.

Acknowledgements
James Sprinks, Robert Houghton, Steven Bamford and Jeremy Morley were supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/G037574/1) and the RCUK's Horizon Digital Economy Research Institute (RCUK Grant No. EP/G065802/1). The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under iMars grant agreement no. 607379. 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.

Author Biographies

Dr James Sprinks
James Sprinks is a Research Associate at the University of Nottingham's Human Factors Research Group, based within the Nottingham Geospatial Institute. His research considers the use of Citizen Science to identify and map geomorphological features on the surface of Mars. He is researching how citizen science platforms can be designed and implemented to both ensure that the data generated is in a usable format and scientifically robust, while still maintaining a user experience that results in volunteers enjoying and learning from the process. This process encompasses elements of planetary science, remote sensing, Human-Computer Interaction, GIS, ergonomics and human factors research.

Dr Jessica Wardlaw
Jessica is a Research Associate at the Nottingham Geospatial Institute where she is currently developing a Citizen Science platform for the iMars EU Framework 7 project, which is developing tools and 3D models for analysis and interpretation of imagery of Mars, to identify surface changes since NASA's Viking Orbiter mission in the 1970s. Her research interests span Geography (including Web GIS, cartography, spatial cognition and knowledge construction, health geography) and Human-Computer Interaction (from applied aspects such as user-centred design, usability engineering and design practice, to cognitive aspects including sense- and decision-making and information visualisation).

Dr Robert J Houghton
Dr Robert J. Houghton is an Assistant Professor in Human Factors in the Faculty of Engineering at the University of Nottingham. He specialises in cognitive and systems ergonomics and has carried out research in both laboratory and field settings relevant to topics such as digital economy services, multimodal interfaces and the development of mobile technology.

Dr Steven Bamford
Dr Steven Bamford is a Lecturer in the School of Physics and Astronomy at the University of Nottingham. His main research interests are in extragalactic astronomy, but through early involvement in Galaxy Zoo he came to be founding Science Director of the Citizen Science Alliance. Through the

Zooniverse, this organisation has brought authentic involvement in scientific research to millions of participants, and led to valuable science results in many fields. Dr Bamford continues to contribute to the coordination and utilisation of citizen science, with a particular interest in improving the efficiency and quality of citizen science.

Jeremy Morley
Jeremy Morley has been the Chief Geospatial Scientist at Ordnance Survey since April 2015 where he leads the Research and Education team who focus on research in collaboration with universities and others, and promoting spatial literacy and geospatial education. He has worked in geospatial research since the mid-90s. His interests in geographic information science include crowd-sourcing and citizen science, 3D GIS, and open and interoperable geographical information services. From 2009 to 2015 he was Geospatial Science Theme Leader in the Nottingham Geospatial Institute during which time he was involved in three different European projects associated with mapping Mars.

References

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S., 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. IEEE Internet Comput. 17, 76–81.

Cai, C.J., Iqbal, S.T., Teevan, J., 2016. Chain Reactions: The Impact of Order on Microtask Chains, in: Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems. ACM, San Jose, USA, p. Pre-press.

Cheng, J., Teevan, J., Iqbal, S.T., Bernstein, M.S., 2015. Break It Down: A Comparison of Macro- and Microtasks, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15. ACM, New York, NY, USA, pp. 4061–4064. doi:10.1145/2702123.2702146

Chung-Yan, G.A., 2010. The nonlinear effects of job complexity and autonomy on job satisfaction, turnover, and psychological well-being. J. Occup. Health Psychol. 15, 237–251.

Cox, J., Oh, E.Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G., Holmes, K., 2015. Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects. Comput. Sci. Eng. 17, 28–41. doi:10.1109/MCSE.2015.65

Crowston, K., Fagnot, I., 2008. The Motivational Arc of Massive Virtual Collaboration. Proc. IFIP WG 95 Work. Conf. Virtuality Soc. Massive Virtual Communities.

Curtis, V., 2014. Online citizen science games: Opportunities for the biological sciences. Appl. Transl. Genomics, Global Sharing of Genomic Knowledge in a Free Market 3, 90–94.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D., 2011. Gamification. Using Game-design Elements in Non-gaming Contexts, in: CHI '11 Extended Abstracts on Human Factors in Computing Systems, CHI EA '11. ACM, New York, NY, USA, pp. 2425–2428.

