Designing Task Workflows To Ensure The Best Scientific Outcomes In Citizen Science

James C Sprinks

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

September 2016

James C. Sprinks: *Designing Task Workflows to Ensure the Best Scientific Outcomes in Citizen Science,* thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, © September 2016.

The author was supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/G037574/1) as well as the RCUK's Horizon Digital Economy Institute (RCUK Grant No. EP/G065802/1).

SUPERVISORS:

Dr. Robert Houghton, Dr. Steven Bamford and Mr. Jeremy Morley.

To my parents, Liz and friends.

In honour of my godson, Sampson James Connibeer; may I be a fitting mentor and guide when needed.

"The intelligence of that creature known as a crowd is the square root of the number of people in it."

Sir Terry Pratchett (1997)

ABSTRACT

Citizen science, or *'public participation in scientific research*', can be described as research conducted, in whole or in part, by amateur or nonprofessional participants often through crowd-sourcing techniques. The advance of modern day internet technology that has made the world a more connected place has resulted in a surge of citizen science projects, especially online platforms which allow volunteers to take part in research in both an asynchronous and geographically abstract manner. Due to this increased interest, citizen science is becoming a distinct field of research in its own right beyond the original scientific problems it was devised to address. Although some of this research has considered interface HCI and functionality, there has been relatively little attention paid specifically to human factors issues.

Through this work we attempt to address this shortfall, by considering citizen science as a form of 'work'. Due to its similarities in terms of repetition with production lines of the early 20th century and more recently on-screen visual inspection tasks, some of the many decades of ergonomics research in this field are applied specifically to the virtual citizen science arena. We make a first step in considering how virtual citizen science systems can be better designed for the needs of the volunteer, exploring how manipulating task flow affects both the quality of information collected, and the volunteers' experience of using the interface.

A hierarchical task analysis of 12 Zooniverse projects revealed that the types of tasks, judgements and the way they are presented to the volunteer varies greatly, independent of the science discipline involved. Furthermore, through differing designs of the Zooniverse's 'Planet Four: Craters' platform, it was shown that task workflow design factors such as autonomy, variety, task type and volunteer judgement required can influence the amount of data collected, the accuracy of this data and both volunteer engagement and motivation. Simpler tasks with fewer volunteer judgements required resulted in a significantly greater volume of data collected, however accuracy is affected with an increase of false-positive classifications. Volunteers reported a preference for greater autonomy and

task variety, a stance reflected in the number of times they visited and returned to the platform, however this also significantly reduced the accuracy of classifications – both in terms of inter-participant agreement and expert judgement comparison.

The interplay of task workflow factors and their effect has been shown to be a complex affair. Through the empirical data collected, a model has been derived predicting the influence of different task workflow configurations on classification numbers over time since a platforms' launch. It demonstrates that when considering task workflow design, developers of future citizen science platforms will need to perform a balancing act. The importance of user engagement, the data needs of the science case and the resources that can be committed both in terms of time and data reduction will need to be weighed, and balanced with the realistic public reach and promotion the science case can be predicted to generate.

ACKNOWLEDGEMENTS

A big 'thank you' first and foremost to my supervisors, Robert Houghton, Steven Bamford and Jeremy Morley at the University of Nottingham for supporting me throughout my time as a PhD student. The advice, care and encouragement have been a great help! Many thanks also to friends and colleagues at the Horizon Centre for Doctoral Training and the Nottingham Geospatial Institute, for their inspiration, collaboration and in many cases going through the same dilemmas and suffering similar crises of confidence! In no particular order (and probably omitting some): Matthew Terrell, David Hand, Will Knight, Martin Porcheron, Kyle Arch, Richard James, Horia Maior, Christopher Ellis, Matt Voigts, and Andreas Reiter. Special thanks also to Emma Juggins, Felicia Black and Alison Mellor for their help through the jungle of university admin!

Many thanks to the Zooniverse development team (past and present) at both Adler Planetarium and the University of Oxford for their advice and support in developing the Planet Four: Craters platform: Robert Simpson, Stuart Lynn, Michael Parrish, Grant Miller, Chris Snyder, Meg Schwamb and Brian Carstensen – whose help, insight and enthusiasm acted as an inspiration for the work. Thanks also to Jennifer Taylor at the University of Bristol for providing a science case and expert point of view, and of course to the thousands of citizen scientists who took part – giving both their invaluable time and feedback to guide the research.

Thank you to my friends and family from Cornwall: Mum, Dad, Granddad, Meg, Clair, Rory, Paul, Polly, Ben B & Ben C, Andy, Bush, Jess, Parsons and Shelly, whose support and love have made this possible, and of course Cornwall itself – *Onen hag oll.*

Last but by no means least, thanks to Spot and Optimus – and Liz, for saying 'yes' (what were you thinking)!

CONTENTS

1 Introduction **20**

1.1 Defining the Problem *22*

1.2 Motivation *23*

1.3 Research Questions *24*

1.4 Research Domains *25*

1.5 Contributions *25*

1.6 Publications of this Work *27*

1.7 Thesis Structure *28*

Part I Background **31**

2 The Growth of Citizen Science **32**

2.1 The Beginnings of Citizen Science *32*

2.2 Citizen Science in the Information Age *33*

2.3 The Continuing Importance of Citizen Science *35*

2.3.1 The Advancement of Data Collection 35

2.3.2 Automated Data Analysis 36

2.3.3 Outreach & Education 37

2.4 The Birth of the Zooniverse *37*

2.5 The Contribution that Citizen Science Results Make *40*

2.5.1 Data Catalogues and New Discoveries 40

2.5.2 Validating and Reducing Citizen Science Data 41

2.6 Citizen Science as a Distinct Field of Enquiry *42*

2.6.1 Virtual Citizen Science Typology & Functionality 42

2.6.2 Gamification 43

2.6.3 Citizen Scientist Motivations 44

2.6.4 Volunteer Behaviour & Patterns of Participation 45

2.7 This Thesis' Definition of Citizen Science *48*

2.8 Summary *49*

3 Citizen Science Design and Methods **51**

3.1 Crowdsourcing *51*

3.2 Psychophysics & Psychophysical Methods *53*

3.2.1 Threshold & Detection 54

3.2.2 Matching & Scaling 55

3.3 Visualisation Factors *56*
3.4 Vigilance *58*
3.5 Operator Variability *60*
3.6 Citizen Science as a Form of Work *61*
3.6.1 Feedback 62
3.6.2 Complexity, Variety & Autonomy 64
3.6.3 Human Processing Units: A New Form of Taylorism 65
3.7 Summary *66*

Part II Approach & Methodology **68**

4 Approach To Studying Task Workflow Design **69**
4.1 Task Workflow Design *69*
4.2 Task Workflow Design Factors and Their Effect *71*
4.2.1 Hackman & Oldham's Job Characteristic Theory 71
4.2.2 Pelli & Farell's Psychophysical Methods Approach 73
4.3 A Conceptual Model of Task Workflow Design *75*
4.4 Task Workflow Design Factors of Virtual Citizen Science *77*
4.4.1 Science Case Influenced Factors 78
4.4.2 Design Influenced Factors 79
4.4.3 Postulates from Related Work 81
4.5 Focus of This Thesis *83*
4.5.1 Focus on Design Influenced TWD Factors 83
4.5.2 Relation of the Model of Design Influenced Factors to the Empirical Work 84
4.5.3 Measuring the Scientific Outcomes of a VCS Platform 85
4.5.4 Research Questions 86
4.6 Summary *87*

5 Methodology of Studying Task Workflow Design **89**
5.1 Overview of Studies and Methods Employed *90*
5.2 The Zooniverse Setting *90*
5.3 Mixed Methods Approach *92*
5.3.1 Hierarchical Task Analysis (HTA) 92
5.3.2 Questionnaires 93
5.4 Planet Four *95*
5.4.1 The Crater Counting Science Case 96
5.4.2 Volunteer-Expert Data Comparison 97

5.4.3 Participant Issues 97
5.4.4 Ethical Issues 98
5.5 Studies in this Thesis *99*
5.6 Summary *100*

Part III Empirical Work **101**

6 The Tasks of Citizen Science Platforms **102**
6.1 Introduction *102*
6.2 Research Questions *104*
6.3 Experimental Design *105*
6.3.1 Current Virtual Citizen Science Platforms 105
6.3.2 Hierarchical Task Analysis (HTA) 106
6.3.3 Measuring Volunteer Behaviour through Website Analytics 109
6.4 Analysis and Results *110*
6.4.1 Hierarchical Task Analysis Findings 110
6.4.2 Website Analytics 114
6.4.3 The Effect of Task Workflow Design on Volunteer Behaviour 117
6.5 Discussion *119*
6.5.1 Limitations 119
6.5.2 Hierarchical Task Analysis 120
6.5.3 Website Analytics 121
6.5.4 Task Workflow Design vs. Volunteer Behaviour 121
6.6 Conclusions *122*
6.7 Summary and Outlook *124*

7 The Volunteer's Point of View **126**
7.1 Introduction *126*
7.2 Research Questions *127*
7.3 Experiment Design *128*
7.3.1 Questionnaire Design 128
7.3.2 Role 130
7.3.3 Participants 131
7.4 Analysis and Results *131*
7.4.1 Design & Usability Results 131
7.4.2 Task & Tools Results 133
7.4.3 Imagery Results 137
7.4.4 Discussion & Feedback Results 140

7.4.5 Questionnaire Responses by Age Group 143
7.4.6 Questionnaire Responses by Zooniverse Experience 144
7.5 Discussion *145*
7.5.1 Volunteer Concerns Regarding Design & Usability 146
7.5.2 Volunteer Concerns Regarding Tasks & Tools 146
7.5.3 Volunteer Concerns Regarding Imagery 148
7.5.4 Volunteer Concerns Regarding Discussion & Feedback 149
7.6 Conclusions *151*
7.7 Summary and Outlook *152*

8 Manipulating Task Workflow Design **154**
8.1 Introduction *154*
8.2 Estimating the Age of the Martian Surface *155*
8.3 Image Processing and Preparation *158*
8.4 Research Questions and Hypotheses *160*
8.5 Experiment Design *161*
8.5.1 Crater Marking Tools 163
8.5.2 Classification Interface Design 164
8.5.3 Other Platform Pages 166
8.5.4 Questionnaire Design 168
8.5.5 Participants 168
8.5.6 Design and Procedure 169
8.5.7 Role 169
8.6 Results and Analysis *170*
8.6.1 Questionnaire Results 170
8.6.2 Crater Marking Results 175
8.6.3 Summary and Hypotheses 185
8.7 Discussion *186*
8.7.1 The Effect of TWD on Participant Engagement and Motivation 186
8.7.2 Greater Coverage: But at What Cost? 187
8.7.3 Volunteers' Time is Precious 188
8.7.4 Be Direct with your Volunteers 189
8.7.5 Unforeseen Advantages of a Laboratory Trial 190
8.8 Conclusions *192*
8.9 Summary and Outlook *194*

9 Task Workflow Design 'In the Wild' 195
9.1 Introduction 195
9.2 Image Processing and Preparation 196
9.3 Research Questions and Hypothesis 198
9.4 Experiment Design 200
9.4.1 Interface & Platform Design 202
9.4.2 Questionnaire & Online 'Talk' Forum 203
9.4.3 Participants 203
9.4.4 Procedure 204
9.5 Results and Analysis: Self-Reported 205
9.5.1 Online Questionnaire: Likert-Type Responses 205
9.5.2 Questionnaire Comments and 'Talk' Forum Posts 208
9.6 Results and Analysis: Behavioural 211
9.6.1 Website Analytics 211
9.6.2 Crater Marking Behaviour 213
9.6.3 Participant Agreement 214
9.6.4 Participant – Expert Comparison 217
9.6.5 Summary and Hypotheses 223
9.7 Discussion 224
9.7.1 The Effect of TWD on Participant Engagement and Motivation 224
9.7.2 Time Gentlemen, Please! 225
9.7.3 The Right Tools for the Right Job 227
9.7.4 Separating the Wheat from the Chaff 227
9.7.5 Other Lessons Learnt from a 'Real World' Study 229
9.8 Conclusions 231
9.9 Summary 233

Part IV Synopsis 235

10 Discussion 236
10.1 Designing VCS Platforms for the Best Scientific Outcomes 236
10.2 Conceptual Model of Task Workflow Design Factors 240
10.2.1 Empirical Contributions to Validating the Model 242
10.2.2 Design-Influenced Task Workflow Design Factors 244
10.2.3 Science Case Influenced Task Workflow Design Factors 245
10.2.4 Interplay of TWD Factors 246

10.3 Implications for Virtual Citizen Science Design *247*
10.3.1 Turn Up the Volume 250
10.3.2 Quality over Quantity 251
10.3.3 A Balancing Act 251
10.3.4 Unforeseen Practical Implications 252
10.4 Summary *254*

11 Conclusions **255**
11.1 Summary *255*
11.2 Key Contributions and Conclusions *256*
11.2.1 Methodology 256
11.2.2 Understanding Task Workflow Design 257
11.2.3 Implications for Design 260
11.3 Critical Reflection *261*
11.3.1 Limitations of the Work 262
11.3.2 Reflection on Task Workflow Design for VCS Systems 263
11.4 Future Work *265*

12 References **267**

A Appendix to Study 1 **281**

B Appendix to Study 2 **283**

C Appendix to Experiment 1 **290**

D Appendix to Experiment 2 **298**

LIST OF TABLES

Table 4.1: Hackman & Oldham's JCT Characteristics 72

Table 5.1: Overview of thesis studies and methods 90

Table 6.1: Zooniverse Virtual Citizen Science platforms 105

Table 6.2: Task type & volunteer judgement categories 108

Table 6.3: Hierarchical Task Analysis results 111

Table 6.4: Task Workflow Factors by discipline 112

Table 6.5: Mean average of each task type per discipline 113

Table 6.6: Mean average of each judgement type per discipline 113

Table 6.7: Website statistics for each Zooniverse platform 115

Table 7.1: 'Free' responses to design & usability section 133

Table 7.2: 'Free' responses to tasks & tools section 136

Table 7.3: 'Free' responses to imagery section 139

Table 7.4: 'Free' responses to discussion & feedback section 142

Table 8.1: Task workflow design configuration of each interface 166

Table 8.2: Number of responses by category and interface 174

Table 8.3: Crater marking agreement for each interface 178

Table 8.4: Number of crater clusters per size bin 180

Table 8.5: Crater identification compared to expert 181

Table 8.6: Participant crater marking sizes compared to expert 184

Table 9.1: Number of responses by category and interface 209

Table 9.2: Crater marking agreement for each interface 214

Table 9.3: Number of crater clusters per size bin 217

Table 9.4: Crater identification compared to expert 218

Table 9.5: Participant crater marking sizes compared to expert 221

Table 10.1: Model of Task Workflow Design Factors, its grounding in related work, and this thesis' contribution 243