Dodd, N.G., Ganster, D.C., 1996. The interactive effects of variety, autonomy, and feedback on attitudes and performance. J. Organ. Behav. 17, 329–347.

Dow, S., Kulkarni, A., Klemmer, S., Hartmann, B., 2012. Shepherding the Crowd Yields Better Work, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 1013–1022. doi:10.1145/2145204.2145355

Dubinsky, A.J., Skinner, S.J., 1984. Impact of job characteristics on retail salespeople's reactions to their jobs. J. Retail. 60, 35–62.

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., Cox, A.L., 2014. Designing for Dabblers and Deterring Drop-outs in Citizen Science, in: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14. ACM, New York, NY, USA, pp. 2985–2994.

Eveleigh, A., Jennett, C., Lynn, S., Cox, A.L., 2013. I Want to Be a Captain! I Want to Be a Captain!: Gamification in the Old Weather Citizen Science Project, in: Proceedings of the First International Conference on Gameful Design, Research, and Applications, Gamification '13. ACM, New York, NY, USA, pp. 79–82.

Farell, B., Pelli, D.G., 1999. Pscyhophysical Methods, or how to measure a threshold and why, in: Vision Research: A Practical Guide to Laboratory Methods. Oxford University Press, New York.

Freitag, A., Meyer, R., Whiteman, L., 2016. Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. Citiz. Sci. Theory Pract. 1.

Gerhart, B., 1987. How important are dispositional factors as determinants of job satisfaction? Implications for job design and other personnel programs. J. Appl. Psychol. 72, 366–373.

Ghani, J.A., Deshpande, S.P., 1994. Task Characteristics and the Experience of Optimal Flow in Human—Computer Interaction. J. Psychol. 128, 381–391.

Hackman, J.R., Oldham, G.R., 1975. Development of the Job Diagnostic Survey. J. Appl. Psychol. 60, 159–170.

Hand, E., 2010. Citizen science: People power. Nat. News 466, 685–687.

Harrower, M., Sheesley, B., 2005. Designing Better Map Interfaces: A Framework for Panning and Zooming. Trans. GIS 9, 77–89.

Hennon, C.C., Knapp, K.R., Schreck, C.J., Stevens, S.E., Kossin, J.P., Thorne, P.W., Hennon, P.A., Kruk, M.C., Rennie, J., Gadéa, J.-M., Striegl, M., Carley, I., 2014. Cyclone Center: Can Citizen Scientists Improve Tropical Cyclone Intensity Records? Bull. Am. Meteorol. Soc.

Huang, C.-Y., 2002. Distributed manufacturing execution systems: A workflow perspective. J. Intell. Manuf. 13, 485–497.

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

Iacovides, I., Jennett, C., Cornish-Trestrail, C., Cox, A.L., 2013. Do Games Attract or Sustain Engagement in Citizen Science?: A Study of Volunteer Motivations, in: CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13. ACM, New York, NY, USA, pp. 1101–1106.

Jordan, R., Crall, A., Gray, S., Phillips, T., Mellor, D., 2015. Citizen Science as a Distinct Field of Inquiry. BioScience biu217.

Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J., 2013. The Future of Crowd Work, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13. ACM, New York, NY, USA, pp. 1301–1318.

Kulkarni, A., Can, M., Hartmann, B., 2012. Collaboratively Crowdsourcing Workflows with Turkomatic, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 1003–1012. doi:10.1145/2145204.2145354

Kulkarni, A.P., Can, M., Hartmann, B., 2011. Turkomatic: Automatic Recursive Task and Workflow Design for Mechanical Turk, in: CHI '11 Extended Abstracts on Human Factors in Computing Systems, CHI EA '11. ACM, New York, NY, USA, pp. 2053–2058. doi:10.1145/1979742.1979865

Lahav, O., Naim, A., Buta, R.J., Corwin, H.G., de Vaucouleurs, G., Dressler, A., Huchra, J.P., van den Bergh, S., Raychaudhury, S., Sodré, L., Storrie-Lombardi, M.C., 1995. Galaxies, human eyes, and artificial neural networks. Science 267, 859–862.