LIST OF FIGURES

Figure 1.1: Venn diagram of relevant research disciplines and thesis contributions 26

Figure 2.1: The Pareto Distribution 46

Figure 4.1: Model of the factors of task workflow design. Factors within the red box are adapted from psychophysical methods work, whilst the remaining come from Hackman & Oldham's JCT 76

Figure 4.2: Model of Design Influenced TWD Factors 84

Figure 6.1: HTA flow diagram for Seafloor Explorer website 106

Figure 6.2: Attrition of visits in VCS projects 114

Figure 6.3: Visit frequency of volunteers 116

Figure 6.4: Task type verses volunteer behaviour 117

Figure 6.5: User judgement verses volunteer behaviour 118

Figure 6.6: Variety and freedom verses volunteer behaviour 119

Figure 7.1: Planet Four Interface 130

Figure 7.2: Participant responses to Design & Usability section 132

Figure 7.3: Participant responses to task mental and physical demand 134

Figure 7.4: Participant responses to pace, success, effort & frustration 135

Figure 7.5: Participant responses to 'fan' and 'blotch' tool use 135

Figure 7.6: Participant responses to imagery statements 138

Figure 7.7: Participant responses regarding profile, discussion & blog 140

Figure 7.8: Participant responses regarding understanding & contribution 141

Figure 7.9: Average response to blog questions per age group 144

Figure 7.10: Participant responses to tool 'ease of use' and discussion page statements by Zooniverse experience 145

Figure 8.1: Isochrons for Martian Crater Populations 157

Figure 8.2: Original G05_020119_1895_XN_09N198W image (left) and first row of slices (right) 160

Figure 8.3: Overview of the experimental design in relation to the model of Task Workflow Design factors 162

Figure 8.4: Platform tools (above) and annotations (below) 163

Figure 8.5: Flow diagram of tools available to the user for each interface 165

Figure 8.6: Screenshots of Planet Four interfaces 166

Figure 8.7: Participant responses to NASA TLX statements (with standard error shown) 171

Figure 8.8: Participant responses per interface (with SE shown) 172

Figure 8.9: Response differences per interface (with SE shown) 173

Figure 8.10: Crater marking results (with SE shown) 176

Figure 8.11: Comparison of marking agreement per interface 179

Figure 8.12: Participant crater identification agreement 179

Figure 8.13: Expert and participant average crater markings per interface 182

Figure 8.14: Position and diameter difference of participant markings compared to expert 183

Figure 8.15: Cluster participant contribution verses fraction of true positives 185

Figure 8.16: Example of false positive participant markings (centre) 191

Figure 9.1: CTX image G23_027332_1907_XN_10N202W (left - top section, right - bottom section) 198

Figure 9.2: Overview of the experimental design in relation to the task workflow design model 201

Figure 9.3: Planet Four: Craters home page 202

Figure 9.4: Participation in other projects - by site (top) and number (bottom) 206

Figure 9.5: Participant responses to NASA TLX statements (with standard error shown) 207

Figure 9.6: Participant responses to statements regarding platform imagery (with standard error shown) 208

Figure 9.7: Time spent on site and number of images classified (with standard error shown) 212

Figure 9.8: Number of return visits per interface 212

Figure 9.9: Crater marking results (with SE shown) 213

Figure 9.10: Comparison of marking agreement per interface 216

Figure 9.11: Participant crater identification agreement 216

Figure 9.12: Expert and participant average crater markings per interface 219

Figure 9.13: Position and diameter difference of participant markings compared to expert (with standard error shown) 220

Figure 9.14: Cluster participant contribution verses fraction of true positives compared to expert 222

Figure 9.15: Examples of false positive participant markings (larger blue circles – marking peaks rather than craters) 230

Figure 10.1: Effect of Design Influenced TWD Factors on VCS Outcomes 245

Figure 10.2: Number of classifications (images) completed over time using each TWD interface for each scenario 249

LIST OF ABBREVIATIONS AND ACRONYMS

ANOVA	Analysis of Variance
API	Application Programming Interface
BBC	British Broadcasting Corporation
CSA	Citizen Science Alliance
CTX	Context Camera
DOT	Dictionary of Occupational Titles
DV	Dependent Variable
F(1..*n*)	Study Participant Number – using Full interface
GIS	Geographical Information System
HCI	Human Computer Interaction
HiRISE	High Resolution Imaging Science Experiment
HTA	Hierarchical Task Analysis
InSight	Interior Exploration using Seismic Investigations, Geodesy and Heat Transport
ISIS	Integrated Software for Imagers and Spectrometers
IV	Independent Variable
JCT	Job Characteristic Theory
Ma	Mega-annum (million years)
MRO	Mars Reconnaissance Orbiter
MUG	Microsoft Usability Guidelines
NASA	National Aeronautics and Space Administration
P(1..*n*)	Study Participant Number
PDS	Planetary Data System
R(1..*n*)	Study Participant Number – using Ramped interface
S(1..*n*)	Study Participant Number – using Stepped interface

SD	Standard Deviation
SE	Standard Error
SDSS	Sloan Digital Sky Survey
TWD	Task Workflow Design
TLX	Task Load Index
USGS	United States Geological Survey
VCS	Virtual Citizen Science

List of Appendices

A Appendix to Study 1 281

B Appendix to Study 2 283

C Appendix to Experiment 1 290

D Appendix to Experiment 2 298

1

INTRODUCTION

The information age has transformed the process of collecting and storing scientific data, with increasingly sophisticated instruments able to study the universe around us in amazing detail at an exponentially increasing rate. In fact, this abundance of data is proving a headache for the professional scientific community, with potentially many terabytes of information sitting on hard drives remaining unexplored for years, even decades, with its potential unrealised.

However, the current digital revolution has made the sharing of data and information increasingly straightforward, in a way that is unrestricted by constraints both geographical and temporal. Over the past decade the scientific community has started to use this to their advantage, sharing their data with the public through specially designed platforms and in return asking them to take on a portion of the analytic workload. As technology continues its relentless advance, it is clear that novel approaches will be needed to deal with the 'data avalanche', and 'citizen science' will continue to grow as a key part of the solution.

One method by which scientists are using the public to assist with scientific analysis is through web-based citizen science platforms. Through allowing the general public to access and annotate varying types of scientific data from the comfort of their own home, it could be argued that this above all else has been responsible for a huge rise in citizen science participation in recent times. Currently there are millions of registered citizen scientists around the world contributing to such platforms across a range of scientific disciplines; marking features on the surface of other planets, transcribing the logbooks of historical shipping voyages, and identifying the calls of different bat species are just a few of the scientific endeavours that are being undertaken by volunteers. This diversity of pursuit is only likely to

increase as other research areas investigate the gains that can be made through using citizen science.

The success of citizen science platforms is not restricted to the size of communities they have built and the subsequent increase in public engagement, but also stretches to their scientific impact. Huge data catalogues have been produced on the back of volunteers' contributions, resulting in a compelling increase in citizen science-based publications over recent years (Follett & Strezov, 2015). However, this increase in presence has put citizen science under the spotlight, and although evidence increasingly supports the process questions still exist regarding its legitimacy – as highlighted by Gollan (2013):

> *"Many scientists question the quality, reliability and in general, the utility of data.*
>
> *Evidence for the stigma associated with citizen science projects comes from different sources. One was a long-standing community-based program to survey diversity of bird species. Researchers found the estimated numbers of birds changed through time simultaneously with changes in the observers. It was concluded that the trends detected were not likely to represent real changes in bird abundance, but were more likely due to prejudices of the individual observers."*

With the advent of this increased research presence, an increasing number of disciplines are realising the potential of the citizen science approach in regards to furthering their own work. As such its influence is growing, to the point where citizen science associations have been formed in North America, Europe and Australia[1] - with a dedicated conference being held for the first time in 2015. The importance of citizen science has also been recognised at a governmental level, with the U.S. Federal Government introducing the Crowdsourcing and Citizen Science Act in 2015[2] to

---

[1] Citizen Science Association, European Citizen Science Association & Australian Citizen Science Association - http://citizenscience.org/, http://ecsa.citizen-science.net/ and http://csna.gaiaresources.com.au/wordpress/ respectively

[2] Full description of act can be found at: https://www.congress.gov/bill/114th-congress/senate-bill/2113

encourage the use of citizen science across a range of departmental frameworks.

In short, citizen science practices are becoming increasingly interwoven into the fabric of current scientific research. This being the case, it is ever more important that scepticism is addressed head on and that citizen science findings withstand the same academic rigour as traditional techniques.

1.1 DEFINING THE PROBLEM

The development of citizen science platforms can be portrayed as the coming together of many different stakeholders (the scientist, the developer, the volunteer etc.), linked by a web-based service and the data it displays – with the ultimate goal being a mutually beneficial system that is both engaging to the volunteer and produces valid scientific results.

In the earlier days of online citizen science, platforms were predominantly designed by web developers informed by the scientists requiring the data, and as such understandably were designed with this primary objective in mind. Volunteer interfaces, tasks and tools were created to merely capture this data, with perhaps less consideration given to the differing ways it could be achieved. In terms of validating the data retrospective approaches have been the norm, either through identifying better-performing volunteers (Freitag, Meyer, & Whiteman, 2016) or using data sampling techniques (Chambert, Miller, & Nichols, 2014). This is conceivably an irony, creating work for the citizen science developers and expert scientists within a process that essentially is designed to reduce the workload.

Although online citizen science is a relatively new endeavour, the related practices of crowdsourcing (Howe, 2006) and the atomisation of work (Cheng et al., 2015) have a large body of human factors-related research to guide their development. Although growing in consideration, as yet citizen science has not made full use of this knowledge, with little consideration given to how the types of volunteer tasks, their progression and their presentation can influence both user engagement and performance.

However at the same time, the expansion of the citizen science field and the increase in its participation allows projects to capture differing aspects of

volunteer behaviour in increasing detail. This provides an exciting opportunity to tackle the long-standing problem in citizen science: to make non-expert collected data as robust as the expert equivalent.

Building on the study of citizen science volunteer tasks and their workflow, this thesis also seeks to investigate how opportunities in online citizen science development can be exploited to create platforms that maximise their scientific outcomes.

1.2 MOTIVATION

In parallel to the growing contribution and influence of citizen science research within the scientific domain, the discipline is rapidly growing as a distinct field of enquiry in its own right. An increasing amount of work has put the volunteer at the centre, considering Human-Computer Interaction, functionality (Prestopnik & Crowston, 2012) and public motivation (Raddick et al., 2009) and its effect on engagement and performance. The motivation of this thesis is to contribute to this body of work, and to the design of citizen science platforms that lead to improved outcomes in the view of the scientists and developers involved. It aims to do this by applying the lessons learnt through the research of related processes to the citizen science domain.

For instance, the design of the tasks and workflows carried out by citizen science volunteers provide an intriguing space in which to study volunteer behaviour, as they have the ability to substantially interfere with or at least shape the volunteers' actions. Furthermore, as shall be expanded upon from the literature (chapter 3), task workflow design and the factors that describe it (explained in 4.3) can have both a beneficial and detrimental effect on engagement and performance. By understanding these relationships it is envisaged that the design of citizen science platforms can be influenced perhaps on a more practical level, compared to other influences that are less controllable due to the nature of citizen science. Through simple design changes regarding the tasks presented to the volunteer, a disproportionate benefit to the outcomes of a project could result compared to the development effort required.

Although task design and its workflow is a well-defined and researched area in several disciplines, including psychology, computer science and

human-computer interaction, it still has many open research questions regarding its use in technical solutions – such as those within citizen science design as examined in this thesis.

1.3 RESEARCH QUESTIONS

The task workflow followed by volunteers is pervasive to citizen science platforms, and due to the broadness of the science addressed can take the form of multiple configurations. As such, the need emerges to understand the effect of different configurations on the scientific outcomes of citizen science. The focus on the affect of Task Workflow Design (TWD) is grounded in the literature review in chapter 3, whilst the concept and approach to studying the manipulation of different task workflow design factors is explained in chapter 4. With the goal of understanding task workflow design, the primary research question addressed by the empirical studies in this thesis is:

- What task workflow designs ensure the best scientific outcomes?

In order to understand the different ways task workflow design can be manipulated and systematically studied in this work, the constituent factors that categorise TWD were synthesised from the related literature described in chapter 3 – and then transformed into a research framework in chapter 4 by asking the following question:

- Which task workflow design factors influence user engagement and performance?

The investigated TWD strategies within this work focused on adapting design-influenced factors. This focus is not just related to the goal of understanding the influence of task workflow design, but also because they are identified as being the most practical adaptations to be of use to the developers of citizen science platforms (see 4.3). Therefore, this thesis' secondary goal is to learn about the design of online citizen science platforms by answering the question:

- What are the design implications for citizen science platforms that may be drawn from the empirical work?

Overall, the goal of this thesis is to develop an in-depth understanding of citizen science task workflow design, and how it can be manipulated to influence and ultimately improve the scientific outcomes of citizen science activities.

1.4 RESEARCH DOMAINS

Due to both the multidisciplinary and practical nature of the thesis, several methods are utilised to make a contribution gained through extracting from, and positioning the thesis with the following research domains:

- *Citizen Science Design* – The study of the design of citizen science platforms, specifically regarding the tasks required of the volunteer and their presentation, describe the application sphere of this thesis.
- *Human Factors* – The study of human behaviour within the citizen science technique and the application of previous knowledge is the over-arching research area of the thesis.
- *Psychophysical Methods* – The thesis is inspired by terminological, conceptual and methodological approaches from related work regarding sensory science, physiology and optics in order to inform experimental design and analysis.
- *Planetary Science* – In order to carry out the empirical work of this thesis, a bespoke citizen science platform was developed – studying craters on Mars. Therefore the thesis both draws from and directly contributes to the discussion of developing citizen science projects for planetary science.
- *Hierarchical Task Analysis* – The method of review employed in the study of current citizen science platforms (chapter 6), in order to define different task workflow design configurations.

1.5 CONTRIBUTIONS

The diagram shown in figure 1.1 presents an approximate thematic overlap of the research areas, with their influence on the thesis represented by their size. The key contributions this thesis (Marked A to E) makes to further knowledge are positioned within the areas where their main

contribution sits. Broadly, A – C contribute to the understanding of task workflow design, while D and E contribute implications for the design of citizen science platforms.

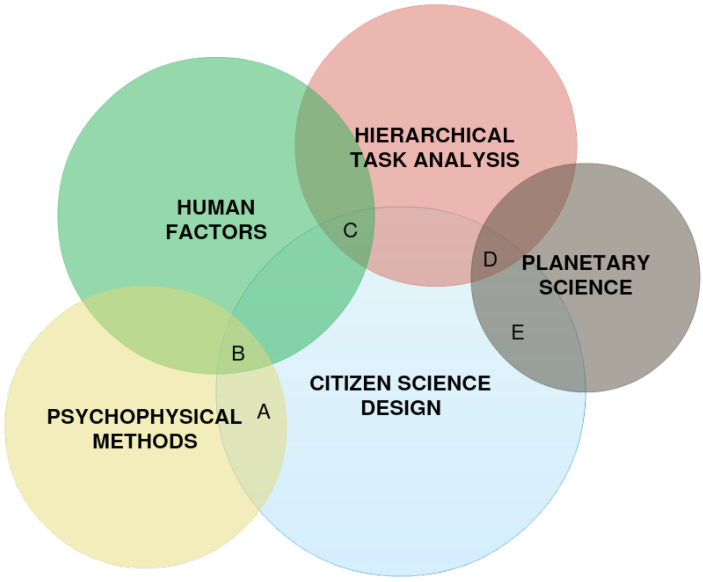


Figure 1.1: Venn diagram of relevant research disciplines and thesis contributions

A: *Adaption of Psychophysical Methods and Job Characterisation Theory (JCT) research to the citizen science domain* – Human factors research regarding JCT and Psychophysical Methods is extended through mapping their constituent factors with the citizen science alternative (see 4.2).

B: *A conceptual model of task workflow design factors* – A classification that serves the practical purpose of identifying factors that can be manipulated at the development stage of a citizen science project. This is derived through a review of related literature and its authenticity verified through the empirical work (see 4.3).

C: *Task workflow design factors interaction* – The empirical work of this thesis expose several ways in which task workflow design factors and their interplay influence the outcomes of a citizen science platform.

D: *Task workflow design management* – The thesis highlights several implications regarding the design of online citizen science platforms. This is both specifically regarding the planetary science case and potentially more broadly across other disciplines. Design

recommendations are stated and a model derived predicting volunteer classification behaviour over time (see 10.3).

E: *Planet Four: Craters citizen science platform* – In order to investigate the manipulation of task workflow design, a bespoke citizen science platform was developed addressing an associated science case.

1.6 PUBLICATIONS OF THIS WORK

Several parts of this thesis have been accepted by peer-review for publication in conference proceedings across several disciplines, or are currently in submission (listed in contents order):

- Parts of the introduction (chapter 1) have been expanded upon and appeared as a conversational piece in *J. Sprinks (2015). Human Factors in Citizen Science, The Ergonomist, 539, pp4-5.*
- The review of citizen science mechanisms and their relation to other human factors practices (chapter 3) has contributed to *R. Houghton, J. Wardlaw, J. Sprinks, S. Bamford, S. Marsh (2016). Martian Factors: A systems ergonomics approach to citizen science, Human Factors in Complex Systems, June 2016, Nottingham, UK.*
- The review of current citizen science platforms, in terms of their task workflow design (chapter 6) has been published as *J. Sprinks, J. Morley, S. Bamford, & R. Houghton (2015). Keeping Citizen Scientists Interested: The Importance of Task Workflow Design, Citizen Science Conference, Feb 2015, San Jose, CA, United States.*
- The model of task workflow design factors, its derivation (chapter 4) and its subsequent application (chapter 6) has been submitted as: *J. Sprinks, R. Houghton, S Bamford, J. Morley, (2016). The Diverse Zooniverse: The Different tasks of Citizen Scientists, Citizen Science: Theory & Practice, under review (submitted Sept 2016).*
- The findings of the laboratory experiment (chapter 8), specifically regarding the self-reported view of the participants, have been published as *J. Sprinks, R. Houghton, S. Bamford, J. Morley, (2015). The Impact of Task Workflow Design on Citizen Science Users and Results, Contemporary Ergonomics and Human Factors 2015, pp371-378, Apr 2015, Daventry, UK.*

- The findings regarding remotely sensed image analysis in chapter 8 have been presented as *J. Sprinks, J. G. Morley, S. Bamford, R. J. Houghton, (2015). The Impact of Task Workflow Design on VGI Citizen Science Platforms, GIS Research UK, Apr 2015, Leeds, UK.*
- The findings regarding the effect of task workflow design on volunteer crater counting analysis (chapter 8) were presented at *J. Sprinks, J. Morley, S. Bamford, R. J. Houghton, (2015). The Impact of Task Design on Citizen Science Results, 2015 AAAS Annual Meeting, Feb 2015, San Jose, CA, United States.*
- A full account of experiment 1 (chapter 8), incorporating parts of the thesis approach (chapter 4) and discussion (chapter 10) have been submitted as *J. Sprinks, J. Wardlaw, R. Houghton S. Bamford, J. Morley, (2016), Impacts of interface design on motivation and performance in citizen science tasks, International Journal of Human-Computer Studies, under revision (submitted June 2016).*
- A full account of the real-world experiment (chapter 9), incorporating the predictive model of classification behaviour (chapter 11) has been submitted as *J. Sprinks, R. Houghton, S. Bamford, J. Morley, (2016). Planet Four Craters: Optimising task design to improve volunteer crater-counting performance, Meteoritics & Planetary Science journal – Special Issue: Issues in Crater Studies, under review (submitted August 2016).*
- Selected qualitative responses of volunteers from chapters 7, 8 and 9 have contributed to a discussion published as *J. Sprinks, R. Houghton, S. Bamford, J. Morley, (2015). Citizen Scientists: The importance of Being needed and not wasted, CHI PLAY '15 Workshop: The annual symposium on Computer-Human Interaction in Play, Oct 2015, London, UK.*

1.7 THESIS STRUCTURE

This thesis consists of four parts. Part I reviews the relevant background literature. Chapter 2 introduces the beginnings of citizen science, the methodologies used, the contributions made and issues raised through the use of non-expert data in the scientific domain. With these issues in mind, Chapter 3 reviews the mechanisms and techniques used by citizen science,

and relates them to approaches and findings from other related bodies or knowledge (such as crowdwork, visual inspection etc.). It considers the large catalogue of human factors research relating to such paradigms, and introduces the analogy of considering citizen science as a form of work. In doing this it frames and considers important concepts such as autonomy, task variety and task complexity – all of which have been shown to influence performance and motivation in other settings.

Part II develops the approach and methodology employed in order to apply the lessons learnt through related human factors research to the citizen science case. Chapter 4 derives a model of task workflow design factors that is central to the thesis, by drawing on the concepts of Job Characterisation Theory and Psychophysical Methods and the implications regarding performance, motivation and engagement claimed in the literature. Furthermore, the chapter establishes the focus of the empirical work and relates it to the principal research question. Chapter 5 introduces the methodology used in order to study the effect of task workflow design on the scientific outcomes of citizen science – relating it back to the conceptual foundations grounded through the human factors work. The Zooniverse is introduced, as is the development of the *Planet Four: Craters* platform in order to study task workflow design in a real world scenario. A mixed methods approach is described utilising hierarchical task analysis, enhanced through questionnaires and expert comparison in order to capture volunteer motivation and performance.

Part III contains the main empirical body of the thesis. The first half acts as an overview of the current position regarding citizen science. Chapter 6 reviews several existing Zooniverse projects, and considers their task workflow design configurations. Chapter 7 considers the opinion of the volunteer, canvassing the community of an existing project (Planet Four) and relating their thoughts and concerns to task workflow design considerations. Based on the findings, the second half directly experiments with task workflow design factors. Chapter 8 describes a laboratory experiment testing different citizen science task workflow design configurations and their effect on volunteer motivation and performance. Finally, chapter 9 takes this approach into the real world, directly investigating the influence of task workflow design on the motivation and

performance of an online citizen science community through the deployment of a purpose-built citizen science platform.

Part IV provides a contemplation of the empirical work of this thesis. Chapter 10 discusses each of the empirical chapters in relation to each other, and relates them back to the broader issues of task workflow, concluding with implications for citizen science platform design. As part of this process a predictive model of volunteer analysis behaviour is presented, in order to inform the design of future citizen science platforms. Chapter 11 ends the thesis with a brief summary – stating the key conclusions of the empirical work in terms of the methodology used, understanding the influence of task workflow design and the implications for the design of citizen science platforms. Finally future work directions to further the field are outlined.

PART I
BACKGROUND

2

THE GROWTH OF CITIZEN SCIENCE

This chapter looks at the beginnings of citizen science, and how its position within academia and society has changed over time. An understanding of the profound impact digital technology has had on the uptake of citizen science, both by the scientific community and the general public, is developed. Importantly, this backdrop illustrates the wide range of scientific disciplines that now utilise citizen science projects, why such an approach is needed, and their effect on scientific research.

A subsequent discussion of the research that has been carried out regarding these projects, both in terms of their design and the communities that have been created, is offered to help place the thesis' contribution within the broader context of research into citizen science.

2.1 THE BEGINNINGS OF CITIZEN SCIENCE

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 in large, organised groups. It is a new term for an old practise, with science originally being practised by self-funded gentleman scientists such as Galileo Galilei and Johannes Kepler (Ede & Cormack, 2012). Prior to the 'professionalisation' of science in the late 19th century most scientific research was conducted by amateurs (as in unpaid), however they were considered experts of their chosen field (Vetter, 2011). Throughout the 19th century, there have been many examples of these amateur experts recruiting non-expert volunteers to contribute to their research. For example, in 1844 German astronomer Friedrich Argelander requested 'friends of astronomy' to submit their observations of variable stars (Grier, 2005), whilst in 1860 George Symons published his first survey of British Rainfall with the help of 500 volunteer

observers (Anderson, 2005) – a number that would rise to over 3400 by 1900. Similar contributions by other non-trained participants have helped create valuable collections of plants, fossils rocks and other ecological specimens (Miller-Rushing, Primack, & Bonney, 2012). The first well-known example of science research being carried out by a 'crowd' of truly amateur participants is the Audubon Society's 'Christmas Bird Count', first performed in 1900, and continuing to the present day. The original event involved 27 observers counting the number of birds they spotted in the North-Eastern United States, and has grown to include many tens of thousands of observers taking part each year across the whole of North America.

By the middle of the 20th century, the landscape of scientific research had changed considerably. The majority of research was now carried out by universities and government-funded bodies, and the need for and participation in citizen science had declined in comparison. Whether this change was for the better or worse has been constantly questioned, the philosopher Feyerabend (1982) stating a need for a *"democratisation of science"*, and biochemist Chargaff (1978) wanting science *"dominated by amateurship"* – 'lovers of' the subject, instead of money-biased bureaucrats.

2.2 Citizen Science in the Information Age

In the latter part of the 20th century, technology has played a big part in bringing another change to the landscape of citizen science. The Internet in particular has allowed amateurs to contribute to scientific research remotely, resulting in a huge rise in citizen science participation. In addition, mobile technologies such as smart phones and tablets have meant that the world has never been more connected, with almost half of the world population[3] able to both access and provide data anywhere at any time. Through these technologies, two roles have arisen that are required of the citizen scientist, *observers* and *classifiers* (Marshall, Lintott, & Fletcher, 2015).

[3] According to www.internetlivestats.com, accessed 12th September 2016

The *Observer* role requires citizen scientists to act as remote data collectors or sensors for citizen science projects, collecting an array of geographical, biological and natural data typically relating to the location where they are based or familiar with. These data are then uploaded to the citizen science website by the observer, at which point experts can remotely analyse data from a larger geographical area or population than would be feasible to cover by other means. This method has been used in a range of scientific research, from environmental studies regarding air and water quality (Dutta et al., 2009; Firehock & West, 1995) to ecological studies mapping plant species and measuring biodiversity (Hawthorne et al., 2015; Joost et al., 2014) and biological studies looking at trends in populations and migration (Barlow et al., 2015; Muratet & Fontaine, 2015 and Lagoze, 2014).

Classifiers take on the role of the 'human' in human computation systems, through the use of specifically designed Virtual Citizen Science (VCS) platforms (Reed, Rodriguez, & Rickhoff, 2012). For the purpose of this thesis, human computation can be described as using the internet population to perform tasks that analyse and interpret complex or contextually rich data to solve difficult problems, that cannot be solved by known computer algorithms alone (Ma et al., 2009). VCS platforms achieve this in an inverse approach to that of the previous method mentioned, with the participant acting this time as the analyst, usually studying data previously collected by the professional research community displayed to them through a website interface. The scientific research addressed by VCS systems is wide-ranging, with participants asked to (for example) classify different types of galaxies in images obtained using large telescopes (Lintott et al., 2011), transcribe historical ships logs and weather readings (Eveleigh et al., 2013), or measure the size of creatures found on images of the sea floor (www.seafloorexplorer.org).

In addition to projects that concentrate on one of the two roles described, a number cross both categories and use volunteers both to collect data and to analyse it once collected. Disciplines including ecology (Dickinson & Bonney, 2012) and astronomy (Hueso et al., 2010) provide a number of examples, combining the access to data collected from a wider range of

locations and instruments with the asynchronous, increased rate of data analysis.

It could be argued that VCS platforms have had the biggest impact in the growth of citizen science, with participants able to take part through a simple online connection. While projects using observers often get volunteer numbers in the region of tens to hundreds, VCS-centred projects can get tens of to hundreds of thousands. One such example, the Zooniverse[4] (see 2.4), has surpassed a million registered users (March 2014) since its beginning in 2009.

2.3 THE CONTINUING IMPORTANCE OF CITIZEN SCIENCE

While it is clear how modern technologies have resulted in a large increase in accessibility and participation in citizen science, a question that arises is: *do scientific communities still need citizen science?* The advantages of citizen science platforms that utilise observer volunteers may be obvious (local area knowledge, geographical/temporal coverage etc.), but the need for classifier type VCS platforms is less clear in modern times, where a huge network of universities and government-funded research bodies exist.

2.3.1 The Advancement of Data Collection

The seemingly relentless advance of modern day technology has not only made the world a more connected place, but has also increased our capacity to collect and store information to an unprecedented level. This has resulted in a flood of data being produced, particularly by increasingly advanced and automated instruments carrying out large-scale surveys. Describing one such dataset, Lahav et al. (1995) states that:

"Classifying very large data sets is obviously beyond the capability of a single person".