Lasecki, W.S., Marcus, A., Rzeszotarski, J.M., Bigham, J.P., 2014. Using Microtask Continuity to Improve Crowdsourcing. Carnegie Mellon Univ. Hum.-Comput. Interact. Inst. - Tech. Rep. CMU-HCII-14-100.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R.C., Raddick, M.J., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J., 2011. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. Mon. Not. R. Astron. Soc. 410, 166–178.

Mankowski, T.A., Slater, S.J., Slater, T.F., 2011. An Interpretive Study Of Meanings Citizen Scientists Make When Participating In Galaxy Zoo. Contemp. Issues Educ. Res. CIER 4, 25–42.

Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M.E., Lintott, C.J., Smith, A.M., 2013. Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing, in: First AAAI Conference on Human Computation and Crowdsourcing. Presented at the First AAAI Conference on Human Computation and Crowdsourcing.

Nov, O., Arazy, O., Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. Proc. AAAI Int. Conf. Weblogs Soc. Media 249–256.

Oldham, G.R., Hackman, J.R., 2010. Not what it was and not what it will be: The future of job design research. J. Organ. Behav. 31, 463–479.

Pelli, D.G., Farell, B., 2010. Psychophysical Methods, in: Handbook of Optics. McGraw-Hill, New York, p. 3.1-3.12.

Ponciano, L., Brasileiro, F., Simpson, R., Smith, A., 2014. Volunteers' Engagement in Human Computation for Astronomy Projects. Comput. Sci. Eng. 16, 52–59.

Prather, E.E., Cormier, S., Wallace, C.S., Lintott, C., Raddick, M.J., Smith, A., 2013. Measuring the Conceptual Understandings of Citizen Scientists participating in Zooniverse Projects: A First Approach. Astron. Educ. Rev. 12.

Prestopnik, N.R., Crowston, K., 2012. Citizen Science System Assemblages: Understanding the Technologies That Support Crowdsourced Science, in: Proceedings of the 2012 iConference, iConference '12. ACM, New York, NY, USA, pp. 168–176.

Raddick, J., Bracey, G., Gay, P.L., Lintott, C.J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J., 2013. Galaxy Zoo: Motivations of Citizen Scientists. Astron. Educ. Rev. 12.

Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C.J., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J., 2009. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. ArXiv09092925 Astro-Ph Physicsphysics.

Reed, J., Raddick, M.J., Lardner, A., Carney, K., 2013. An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects, in: 2013 46th Hawaii International Conference on System Sciences (HICSS). Presented at the 2013 46th Hawaii International Conference on System Sciences (HICSS), pp. 610–619.

Reed, J., Rodriguez, W., Rickhoff, A., 2012. A Framework for Defining and Describing Key Design Features of Virtual Citizen Science Projects, in: Proceedings of the 2012 iConference, iConference '12. ACM, New York, NY, USA, pp. 623–625.

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., Jacobs, D., 2012. Dynamic Changes in Motivation in Collaborative Citizen-science Projects, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 217–226.

Schmidt, M.-T., 1998. Building Workflow Business Objects, in: Patel, D.D., Sutherland, D.J., Miller, J. (Eds.), Business Object Design and Implementation II. Springer London, pp. 64–76.

Sprinks, J., Morley, J., Bamford, S., Houghton, R., 2015. Keeping Citizen Scientists Interested: The Importance of Task Workflow Design. Citiz. Sci. Conf. San Jose CA 11-12th Febr. 2015.

Sprinks, J., Morley, J.G., Houghton, R., Bamford, S., 2015. The Impact of Task Workflow Design on VGI Citizen Science Platforms. GIS Res. UK 2015 Leeds UK.

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

Tinati, R., Van Kleek, M., Simperl, E., Luczak-Rösch, M., Simpson, R., Shadbolt, N., 2015. Designing for Citizen Data Analysis: A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15. ACM, New York, NY, USA, pp. 4069–4078. doi:10.1145/2702123.2702420