Citizen science and VCS platforms can help address this issue through scalability. By enabling large numbers of volunteers to work on a project,

[4] The Zooniverse suite of VCS projects is available at http://www.zooniverse.org, and we shall see later in this chapter its importance to the discipline.

analysis can be completed in much shorter timescales whilst also reducing the reliance on an individual or small team of expert researchers.

2.3.2 Automated Data Analysis

In order to address the issue of too much data, some research has looked into utilising automated systems to undertake analysis. Generally, these systems involve a computer program or algorithm systematically running through each piece of data, picking out and reporting back any relevant result it finds. There has been success with automated analysis (e.g. Statella et al., 2014; Di et al., 2014) – and it will be used when possible due to its well-defined, objective, repeatable and quantitative nature. However, not all tasks can yet be easily automated and often require human-derived training data to achieve high accuracy (Ding et al., 2010), that can deteriorate when the subject being studied increases in complexity – due to the difficulty in creating rules to define it (Barnard et al., 2014).

Another limitation of automated analysis is its rigidity. The results gathered depend solely on the algorithm used and the pre-defined rules it must follow. An advantage of human analysis (and therefore citizen science platforms) is that there is still the opportunity for 'chance' discoveries; unusual results or objects that were not predicted or thought of as important when the experiment was originally defined. Human classifiers will, more often than not, notice or even consciously look for unusual objects, while automated systems can only look for what they have been instructed to (Christian et al., 2012).

The recognition that humans and machines are better suited to different tasks, and the challenge of correctly allocating functions to each in a system is not exclusive to present-day scientific endeavours. Over sixty years ago, Fitts (1951) ignited function allocation research through an 11-point list – highlighting the advantages and disadvantages of both human and machine in an air traffic control scenario. In the proceeding years, Fitt's list has been adapted to inform a range of solutions (Price, 1985; Reason, 1987), and despite varying opinion regarding its validity (Sheridan et al., 1998) the approach persists to inform human-machine function allocation in the present day (Winter & Dodou, 2011).

In the future automated analysis will continue to develop to be able to identify more complex objects and results with a higher degree of accuracy. As of yet however it is not a viable solution to some elements of the big data problem, especially without human collaboration in the form of training and validation.

2.3.3 Outreach & Education

Although this thesis concentrates on the benefits to the scientific community in terms of results and discoveries, citizen science can also have other benefits and impact regarding the volunteers involved. Citizen science projects can be influential in expanding the scientific knowledge and literacy of the volunteer community it creates (Bonney et al., 2009; Cormier et al., 2011), where even low-level participation can result in greater knowledge of the discipline and increased public engagement (Heaton et al., 2014). As such, citizen science programmes are now being considered as a tool for formal teaching practices (Britton & Tippins, 2015; Borden et al., 2012) and as an important medium for science communication between experts and the public (Cormick et al., 2014).

Through increased knowledge of the subject concerned, and the ability to be able to communicate directly with the professional community, there are cases where citizen scientists have been able to directly influence scientific policy (Harris & Russo, 2014) – influencing future scientific goals, and even initiating funding campaigns in order to finance continuing research. It seems citizen science can both educate the volunteer, and also give them the platform for their voice to be heard.

2.4 THE BIRTH OF THE ZOONIVERSE

Two of the disciplines that suffer the most from the 'floods of data' issue explained in section 2.3 are Astronomy and Planetary Science. Since the space race of the 1960's over a hundred missions have been undertaken, studying varying planetary bodies both within our solar system and

beyond. Mars alone has been the subject of 16 NASA missions[5], with more planned in the future, each carrying more advanced instrumentation able to collect data in greater abundance with unprecedented levels of detail. The High Resolution Imaging Science Experiment (HiRISE) of the Mars Reconnaissance Orbiter (MRO) mission alone has collected over one million images of the planetary surface[6], revealing the presence of several geomorphological features similar to those found on Earth.

How best to analyse and process this abundance of collected data has long been an issue for both the astronomy and planetary science communities, amplified by the relatively small number of professional scientists involved in active research. It is estimated that there are only around 10,000 professional astronomers worldwide (Forbes, 2008), who are unlikely to be able to satisfy the demand for data analysis, especially beyond their own research interests, nor arguably is it a good use of their time. This makes the fields of astronomy and planetary science prime candidates for, and adopters of, citizen science and more specifically VCS platforms. One of the first examples of astronomy's use of the online public is the SETI@home project. Launched in 1999, it uses volunteers in a passive capacity, harnessing their computers spare CPU power to help in the search for extra-terrestrial life (Anderson et al., 2002). Harnessing the public more actively, in 2000 NASA piloted a VCS platform asking volunteers to analyse craters found on Martian HiRISE imagery, titled 'NASA Clickworkers'. Deemed a great success (Szpir, 2002), a beta version of the platform is currently in production.

In the following years a number of astronomy-based citizen science projects came and went (Stardust@Home, Digital Access to a Sky Century @ Harvard), with varying success (Nov et al., 2011a), until the launch of Galaxy Zoo. Developed in 2007 by a collaboration including the Universities of Oxford, Portsmouth and John Hopkins, MD it asked volunteers to classify

[5] An overview of NASA Martian missions is available at http://www.nasa.gov/mission_pages/mars/missions/

[6] Imagery can be accessed through NASA's Planetary Data System (PDS) library, available at http://pds.nasa.gov/

galaxies found in images collected by the Sloan Digital Sky Survey[7] (SDSS) by answering simple questions about their appearance (is it round, spiral etc.). The platform perhaps captured the public's imagination more than any other that went before, and within 2 years over 200,000 volunteers had helped to perform over 100 million galaxy classifications (Raddick et al., 2009). The early success of Galaxy Zoo gave encouragement to its developers that it was a viable mechanism for scientific research, resulting in the birth of the 'Zooniverse'.

Created by the Citizen Science Alliance[8] (CSA) in 2009, The Zooniverse was originally a suite of astronomy-led VCS platforms (including Galaxy Zoo), that not only had volunteers studying galaxies but also other targets such as the Sun and Moon. Over time the Zooniverse has grown to include more projects with an increasing number of registered users taking part. This growth has resulted in increased media interest in the platform, culminating in projects being heavily featured in the BBC's annual 'Stargazing Live' broadcast[9] from 2012 onwards. The exposure on such a flagship show has a profound effect, and during peak broadcast times the featured project can receive over a million classifications from volunteers an hour (Angus, 2014).

Subsequently, the Zooniverse is currently the world's leading virtual citizen science platform, with over a million registered participants and a similar number of unregistered taking part in over 20 current projects, that not only involve astronomy but cover a range of scientific disciplines including nature, biology, medicine, history, climate science and humanities (Simpson, Page, & De Roure, 2014).

---

[7] Information regarding the Sloan Digital Sky Survey and its data can be found at http://www.sdss.org

[8] An overview of the Citizen Science Alliance and its structure is available at http://www.citizensciencealliance.org/

[9] Information on the Stargazing Live broadcast and featured content is available at http://www.bbc.co.uk/programmes/b019h4g8

2.5 THE CONTRIBUTION THAT CITIZEN SCIENCE RESULTS MAKE

Citizen science and VCS platforms can clearly have a profound effect on the rate that data can be analysed and results collected, with successful projects able to complete in weeks the work that would take professional scientists several years (Pinkowski, 2010), but is it any use? Can volunteered results supplied by untrained analysts be used to further scientific research in the same way that expert analysis can?

2.5.1 Data Catalogues and New Discoveries

Through the collection of data on VCS platforms, several data catalogues have been produced and subsequently published into the scientific domain. It is claimed that these catalogues are comparable to existing expert counterparts (Simpson et al., 2012), are larger and more complete (Lintott et al., 2011), and in some cases contain added detail including bias and accuracy measures (Willett et al., 2013). Citizen science data catalogues have also been used to correct existing datasets created by automated mechanisms where limitations exist due to complexity (Hennon et al., 2014). Furthermore, through using citizen science data, research conducted by the scientific community has resulted in a number of publications. Data collected by the Zooniverse alone has been associated with over 70 papers[10] in a number of scientific fields.

Beyond the systematic task of collecting data and creating catalogues, VCS platforms have also been responsible for discovering new scientific features and phenomena. These discoveries have either been a main target of the platform, and as such help define its success (Fischer et al., 2012), or have been an 'added bonus' – found serendipitously by human classifiers in a way automated systems as of yet cannot (Lintott et al., 2009; Cardamone et al., 2009; Boyajian et al., 2016). Perhaps due to the unplanned nature of these discoveries (in the eyes of the platform developers at least), they are often not made through the main mechanism of the website, but

[10] A full list of Zooniverse publications is available at https://www.zooniverse.org/publications

collectively through community forums and threads associated with the project (Tinati et al., 2014).

2.5.2 Validating and Reducing Citizen Science Data

Although citizen science data has proved a valuable resource for scientific research, there is a caveat attached. Due to the mostly untrained expertise of the volunteer and the unsupervised nature of the collection, combined with a lack of a 'ground truth' comparison, the data at the very least needs to be validated by an expert, and often reduced and 'cleaned up' to remove errors. Depending on the task asked of the volunteer (be it acting as an *observer* or *classifier*), different sources of error could occur. In the case of the classifier role, these can include simple detection and classification mistakes (false-positives etc.) that can become more prevalent as the subjectivity/subtlety of the decision increases (Swanson et al., 2016). More associated with the observer role, environmental conditions could add a bias to the data, caused by the method of data collection or by location-based demographic factors related to the citizen science community (Ward, 2014). While it is widely assumed that the majority of errors are due to legitimate mistakes, there have been examples of malicious behaviour by volunteers (Arcanjo et al., 2014), a risk with un-policed online participatory platforms.

In order to reduce the amount of unusable data, and therefore the demands on the scientist to identify it, various techniques have been considered. Regarding observer data, different models of statistical testing have been shown to increase the efficiency of data collecting tasks (Arcanjo et al., 2014), whilst with classifier data variations in data sampling design can help remove biases that might exist (Chambert et al., 2014). Tackling the issue at source, other approaches involving both classifier and observer data have attempted to learn from the errors and biases found, in order to better understand the training needs of VCS volunteers for future projects (Nerbonne & Vondracek, 2003; See et al., 2013).

Despite the number of scientific publications that have resulted from citizen science data, and the varying methodologies being developed to make the data increasingly accurate, sometimes a citizen science project is destined to fail (Kaufman, 2014). It is therefore imperative that citizen

science developers and science teams consider the needs of the volunteers and the associated errors and biases that might arise at the design stage, both to reduce the data demands on the scientist and in extreme cases prevent the total failure of no usable data being collected.

2.6 CITIZEN SCIENCE AS A DISTINCT FIELD OF ENQUIRY

Despite virtual citizen science being a relatively new form of work, there has been a growing field of research considering citizen science practices in their own right, beyond the scientific problems they address. Particularly, studies involving interface Human Computer Interaction (HCI), platform functionality and public engagement have grown in number, contributing to a growing body of citizen science scholarship (Jordan et al., 2015).

2.6.1 Virtual Citizen Science Typology & Functionality

Early research regarding virtual citizen science defined platforms simply by design and functionality. By considering VCS platforms as system assemblages, a group of functionalities and technologies used together to achieve a goal, Prestopnik & Crowston (2012) compared the factors that contribute to their selection. It was found that functionalities were selected based on higher-level concerns, such as the science goals and participant motivations, while technologies were chosen for more pragmatic reasons such as cost or availability. New VCS platforms would benefit more by looking at participant satisfaction, informed by well designed interfaces and support features, rather than motivation, which is predominately affected by the scientific goal. Further investigating these differences, Reed et al. (2012) compared the usability between several Zooniverse projects. Tailoring five usability categories inspired by the Microsoft Usability Guidelines (MUG), each Zooniverse platform was scored by two experts in website usability for each of the categories: content, ease of use, made for the medium, aesthetics and emotion. It is hoped that follow up research will be able to relate the usability of VCS platforms to other user experience factors.

2.6.2 Gamification

Gamification is a relatively new form of design process based on the use of game design elements in non-gaming contexts, for example *"piggybacking game play to solve human information tasks such as tagging images"* (Deterding et al., 2011). While using this technique to solve information tasks can be said to be 'strong gamification' based on the idea that ludic elements provide an intrinsic motivation to play a game (i.e. it is inherently fun to participate), an alternate 'weak gamification' also exists where extrinsic motivation cues such as rewards, feedback, scores and competition are combined with non-game work tasks. Both forms have been integrated within virtual citizen science platforms, with varying success.

The field of biological science has a number of examples of VCS projects that incorporate the 'strong gamification' approach, where what were very complex research problems have been re-imagined as a type of online multi-player game. This has proved very successful, the collective problem-solving abilities of the crowd being harnessed to combine a wide range of approaches that accelerate the understanding process (Curtis, 2014). The 'weak gamification' approach has been used by VCS platforms that often involve simpler, more repetitive tasks, in order to make them more enjoyable and therefore motivating and sustaining participation (Eveleigh et al., 2013). In fact there have been examples of volunteers creating their own games beyond the design of the project (Greenhill et al., 2014) in order to add interest to classification systems they have found to be dull and repetitive (Prestopnik & Crowston, 2011).

While the argument for the use of 'strong gamification' within certain VCS projects is compelling, allowing previously overly complex research problems to be addressed by non-expert volunteers, it is less clear-cut regarding 'weak gamification'. Evidence suggests it can help sustain the engagement of some volunteers, motivated by both the competition aspect and the ability to see their progress, while other volunteers are similarly put-off by the 'race to the top', slowing their contribution and in some cases leaving the platform altogether. Furthermore, beyond the established volunteer community, there is little evidence that weak gamification elements are necessary to attract new users (Iacovides et al., 2013). Even

less is known regarding the effect of gamification, strong or weak, on either the quantity or quality of the volunteer results, and it is clear that this is a research 'gap' that needs addressing in the future – especially due to the complexities and difficulties that exist when attempting to design a 'fun' experience (Knaving & Björk, 2013).

In a worse 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 favour of more 'serious' interface designs (Bowser, Hansen, & Preece, 2013). Even if a balance could be struck between the motivational benefits of gamification and any potential downside, it is a complex undertaking and would require a significant amount of effort and consideration (Rojas, Kapralos, & Dubrowski, 2013) that could at least in part negate the point of a citizen science solution – to reduce the workload of the scientist. As such, gamification is not a theme of this thesis, with the focus instead being on design factors that have the potential to result in proportionally greater gains compared to the effort required.

2.6.3 Citizen Scientist Motivations

One of the earliest surveys into the motivations of VCS platform participants was conducted by (Raddick et al., 2009). Through the interviews of 22 participants from the platform Galaxy Zoo, a Zooniverse project asking people to classify galaxy types shown in imagery gathered from a variety of telescopes, a list of 12 motivation categories were defined that included scientific contribution, learning, discovery and being part of a community. A follow up study led by Raddick et al. (2013), expanded the survey to 11,000 Galaxy Zoo participants. Volunteers were asked to report their motivations using the 12 categories defined previously, and were given freedom to add others if they wished. It was found that 40% of respondents marked 'contributing to science' as their primary motivation to participation. Reed et al. (2013) expanded this type of study into participant motivation across several different VCS platforms of the Zooniverse group, across several different scientific disciplines. 199 registered users of the Zooniverse completed a survey asking them to report freely their motives for participation, and their replies were

grouped into three main factors; social engagement (being involved with a like-minded group), interaction with the website (ease of use, aesthetics) and helping (feeling important, making a contribution). Again specifically regarding the participants of Galaxy Zoo, Mankowski et al. (2011) studied responses to a forum question asking 'why are you taking part?'. Through the analysis of 200 posts in response, it was found again that one of the key motivations is contributing to a larger scientific project. In addition to this however a second motivation was discovered much more uniquely associated with the platform itself, the access to current professional astronomical imagery. This result was taken further, the study suggesting using the most aesthetically pleasing imagery as a positive reinforcement to maintain participation rates over time.

In a comparison of paid and volunteer activity within a citizen science task, using the Zooniverse's Planet Hunters project, Mao et al. (2013) noted that while different incentives based on time, task or classification could alter the profile of user behaviour in a pragmatic fashion (where speed/accuracy trade-offs could be manipulated), volunteers and paid contributors achieved a similar level of performance. The one difference that did exist was that volunteers tended to spend significantly longer on a task – perhaps unsurprising as paid-for participants were paid per each unit completed.

The general design implications surrounding the findings regarding citizen scientist motivations are that a citizen science project should involve an actual, valid science case and that this should be communicated with users who in return receive credit (either within platform or co-authorship of papers etc.) for their contribution. Unsurprisingly, citizen scientists want to be scientists at some level making contributions that further science, rather than seen as the Von Ahn style view of 'human processing units' (Von Ahn, 2005; Davis et al., 2010), and the conduct of the project should respect this.

2.6.4 Volunteer Behaviour & Patterns of Participation

It is widely accepted that the majority of virtual citizen science projects – in common with most Internet services – have a user profile (number of volunteers against time) taking the form of a power law that approximates to the Pareto distribution (figure 2.1). In practical terms this means that

the typical citizen scientist only visits the site once or twice at most, but those who remain engaged end up over time performing the majority of the work or tasks.

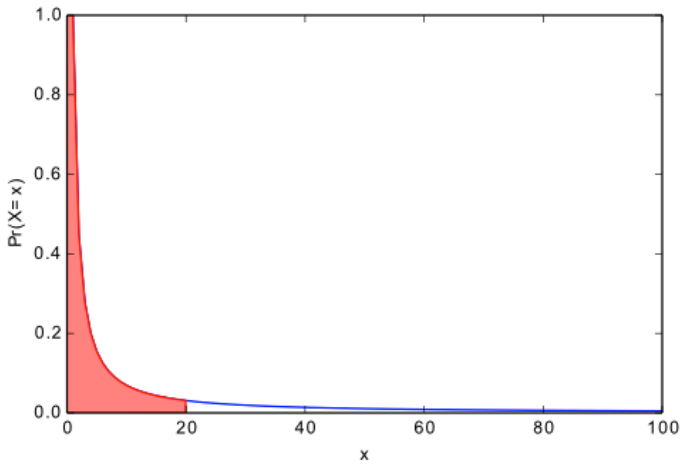


Figure 2.1: The Pareto Distribution

Considering two Zooniverse sites, Galaxy Zoo and the Milky Way Project, 70% of volunteers visited only one day, contributing, in total, around 20% of classifications (Ponciano et al., 2014). A confirmatory correlation of 0.7 was found between time spent and visits, suggesting that committed volunteers are responsible for the majority of classifications. The existence of (at least) two distinctive types of citizen scientist, the committed volunteer who visits many times and the one-time visitor or 'dabbler' (Eveleigh et al., 2014) is in agreement with studies concerning other online volunteer-based systems such as Wikipedia and open-source software development that describe a "motivational arc" (Crowston & Fagnot, 2008).

Rotman et al. (2012) looked at how volunteer motivation evolved depending on the length of participation. Conducting a survey of 142 ecological VCS platform participants, it was found that volunteers are motivated by a number factors affected by personal interests, attribution and acknowledgement. These factors dynamically change over time, throughout the volunteer's work on the project. Two key moments were identified; the volunteer's first encounter with the platform, and the point where the project or task involved changes, and considering these shifts during platform design would result in a better relationship between scientists and volunteers. Rather than studying a particular platform, or a

range of disciplines, Nov et al. (2011b) looked at the difference of participant motivation depending on their level of interaction with the task to be performed. Comparing motivations from a high interaction level (active-participation) and a low interaction project (passive participation), it was found that the active participants are more strongly associated with collective motives (the science goal) while passive participants were motivated by identifying with the community. Furthermore, it was suggested that this relation between task interaction and motivation implies a need for more dynamic VCS platforms, which allow the participant to start on low-level interaction tasks and progress to higher-level tasks over time.

An alternative approach to this two-level perspective of user recruitment is to design the platform, and the associated tasks required to be performed, specifically for the casual 'dabbler' that makes up the majority. The argument behind this approach is that as they are the largest user group that are associated with a VCS project, it could pragmatically be sustained on a 'leaky bucket' basis, ensuring their contributions are realised to the full (Wiggins & Crowston, 2010; Eveleigh et al., 2014). Although casual volunteers only perform 20% of the work on a project, it could be argued this is due to the platform setups not being designed for the casual user, and a more sympathetic version could extend their participation by at least 'a few more' classifications. The sheer size of this user group would then lead to a considerable increase in the rate of tasks completed. By breaking down tasks into smaller units of work that do not require a large investment in time and effort, and providing relevant feedback, it could *"make 'dabbling' easier and help dabbers to feel that their contribution is valuable and valued"* (Eveleigh et al., 2014).

Beyond considering volunteer behaviour and participation regarding the scientific task and goals of VCS projects, a growing body of research has studied participation regarding the communities and forums that have been created alongside the main interface, as described in section *2.5.1.* As with participation in the main scientific task, user profiles again take the form of a power law approximating to the Pareto distribution, with a highly engaged few contributing the most 'posts' and communication. Furthermore, a 'super group' of volunteers exist; whose contributions both

to the main scientific task and the community is a factor of 10 higher than the average (Luczak-Roesch et al., 2014). The behaviour of volunteers highly engaged in such forums has evolved over time, from beginning with knowledge acquisition activities (completing the scientific task), to sharing this knowledge with other volunteers through the associated forum, and finally discovering their own 'roles' and responsibilities within the community (Jackson et al., 2015). In essence a citizen science 'culture' has emerged (Newman et al., 2012) where more experienced volunteers act as mentors for new participants when guidance is sought, ensuring they follow the established 'norms' of performing the scientific work. The social interaction provided by this process acts as a motivational factor for participation beyond scientific contribution (Van Den Berg, Dann, & Dirkx, 2009).

In general there is a positive relationship between the number of classifications completed and contributions to discussion forums (Tinati et al., 2014). However, there exists a 'super group' of volunteers that have contributed highly to both (over 6000 classifications and 140 forum posts) – and once at this level volunteer task completion behaviour can be split into 3 categories; sporadic classification, consistent classifying and very infrequent classifying. While at first glance it might be seen as a drawback to lose scientific contributions from such highly-engaged volunteers, it could be argued that they perform an equally important role of offering support, advice and information to new participants through the forum – a task that initially would have been the responsibility of the developers and science team who created the project.

2.7 THIS THESIS' DEFINITION OF CITIZEN SCIENCE

At the beginning of this review, Hand's (2010) rather broad definition of citizen science was quoted. It has subsequently been shown throughout this review how many different forms citizen science can take, from the self-funded gentlemen scientists many centuries ago working alone, to the mass-participation online projects where up to hundreds of thousands of volunteers work towards a shared scientific goal. The variations in the type of role citizen scientists can take have also been demonstrated, either being the *observer*; contributing to the collection of data at a greater rate and

breadth than possible by scientists alone, or the *classifier*; analysing large datasets previously collected by the scientific community. This large variation in citizen science endeavours is beyond the scope of investigation for a single research thesis, and as such the definition of citizen science for this work is as follows:

- Interaction with the volunteer is through a Virtual Citizen Science (VCS) system (Reed et al., 2012), accessible through a web browser independent of location and with no need of specialist instrumentation beyond access to the Internet.
- Volunteers take on the role of a *classifier* (Marshall et al., 2015), analysing data already collected by the scientific community rather than contributing new datasets.
- The volunteer community work towards a joint scientific goal determined by a science team on a platform purposely designed for the task, as opposed to more serendipitous use of the data/analysis collected (Freitag, 2016).

By focussing on VCS systems, a number of environmental and geographical factors that can effect other types of 'in the field' projects (Ward, 2014) which are hard to control can be negated. Likewise, by concentrating on citizen science that only involves the *classifier* type role, issues regarding the types of datasets involved and the environmental conditions and equipment used to collect them can also be controlled. By narrowing the field of citizen science in this way, and concentrating on a specific scientific goal and associated task, this thesis can concentrate more on design-led factors that may influence the outcomes of a CS project.

2.8 SUMMARY

Sections 2.1 and 2.2 explain the history of citizen science, and its subsequent rise to prominence due to the interconnectivity afforded by the information age currently taking place. Specifically the importance of virtual citizen science platforms is outlined, progressing participation rates from the hundreds and thousands to millions of volunteers. Moving on from the how, section 2.3 outlines the importance of citizen science to the scientific community, and in doing so underpins the broader interest of this thesis beyond the disciplines of human factors and planetary science. The

introduction of the Zooniverse platform in section 2.4, that encompasses a range of different scientific research and has to date over 2 million participants, demonstrates both its success and its position at the forefront of the citizen science movement. This in turn acts as an explanation for introducing the Zooniverse as the research context for the thesis, especially regarding the studies carried out in chapters 6, 7, 8 and 9.

In considering the contribution that citizen science results make to the advance of science, section 2.5 highlights some of the consequences of using volunteered scientific data. Limitations in terms of data accuracy and coverage become apparent, as does the effect of poor recruitment. As a result there is a growing movement considering how to improve and validate volunteered data – one of the central phenomenon under investigation in this thesis. Based on the growth, importance and subsequent issues raised regarding citizen science, section 2.6 outlines how the discipline has given birth to a distinct field of enquiry in its own right. While early research has to date concentrated on interface HCI and platform functionality, and its influence on volunteer motivations and behaviour, there is a growing interest in scientific task design and how to cater it towards specific volunteer types. Finally, section 2.7 narrows the definition of citizen science for the purpose of this thesis to VCS systems involving *classifier* type tasks.

Having established the general background of this thesis in terms of its citizen science underpinnings, online communities, volunteered scientific data and its application to address some of the data validation issues, the following chapter turns its attention to the mechanisms and associated disciplines involved in the derivation of a virtual citizen science project. Specifically it will discuss human factors considerations that have been researched regarding these mechanisms, which have yet to be directly applied to the citizen science example.

3

CITIZEN SCIENCE DESIGN AND METHODS

Following on from the previous chapter, it can be concluded that while citizen science, and virtual citizen science platforms, have become a tool increasingly used by the scientific community and have contributed to furthering knowledge, there are issues that arise when using volunteered scientific data. These issues are intrinsically linked with the volunteer, and as such a growing area of research is investigating their motivations and behavior. Despite this, there has been relatively little attention paid specifically to human factors issues. This is perhaps ironic given the importance of the 'citizen' part of the discipline, 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).

It is worth considering that virtual citizen science platforms involve processes, mechanisms and methodologies that have been used historically in other similar systems, and as such there is a wealth of research regarding their design and implementation. This chapter formally introduces these mechanisms, and presents related work that has studied various human factors considerations including their effect on motivation and performance. The discussed techniques, approaches and findings regarding these mechanisms inform the methodological approach of this thesis and the development of a model of Task Workflow Design detailed in chapter 4.

3.1 CROWDSOURCING

Crowdsourcing, as the name suggests, refers to various approaches in sourcing work or information from a crowd. It was originally considered a

method of outsourcing a task outside an organisation, taking a task traditionally performed by an employee and outsourcing it to an undefined and generally large group of people in the form of an open call (Howe, 2006, 2008). In practice, crowdsourcing has since become associated with using the Internet to mediate an open call and the interactions of a large community. Examples include the development of open-source software or information (e.g. Wikipedia), seeking product development suggestions and assessing market interest (e.g. ideas.lego.com) through to obtaining funding for nascent projects (e.g. Kickstarter). Arguably Crowdsourcing can be seen as a broader term that captures a wide range of crowd-based activities mediated via the web, that therefore encompasses citizen science and virtual citizen science platforms.

As a range of different types of crowdsourcing models now exist, it stands to reason that they differ in terms of their design. Criteria such as incentive mechanisms, task assignments, required qualifications and reputation, user demographics and automation have been found to vary across such platforms, depending on their goals and aims (Vakharia & Lease, 2013). The design of crowdsourcing platforms should be carefully considered however, to ensure crowd work is complex, collaborative and sustainable, rather than becomes a form of assembly-line piecework (Kittur et al., 2013). Factors including workflow, task design, motivation and reputation are all important in deciding the success of a project, for instance by simply having knowledge of workers reputations and assigning tasks accordingly, system performance can be greatly increased (Tarable et al., 2014). Similarly, by considering microtask continuity in workflow design, gains can be made as large delays between tasks and contextual changes in task type can result in a slowdown of completion time (Lasecki et al., 2014). The introduction of micro-breaks however can improve worker retention rates and increase engagement over longer (over an hour) periods (Rzeszotarski et al., 2013).

As a form of crowdsourcing, citizen science projects can directly learn from the research that has been conducted regarding the performance of such systems, and the effect that various criteria and design factors can have.

3.2 PSYCHOPHYSICS & PSYCHOPHYSICAL METHODS

Psychophysics can be described as the study of the relationship between stimulus and sensation (Gescheider, 1997). Originally derived through Fechner's *Elements of Psychophysics* in 1860 (see English translation - Fechner, 1966), the field of study provides both methods and theory for the measurement of sensation – providing psychology with the basic implements to study the mind. In the following years, psychophysics has continued to play an important part in experimental psychology, a key development being the theory of signal detection. Originally conceived through research into radar operation (Peterson, Birdsall, & Fox, 1954), it provides the means to quantify the ability to discern between information patterns (the stimulus) and random background patterns (known as 'noise'). This advance has broadened the applicability of psychophysics to practical fields beyond the original aim of sensory measurement. Subsequent work has both considered operator performance (Nevin, 1969) and informed the improvement of information collection and presentation techniques. The field of optics has particularly relied on signal detection theory, deriving adaptive methods to counteract the noise effects of light scattering and distortion (Hardy, 1998). As such, it has played a key role in the advancement of disciplines such as astronomy (Lohmann, Weigelt, & Wirnitzer, 1983) and those relying on remote sensing techniques (Andrews, 2004).

Citizen science, and VCS platforms, often require the volunteer to view and annotate in some way or another a set of remotely sensed imagery. While in the case of citizen science this condition has appeared opportunistically due to science case requirements, substantial psychophysics literature exists regarding explicit evaluation techniques. In order to measure both perception and performance when carrying out such work, a set of tools exist known as psychophysical methods. They can be used to both assess how a participant is performing, and also to define the characteristics and formatting of the display or image. Observers have to carry out a set of different task types, and make a number of associated judgements, many of which bear resemblance to the tasks and judgements asked of VCS participants. These tasks include (Pelli & Farell, 2010): detection/threshold (is a stimulus present/identifiable), discrimination (the difference between

two stimuli) and matching (adjusting an attribute of two stimuli until they are equal). Corresponding judgements such tasks require the observer to make include 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). There has been a large body of research looking into how the combinations of these tasks and judgements affect observer performance, which could be used to inform VCS design.

3.2.1 Threshold & Detection

One of the main tasks many VCS platforms ask of the participant is to detect if a certain feature or object is present or identify a feature in an image. There has been a large body of research looking at variability in the threshold contrast difference required to detect stimuli, across different judgement types and observer factors. Thomas (1985) compared the threshold required across two tasks, detection and identification. Asking the participants to complete the two tasks using a number of different methods (rating, forced choice etc.) it was found that there was no significant difference in the sensory cut-off threshold, and that accuracy in completing both tasks is more closely related to the signal strength of the stimuli or feature being observed. Observer uncertainty also plays a big part in detection. Pelli (1985) defined an uncertainty model for visual detection, assuming that uncertainty exists when an observer is confronted with any stimuli, and they choose what they define as the likeliest. Although not tested on human participants but relying on simulation data, it was found that this uncertainty is a power function of contrast, and also predicts the effect of other factors such as noise and near-threshold contrast stimuli.

Considering the experience of the observer, Jäkel & Wichmann (2006) tested 'naïve' participants using both detection and discrimination tasks. Using a number of different spatial judgement combinations, participants were asked to either *detect* or *discriminate between* stimuli that appeared on a 3x3 cell grid displayed on a CRT screen. They had to make an 8-Alternative forced-choice (8-AFC) judgement (possible stimulus locations being every grid cell apart from the centre), a 4-AFC (the four corner cells),

or a 2-AFC (the left and right cells from the centre) judgement. It was found that 4-alternative forced-choice methods resulted in the best threshold levels for detection tasks, in disagreement to the accepted 2-interval forced-choice method accepted for experienced observers. It was also found that 4-AFC resulted in better efficiency, with the tasks being completed over 3 times more quickly. Interestingly, the results were the complete opposite for discrimination tasks, implying an influence of the type of task being performed. Sometimes it is not the differences in observer or task that varies detection threshold levels, but differences in the feature itself. Studying the legibility of different letters, Elliott, Whitaker & Bonette (1990) looked at the results of almost 500 threshold-sensitivity measurements taken during optometric practise. It was found that the success rate of identifying 2 out of 3 grouped letters varied from between 67-97% depending on the letters included, hence suggesting some letters seem harder to identify than others.

3.2.2 Matching & Scaling

Other tasks that are frequently required of VCS participants are to either match two images in terms of some attribute value (brightness, feature density etc.) or to rank imagery based on a pre-defined scale. Research exists that considers the factors that influence human perception in terms of image matching and scaling, and as such could be of benefit to the citizen science and VCS platform community.

Mojsilovic, Hu & Soljanin (2002) derived a statistical method to identify and extract perceptually important colours for image matching. By comparing the statistical results for a set of interior design images with the opinion of 15 participants asked to match them, good agreement was found in comparison to other methods, therefore suggesting that the colours identified by the method are more influential in regards to human perception of image matching than others. Incorporating both colour shading with image content, Rogowitz et al. (1998) looked at the results of 15 participants ask to match by similarity a total of 97 digital photographs, incorporating a range of colours, viewing distances and content. Although good agreement was found across the participants, the results were not symmetric. For instance when looking for images similar to a nature scene,

nature scenes including children were selected. However, when looking for images similar to nature scenes including children, images including children were selected and not other nature scenes. The study concluded that although the overall colour of an image is an important factor, content dimensions such as 'human versus non-human' or 'natural versus manmade' also play a part when judging the similarity of imagery.

Looking at magnitude scaling, and how it is influenced by context, Marks (1988) asked participants to judge the loudness of different sounds, played at different frequencies. It was found that if you kept the context (in this case frequency) constant, the resulting magnitude scaling had good agreement across the results. However, if the context is varied (from 500-2500hz) there is a shift in the magnitude scaling depending on the proportion of presented sounds in each context. It was concluded that methodologies that use magnitude scaling should only compare stimuli whose subjective context are constant.

In one of the few studies directly involving the citizen science case, Hutt et al., (2013) performed a comparison of 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 vs. 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 expert observers in terms of overall quality.

3.3 VISUALISATION FACTORS

Another area of research that could directly inform VCS platforms concerns factors regarding the images themselves. The size of the image displayed to the user, its spatial resolution and attributes such as brightness and content have been found to affect analytic performance, as well as technical specifications regarding the display equipment used.

Hodgson (1998) demonstrated that changing something as simple as the window size in which the image is displayed can have a dramatic effect on classification accuracy. By changing the size, from the previously defined ideal of 9 by 9 pixels derived for automated classification systems to a

much larger 40 by 40 area, accuracy increased from a range of 32-54% to 88-100% when classifying urban land use. Considering a more ecologically based land-use task, Hodgson presented areas of Denver to trained analysts at different window sizes ranging from 10 by 10 pixels to 100 by 100 pixels, again with performance increasing with window size. This can be explained in terms of larger windows allowing more top-down reasoning:

"In normal circumstances, the photo interpreter may proceed from known to unknown parts of a photograph, thereby using information gained from one part of a scene in classifying another… this calls into question the notion of using small windows, regardless of the classification algorithm… if the human cannot correctly classify small sub images, then should we expect an automated classifier to do so?" (Hodgson, 1998).

The spatial resolution of imagery can play an important part in image analysis, and the requirements can vary depending on the subject being studied. For instance, hazard practitioners have suggested certain guidelines for disaster emergency response. For most interpretation tasks, 0.25 – 2 metres should be used (Schweitzer & McLeod, 1997), for assessing damage to housing, utilities and roads, a resolution of 0.3 – 1 metre is needed (Cowen & Jensen, 1998) and for identifying roof damage, a resolution of 0.2 – 0.3 metres is required (Bolus & Bruzewicz, 2002). Taking a more cognitive approach, Battersby, Hodgson & Wang (2012) evaluated the performance of participants with varying image interpretation experience. Asking them to analyse imagery of several different resolutions, it was found that a resolution of 1.5m was the coarsest size for the interpretation of structural damage. Display resolution, although not controllable by the VCS system, should also be considered due to the numbers and diversity of citizen science participants.

Ziefle (1998) experimented with three different CRT display resolutions. It was found that participants using the higher resolution display had lower fixation times, faster reaction times and as a result much lower fatigue levels. It was concluded that the use of higher-resolution displays would result in making prolonged on-screen activities more comfortable. Further studies (Hansen & Essock, 2012; Essock et al., 2004) consider the influence of image attributes such as colour rendering and brightness on human

perception. Assessing the performance of participants while studying night imagery of natural scenes, it was found that type of scene content, type of colour rendering, the distribution of magnitude and image amplitude all play roles in determining perceptual performance. Although not defining any direct relationships, it is argued that further research should involve testing these factors against a number of differing perceptual tasks, and future image analysis applications should benefit from their consideration.

3.4 VIGILANCE

If we accept many VCS platforms require the volunteer to perform tasks that are repetitive in nature, it is useful to review what is known of performance, and particularly visual performance, over extended periods of time. The classic approach to this has been to consider the problem of vigilance, the degradation of performance over the course of a period of time (Mackworth, 1948). This work was originally undertaken in the context of the Second World War, due to concerns regarding the rate at which radar and sonar operators missed contacts. It is generally agreed that there exists a fairly reliable vigilance decrement: *"the deterioration in the ability to remain vigilant for critical signals with time, as indicated by a decline in the rate of the correct detection of signals"* (Parasuraman, 1986). This decrement is believed to take effect after approximately 15 minutes of attention, although this timespan can reduce under conditions of high demand and/or the signal strength of the target feature is relatively low. This becomes particularly relevant to VCS systems when considering volunteers who contribute more (see the Pareto distribution, figure 2.1) and spend more time on the site, or when features to be detected by the volunteer occur relatively rarely. For instance, McCarley et al. (2004) found that airport security scanners missed 15% of hunting knives, even when directly fixated upon and 44% that appeared in a location not fixated upon.

Several possible responses to the vigilance phenomenon have been considered, with varying applicability to the citizen science example. Training the user, which effectively reduces the workload of the task, has been shown to reduce vigilance decrement but this tends to be of an extended type deemed unsuitable outside of a workplace (Koller, Drury, & Schwaninger, 2009). Similarly, "years of analysis experience" (Curran et al.,

2009) can have the same effect, but this does suggest that long-term committed users will show better vigilance performance compared to newer or casual volunteers. Another approach might be to pre-select observers, as some personality traits that can be tested for correlate with resisting vigilance decrement (Shaw et al., 2010) while it has also been found that younger observers tend to be more resistant (Parasuraman & Giambra, 1991; but see Berardi et al., 2001 who found no age related pattern). However, given the democratic, open nature of citizen science this is unlikely to be appropriate, especially in the early stages of the project where the make-up of the volunteer community is unknown.

Changing tack, instead of manipulating the observer group it is possible to artificially manipulate the event occurrence rate with the addition of false signals. This technique has often been used in critical detection tasks, such as airline baggage screening or X-ray interpretation, i.e. a baggage screener is statistically unlikely to detect many real-case firearm offences on any given day but will be shown and expected to detect numerous controlled firearm examples in imagery over the course of a working shift (Durso, 2007). It is open to debate whether such an approach would be suitable for a VCS task, as while it may well be possible to apply such false signals that in turn improve performance, it could be detrimental to the volunteers' perception of contribution. Communication is key, as there are circumstances in which this could be seen as actively misleading, for instance if a volunteer believed they have discovered their 'own' new feature for example. A way around this might be to not use artificial foils, but to incorporate selections from a set of imagery that have already had features detected and annotated on them, and use these trials for output agreement purposes. The general indication found that higher workload can reduce the onset of vigilance decrement can be taken as further reinforcement that good interface design and clear task design that reduces extraneous cognitive load are key to improving results from participants.

Alternatively, it could be true that the best approach is to remove the requirement for vigilance altogether. Defining science cases that regularly require actions by the user, permitting users to shift between different tasks and actions, and allowing breaks can reduce the monotonous nature of the work (As found by Rzeszotarski et al., 2013 in section 3.1). In

contrast Lasecki et al. (2014) demonstrated the advantages of task continuity by combining the motivating effects of task variety with the research concerning vigilance, where a more complex scenario emerges that suggests task design, variety, feature density and breaks all need to be considered to maximise performance gains and minimise losses. It seems a trade-off exists between the operator becoming more skilled and getting bored through the lack of challenge – with the timing of this transition being key.

3.5 OPERATOR VARIABILITY

One area of research that could be of great use to VCS platforms, due to their reliance on 10,000's of online participants, is that of operator variability. Several studies (Powell et al., 2004; Leckie et al., 2003; Zhou, Schwarz, & Cadenasso, 2009) have compared the results of different professional operators performing the same remotely-sensed image analysis task, finding that their results differed considerably. In an attempt to solve this issue, Zhu et al. (2000) used two operators to perform the same analysis on a number of images, using a third to provide the absolute result when there was disagreement. It was found that over 30% of the images had to be referred to the third operator, highlighting the issues of operator variability while also raising another that using multiple operators might not always solve the problem. Sarmento et al. (2009) looked at the relationship between operator confidence and accuracy, suggesting that low confidence would be associated with low accuracy and hence errors in the analysis. It was found that the example classification confidence tended to be low, hence highlighting the variability and uncertainty with image interpretation. Another variable between operators can be defined as their 'strategy' for undertaking the task. Differences in these strategies can be down to the relative experience of the operator (Hoffman, Markman, & Carnahan, 2001), but different strategies have also been shown to exist between operators of the same experience, depending on how they use additional data associated with the image. Some operators tend to gather as much information as possible to help with the analysis, while others like to keep an 'open mind', and only use the image itself as a guide (Scepan, Menz, & Hansen, 1999).

Perhaps bearing an even stronger relation to VCS platforms, rather than studying expert or professional operators of remotely sensed image analysis Van Coillie et al. (2014) looked at differences between online participants. Differences in accuracy were measured across a range of factors, both human (demographics, experience, motivation etc.) and external (tiredness, quality of screen, levels of distraction etc.) for a range of different interpretation tasks. It was found that participant accuracy varied considerably across the tasks performed, averaging out at ~80%. Furthermore it was shown that 26% of this difference could be attributed to human factors, rising to 30% when including external factors. It was concluded that participant performance was mainly affected by demographic and personality factors, and to a lesser extent external factors – and this knowledge could lead to a way of identifying individuals with the skills and abilities required to perform image analysis tasks at a high accuracy for long periods of time.

3.6 CITIZEN SCIENCE AS A FORM OF WORK

Although normally presented to the user as a great opportunity to witness images of phenomena never seen before, contribute to the advancement of humanity, or make brand new scientific discoveries, it could be argued when stripped back to the bare mechanism the use of VCS platforms is a type of 'work'. Participants are often asked to do the same task or tasks repetitively over a seemingly never-ending number of images, in an almost 'data entry' like manner, for no financial reward. There is an analogy between this scenario and that proposed by Taylorism (Taylor, 1911) concerning the mechanisms of industrial work, including the fractionation and atomisation of tasks, the most well known application being the rise of Fordism in the 1920s – originating on car production lines. At this time job satisfaction amongst workers was known to be at a low point, leading to a drop in motivation which ultimately impacted on the quality and quantity of work produced. It became clear that the assumption that motivation at work was powered by financial award alone was unsafe (Trist & Bamforth, 1951).

This can be expressed as what are known as Theory X and Theory Y management approaches (McGregor, 1960). Theory X hypothesises that

workers are inherently lazy and only work for financial reasons, requiring an authoritarian management style that has close control, is punitive and removes control from workers. Alternatively, Theory Y assumes engaged workers capable of self-motivation and control, with a more consensual management approach where initiative is supported and rewarded. While it could be argued some human-computation platforms reflect Theory X, when there is a particular concern for, for example, cheating (Schmidt, 2010), Theory Y would be more appropriate for citizen science, as volunteers would not take part but for their self-motivation to participate.

This turning point in industrial management thinking saw the development of Job Characteristic Theory (Hackman & Oldham, 1975; Hackman, 1980). It identified factors such as task variety, skill variety, autonomy and feedback as key to ensuring work is desirable, engaging and offers intrinsic motivation. While in the years that followed the JCT has been extensively tested, and received criticism, this can be addressed by extensions and variants of the model rather than undermining its core tenets (Oldham & Hackman, 2010). It has also been suggested that the JCT can be specifically applied to crowdwork (Kittur et al., 2013), and due to the overlap described in section 3.1 can be extended to citizen science and VCS platforms.

3.6.1 Feedback

Based on Hackman & Oldham's work, there have been several studies looking at the effect of feedback on job satisfaction and productivity. Kim & Hamner (1976) studied 113 blue-collar workers at a telephone exchange to gauge the influence of both feedback and goal setting on performance. Experimenting with both intrinsic (self-reported), extrinsic (foreman reports) feedback and setting daily targets with the participants, it was found that goal setting alone had a marked positive effect on performance, and combining it with feedback only improved this result. Comparing how the influence of feedback differs depending on task type, Dodd & Ganster (1996) sampled 197 undergraduate students in a laboratory setting. Participants completed a number of differing word-processing tasks, and it was found that the effect of feedback correlated to the autonomy of the task. While feedback enhanced performance significantly for high

autonomy tasks, it was found to have little bearing when concerned with low autonomy counterparts. Studying a more quantitative form of feedback associated with computing tasks, electronic performance monitoring, Davidson & Henderson (2000) looked at the effect of displaying this information to 15 participants as they attempted to solve a number of purpose-built anagrams. The anagrams varied in difficulty, and it was found that while this form of feedback had a positive effect on both task proficiency and participant mood for easy tasks, the opposite was true for more difficult ones, suggesting participants get frustrated by realising their own poor performance.

Perhaps of more direct interest to the VCS community, Wilk & Redmon (1997) looked at the effects of feedback and goal setting on productivity and satisfaction specifically through using visual displays. Measuring the productivity and satisfaction of university administration staff though self-reporting and questionnaire responses respectively, results suggest that performance is improved when feedback and goal setting is initiated, and peaked when task completion metrics were displayed graphically to the user. Forming a different view to a large percentage of feedback and work research, Kluger & DeNisi (1996) concluded that feedback does not always have a positive effect on performance and motivation. Carrying out a historical review of the research, evidence suggests that in certain situations the positive effect of feedback only continues as long as the feedback continues, and if feedback is not continuous a negative effect can result. Alternatives were suggested that might not have this issue, including designing work that encourages trial and error, and therefore maximises the learning from feedback systems. Conducting their own review into the research, Ilgen, Fisher, & Taylor (1979) concisely summed up some of the issues regarding feedback, and the aspects that are important. Considerations such as its acceptance, the way it is perceived and the willingness of the recipient to respond all play a part in determining how effective feedback will be in improving performance and behaviour. Feedback has also been directly implicated specifically within VCS platforms. Mao et al. (2013), when studying paid and volunteer participants, found that the most commonly stated reason for giving up on a task was a concern that they were not performing well at it.

3.6.2 Complexity, Variety & Autonomy

Other factors that influence job satisfaction concern the makeup of tasks that are part of the work. Task variety, autonomy and complexity not only affect feedback as mentioned in the previous section, but can also, alone or in combination, directly effect job satisfaction and performance. Gerhart (1987) investigated the relationship between a range of factors and job satisfaction, through studying the survey results of over 10,000 participants taken in both 1979 and 1981. Participants answered a range of questions including rating their job satisfaction on a 4-point scale, from '*dislike it very much'* to '*like it very much'*, and one of the strongest positive correlations discovered was between this and job complexity, based on the Dictionary of Occupational Titles (DOT) rating system. It was also found that there was a temporal effect, with participants whose job complexity had increased most steeply over time also displaying a larger increase in job satisfaction.

Investigating task variety, autonomy and identity along with feedback effects, Dubinsky & Skinner (1984) surveyed 116 salespeople about their employment experience. Defining satisfaction using Hackman & Oldham's 5-item scale, and comparing the results with scaled responses to questions regarding variety, autonomy and identity, it was found that workers have higher motivation when their perceived job variety and identity (the bigger picture, how it contributes) is higher, and their job ambiguity is lower. Furthermore, it was found that workers job performance increases with perceived higher autonomy and variety. It is worth pointing out however that this survey was carried out using salespeople from one company, which could add skew to the results. Expanding on the theory, Chung-Yan (2010) not only looked at the effect of task complexity and autonomy on job satisfaction, but also considered the interactivity between the two factors. Through the use of an online survey, 259 participants responded to questions on autonomy, complexity and satisfaction through 5 and 7-point scales. Unlike the studies mentioned before, it was found that job complexity was not uniformly a motivator or stressor in relation to work, but depends on task autonomy. When autonomy is at a low level, complexity can be both a cause of motivation and stress, but at high autonomy levels complexity becomes beneficial in terms of job satisfaction.

Taking a different approach, but none the less still of interest to the citizen science and VCS community, Ghani & Deshpande (1994) considered the notion of 'optimal flow' during human-computer interaction. Characterised by intense enjoyment and concentration, the flow of 149 professionals was derived through the completion of 5-point scale questionnaires, and it was found that the users 'sense of control' determined flow. In addition, this sense of control became more important as the scope of the job increased, i.e. its autonomy, variety, complexity and identity.

3.6.3 Human Processing Units: A New Form of Taylorism

Despite the explained similarities between crowdwork and the early industrial models of the 20th Century, it could be argued that not all the lessons learnt have been heeded. There has been a tendency to 'dehumanise' the participant, treating them as human processing units or the Human API (Reeves & Sherwood, 2010) and designing workflows that optimise output rather than consider worker satisfaction (Little et al., 2010; Dai, Mausam, & Weld, 2010). Such approaches are driven by considering what machines cannot do rather than what humans can, referred to as left-over automation (Bainbridge, 1983) – resulting in a step back from the advantages of designing harmonious and productive sociotechnical work (Sharples & Houghton, 2016). It could be argued that this represents a form of *neo-Taylorism*, involving work designs averse to producing the satisfying and meaningful work proposed through Hackman and Oldham's (1975) Job Characterisation Theory.

If citizen science follows such an approach synonymous with similar crowdwork processes, there is a danger that the issues regarding participant job satisfaction, motivation and performance originally associated with early industrial work could again become a factor (Kittur et al., 2013). This danger is perhaps even more prevalent due to the lack of financial motivation required by the original Taylor model, and citizen sciences' reliance on attracting and retaining volunteers to perform the work as effectively as possible. The work presented in this thesis is a first attempt to better understand and mitigate these issues, by investigating the workflows of citizen science platforms, their influence on JCT factors and

ultimately the implications regarding the outcomes of a citizen science project.

3.7 SUMMARY

This chapter has defined key mechanisms that are utilised by citizen science and VCS platforms, which are also present in other similar disciplines. Expanding on this, related work has been reviewed regarding these processes, its applicability to the citizen science case, and when the outcomes learnt can be considered.

The effect of visualisation factors (see 3.3) has been shown to have an impact on both user motivation and performance, but may be of little use when informing VCS design. While there may be scope to optimise display 'window size', variables regarding maximum image resolution, content and colourisation are often fixed by the science case and the original instrumentation collecting the data. Likewise, the importance of training and experience in reducing the speed of vigilance degradation (see 3.4) cannot easily be adapted by citizen science, due to its reliance on workplace practises not suitable in an open, crowd sourced setting. While knowledge of operator variability and the demographic make-up of citizen science volunteers might help VCS developers target citizen science tasks more efficiently (see 3.5), this can realistically only be taken into consideration once the community has been formed (and only then after extra analysis by the developers and/or science team) rather than at the design stage.

The classification of psychophysical tasks and associated user judgements that are often required of VCS platform volunteers (see 3.2), and their effects on motivation and performance, along with the factors identified in Hackman & Oldham's Job Characteristic Theory (see 3.6) can however be used to influence VCS at the design stage. Not only this, but such factors are unlikely to be restricted by any science case or development limitations, and so can be freely adapted for the best outcomes. This classification of factors that are important to user performance and motivation prepares for the discussion of the kind of adaptations that could improve citizen science results, and provides the focus of the empirical work of this thesis.

This concludes Part I of the thesis. While chapter 2 framed the thesis' broader underpinning regarding citizen science and specifically VCS, this chapter has considered the 'nuts and bolts' that make up a VCS platform, and the previous research knowledge that could be applicable. This in turn informs the 'mechanics' of the empirical work, and feeds into the approach detailed in the next chapter.

PART II
APPROACH & METHODOLOGY

4

APPROACH TO STUDYING TASK WORKFLOW DESIGN

While the previous chapters chart the growing influence of virtual citizen science platforms, and outline the various disciplines and mechanisms they involve, this chapter marks the beginning of this thesis' main contribution. First, Task Workflow Design (TWD) is introduced and conceptualised as a construct whose exploration is at the core of the empirical studies in chapters 6-9. This conceptualisation is then grounded in a model of the different factors that determine a citizen science platform's task workflow design, which ultimately is used to achieve its scientific aims. From this conceptual model the methodology of this thesis' studies is then developed.

To inform the development of the model, this chapter considers both Hackman & Oldham's Job Characteristic Theory (1975) and the work of Pelli & Farell (2010) regarding psychophysical methods. The conceptual model of task workflow design is overviewed before its constituent parts – the design factors – are synthesised from the literature in detail in section 4.4. From an examination of the types of adaptations that VCS platforms can make at the design stage, *task type*, *judgement*, *autonomy* and *variety* are identified as the core TWD adaptations to be probed in this thesis. On that basis, the central research questions around task workflow design are then developed for the studies in chapters 6-9.

4.1 TASK WORKFLOW DESIGN

The concept of task workflow design is the core construct in this thesis. 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). While originally a concept associated with the manufacturing and business industries (Huang, 2002; M.-T. Schmidt, 1998), this thesis extends the notion and applies it to the virtual citizen science case, due to the analogy that can be made between them (see 3.6). In this example, the tasks of the workflow are the constituent tasks to be completed by the volunteer in order to achieve the overall work, a scientific analysis result or classification.

By considering task workflow design, more is understood about the specific types of work currently performed by VCS volunteers, beyond the broader scientific hypothesis and discipline involved. Furthermore, the effect of TWD on volunteer behaviour and ultimately the scientific outcomes of VCS platforms are assessed. One of the main objectives of this thesis is to understand what constitutes appropriate task workflow design, gauged by the usefulness of the scientific results collated to the scientific team involved.

Additionally, this thesis' contribution is not only achieved from an observational stance; strategies for task workflow design are developed based on previous literature, and implemented and tested using a real-world VCS platform incorporating an existing citizen scientist community. These task workflow designs are assessed by measuring volunteers' opinions and behaviour (quantitatively and qualitatively) and also by measuring scientific output.

In summary, task workflow design in this thesis is understood and utilised to:

- Capture and study the types of work carried out by volunteers of current virtual citizen science platforms,
- Understand the impact of task workflow design factors on volunteer behaviour and scientific output, and
- Evaluate different TWD strategies by their impact on volunteer behaviour and scientific output.

In studying task workflow design, it raises the question: What are the constituent factors of task workflow design? This includes the differences that can exist regarding these factors, and the effect they might have. In order to develop a model of TWD, and define the factors it includes, this

thesis draws on Hackman & Oldham's JCT and subsequent work, and Pelli & Farell's work regarding psychophysical methods.

4.2 TASK WORKFLOW DESIGN FACTORS AND THEIR EFFECT

As mentioned in 4.1, workflow and its design incorporate the tasks performed to complete the work required, along with the interactions that exist between them and the sequence in which they are performed. Previous literature has attempted to identify the key factors regarding such forms of work, and their effect on user behaviour and output.

4.2.1 Hackman & Oldham's Job Characteristic Theory

Although originally developed for industrial production work processes, JCT has been applied to the design of crowdwork (see 3.6), and as such has been extended to the VCS case. Originally Hackman & Oldham suggested that desirable, engaging work has the following characteristics: *skill variety*, *task identity, task significance*, *autonomy* and *feedback*. Table 4.1 explains these categories, and subsequently applies their meaning to the virtual citizen science example.

It is clear that these five factors are restricted in varying degrees by the scientific case and discipline involved, and therefore so is how much they can be influenced at the design stage of the platform. The adaptation shown in table 4.1 provides the basis for a model of constituent factors regarding task workflow design, which is further compared with Pelli & Farell's work on psychophysical methods.

Table 4.1: Hackman & Oldham's JCT Characteristics

Job Characteristic and Definition	Citizen Science Implication
Skill Variety: *The degree to which a job requires various activities, requiring the worker to develop a variety of skills and talents. Jobholders can experience more meaningfulness in jobs that require several different skills and abilities than when the jobs are elementary or routine*	Offer a range of different tasks for completion at different levels of complexity, in order to complete the overall scientific classification.
Task Identity: *The degree to which the job requires the jobholders to identify and complete a workpiece with a visible outcome. Workers experience more meaningfulness in a job when they are involved in the entire process rather than just being responsible for part of the work*	Define a citizen science activity that is simple enough to be quickly performed, but not so abstract or random as to seem meaningless. Allow volunteers to follow a single piece of work through several stages of analysis, or present them with data regarding a certain area or theme.
Task Significance: *The degree to which the job impacts other people's life. The influence can be either in the immediate organisation or in the external environment. Employees feel more meaningfulness in a job that substantially improves either psychological or physical well-being other than a job that has limited impact on anyone else*	In the context of citizen science, this relates to the primary motivation of volunteers being towards the actual scientific questions being addressed; i.e. it is being undertaken for a reason. This reason should be clearly communicated to the user, along with any resulting scientific findings from the project.
Autonomy: *The degree to which the job provides the employee with significant freedom, independence, and discretion to plan out the work and determine the procedures in the job. For jobs with a high level of autonomy, the outcomes of the work depend on the workers' own efforts, initiatives, and decisions; rather than on the instructions from a manager or a manual of job procedures. In such cases, the jobholders experience greater personal responsibility for their own successes and failures at work*	Regarding citizen science, this could be related to idea of allowing users different ways of interacting with the platform. While some autonomy might result in the benefits stated, too much freedom could be seen as antithetical to both quality control and the science case itself.
Feedback: *The degree to which the worker is provided with clear, specific, detailed, actionable information about the effectiveness of his or her job performance. When workers receive clear, actionable information about their performance, they have a better overall knowledge of the impact of their work activities, and what specific actions they need to take to improve their productivity*	Feedback can take on two forms when considering citizen science. Firstly, as is Hackman & Oldham's intent, feedback can be intrinsic to the work itself, i.e. a bricklayer can see his progress by the height and quality of a wall. Secondly, it can be feedback from a supervisor (i.e. a member of the science team) or some form of calculated scoring.

4.2.2 Pelli & Farell's Psychophysical Methods Approach

Psychophysical methods can be described as tools for measuring perception and performance (see 3.2). The approach has been common place for several decades (Boring, 1942) and during this time has been subject to considerable theoretical discussion regarding its use (Gescheider, 1985; Macmillan & Creelman, 1991). Pelli & Farell's (2010) work presents a more formal treatment, highlighting the importance of firstly defining the task the observer has to perform in order to define a psychophysically answerable question. In doing this various types of decision tasks and associated observer judgements have been identified, and reviewed in terms of their strengths and weaknesses regarding perception and performance.

Although originally developed to understand and specify the required characteristics of a display, tested using abstract visual stimuli such as dots or gratings, there are examples of the method being integrated with more practical applications. Disciplines incorporating medical (Beutel, Kundel, & Van Metter, 2000) and satellite (Hay, Niemann, & McLean, 1996) imaging have adapted the approach to their case, and due to VCS platforms' heavy use of similar remotely sensed imagery it has been extended to the citizen science example for this thesis.

4.2.2.1 Observer Tasks

The different tasks identified by Pelli & Farell are:

DETECTION: The goal of a detection task is to determine the amount (energy, contrast etc.) of a stimulus that makes it just detectable for some observer.

DISCRIMINATION: In a discrimination task, the focus is the smallest difference along some physical dimension that permits an observer to reliably distinguish between two stimuli.

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, even though one or more other attributes may remain different.

IDENTIFICATION: The observer's task is to identify the stimulus on each trial by responding with a pre-assigned value or categorisation. The degree to which the observer is familiar with the stimulus and their corresponding labels can vary.

DIRECT SCALING: The goal of a direct scaling task is to provide a scale of some perceptual magnitude (e.g. loudness, brightness, or size) as a function of the corresponding physical dimension(s).

4.2.2.2 Observer Judgements

In order to achieve the defined tasks, Pelli & Farell further identified the following different judgements required of the observer:

THRESHOLD: The observer has to adjust the stimulus contrast up and down until the point where it is "just barely detectable".

MATCHING: Two stimuli are presented, and the observer has to adjust and judge when one exactly matches the other. The level to which this can be a "perfect match" can vary.

YES NO: Usually used for detection tasks, the observer is either asked to classify the stimulus "was a nonblank stimulus present?" or classify the percept, "did you see it?" The observer is only allowed two response alternatives, yes or no.

RATING SCALE: The observer is asked to either rate the likelihood that a nonblank stimulus was presented, or the magnitude of a physical dimension (brightness, size etc.). This can be via a number of fixed alternative ratings or continuous scale.

TWO-ALTERNATIVE FORCED CHOICE: Similar to the Yes/No judgement, this time the observer is shown two separate stimulus presentations, successively or side by side, and asked if the nonblank stimulus was first or second (or on the left or right).

Perhaps due to the reliance on the display of remotely sensed visual stimulus, and the volunteers' perception and analysis of it, the tasks and judgements required of citizen scientists using VCS platforms bear a remarkable resemblance to Pelli & Farell's taxonomy. As such, this strengthens the motivation to develop a conceptual model of the factors

that make up task workflow design whilst adding two factors (*user task* and *judgement*) that are not represented in a model solely based on Hackman & Oldham. These additional factors are arguably more intrinsic to the constituent tasks that make up the work required, compared to the factors of the JCT regarding their presentation, interaction and contribution to the overall aims of the science case.

4.3 A CONCEPTUAL MODEL OF TASK WORKFLOW DESIGN

This section introduces the conceptual model of this thesis, comprising of the factors underlying the task workflow design of a VCS platform. The model consists of both the intrinsic factors associated with the constituent tasks required of the volunteer, and the broader factors regarding their presentation to the user, their interaction and their significance, both in terms of the scientific outcome and as interpreted by the volunteer. To begin, the model is introduced to give an overview of its components. The model's grounding in empirical work as explained by previous literature is then detailed (4.4). For clarity, the model is introduced before the factors that comprise it are presented, although in reality the model was developed from a synthesis of findings from crowdsourcing and human factors studies in the literature.

The taxonomy of the conceptual factors that influence the task workflow design of VCS platforms presents an analytical tool that could inform related research, by providing an overview of the considerations involved in the design process of such systems. It could also allow researchers to position their own contribution to the field in relation to a broader model.

Depicted in figure 4.1, the model will be extended at various points in this thesis to illustrate its relation to the empirical work carried out. It consists of 7 TWD factors, 5 adapted from Hackman & Oldham's JCT (see table 4.1), and 2 from Pelli and Farrell's psychophysical methods work (see section 4.2.2 – and the red box in figure 4.1). The model broadly distinguishes two types of TWD factors, *science case influenced* and *design influenced*. Science case influenced factors are those whose design is dependent on the science being addressed, and the development team of the platform have little influence. In contrast, design influenced factors are mostly unaffected by the science case, and the development team have the freedom to consider

them as they see fit. The further away a factor is from the centre of the model correlates to an increase in the influence, one way or the other.

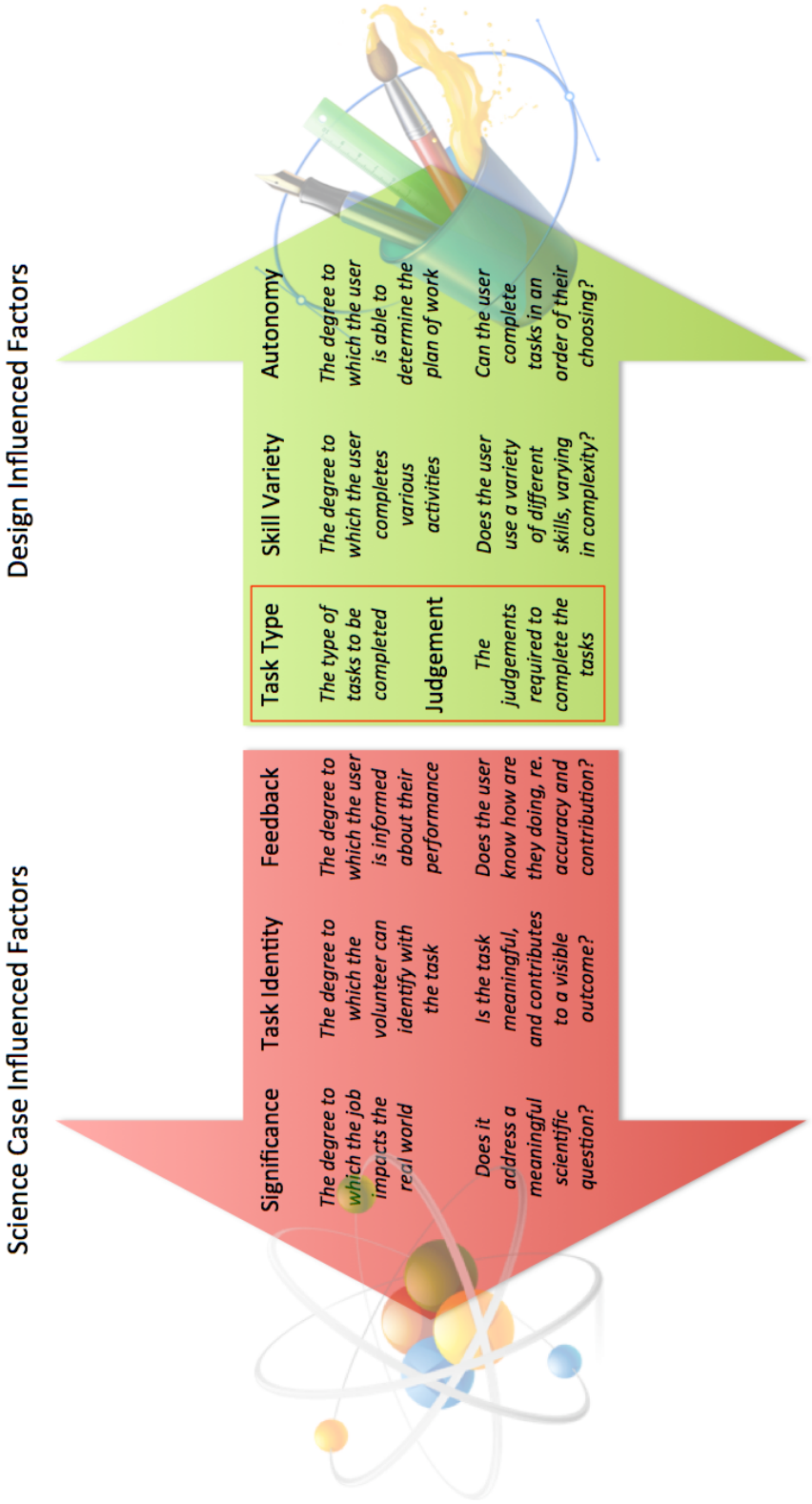


Figure 4.1: Model of the factors of task workflow design. Factors within the red box are adapted from psychophysical methods work, whilst the remaining come from Hackman & Oldham's JCT

For example, the significance of the work completed by the volunteer is highly influenced by the science case addressed, whereas the autonomy of the work is mostly influenced by the design of the platform.

The following example illustrates how the model might be applied in practice:

> When developing a VCS platform, the science case has to be evaluated regarding its a) significance, it addresses a meaningful question that impacts on the real world, b) task identity, the work and tasks carried out by the volunteer are meaningful, and contribute to a visible outcome, and c) feedback, information can be relayed back to the volunteer regarding their contribution and performance.
>
> The design of the interface has to consider a) task type and user judgement, the tasks the user must do, and the decisions they have to make to achieve them, b) the variety of different tasks each volunteer completes, and their complexity, and c) the autonomy of the work, and the independence of the volunteer do take their own path, rather then follow set rules.

It could be argued that science case factors should be considered when deciding if the virtual citizen science mechanism is the right approach, while design factors should be considered to ensure the platform is satisfying to use and results in robust scientific output.

The following section justifies the model by synthesising the contributions of related TWD studies, and applying their contributions to the model of VCS factors.

4.4 TASK WORKFLOW DESIGN FACTORS OF VIRTUAL CITIZEN SCIENCE

The human factors literature outlined in chapter 3 demonstrates the effect that the type, flow and presentation of work can have on user performance and satisfaction. By applying this knowledge to the field of virtual citizen science, the task workflow design of such platforms is broken down into its constituent factors in order to increase our understanding and to inform the experimental design in chapters 6, 7, and 8, and the live study in chapter 9.

4.4.1 Science Case Influenced Factors

Science case influenced factors are those influenced by the overriding scientific goal of the project, and as such the degree to which they can be influenced by design is restricted. The overall scientific goal is in turn influenced by the importance of the research to the discipline, and the lack of an alternative approach to carry out the analysis (see 2.3).

SIGNIFICANCE:

The significance of the work, as stated by Hackman & Oldham, is the degree to which it impacts the real world and effects people's lives. In terms of virtual citizen science, it can be seen as the importance of the scientific question being addressed (and its impact on the scientific community). Whilst this is not directly comparable, the science case and its impact has been shown to be a key motivator for volunteers to visit VCS platforms in the first place (Raddick et al., 2010, 2013; see 2.6.3), and so the reasons for doing the science, and how the analysis will be used needs to be clearly communicated.

TASK IDENTITY:

Identity relates to the degree to which the user can identify with the task, interpret it as meaningful, and visibly see its contribution to the overall outcome. While it could be argued that this can be somewhat influenced by design, i.e. ensuring volunteers can follow a piece of work through several stages of analysis, it is still heavily constricted by the science case. For instance, while the overall goal of the project might be to estimate the degree of geological activity on a planet – the task to achieve this presented to the volunteer is merely to count the number of craters on the screen. The task is somewhat abstract to the aim, and therefore could be hard to identify with and understand its contribution. Identity, and in turn lower task ambiguity, has been shown to increase the motivation of users (Dubinsky & Skinner, 1984; see 3.6.2).

FEEDBACK:

Feedback can take two forms; intrinsic where the user can see their progress - for instance a bricklayer building a wall, or external feedback – i.e. from a supervisor or a score based on performance. While the timing and type of feedback presented to the volunteer can be a design

consideration (interrupt or requested, positive or negative etc.) there are also significant dependencies on the science case. In order to provide intrinsic feedback, the volunteers' contribution has to be given some context that can be easily communicated. While this may be possible where the area of analysis has a physical dimension (geographical region etc.) it could be more difficult where the task result is merely contributing to a 'pool of numbers'. In the case of external feedback (i.e. a score), this requires some sort of ground truth to be known regarding the data analysed – which again is dependent on the science case and/or the scientists involved. Other forms of external feedback can occur outside of the main task workflow, through discussion with peers on project forums. Evidence suggests that over time this can result in a learning effect that could improve performance (Luczak-Roesch et al., 2014) – however this will not provide a benefit to the large proportion of volunteers that do not interact with discussion within the community. Feedback can be key to improving performance (Kim & Hamner, 1976), an effect that increases with the autonomy of the tasks involved (Dodd & Ganster, 1996), and as such is an important factor to be considered regarding task workflow design.

4.4.2 Design Influenced Factors

Design influenced factors are those that can be, for the most part, directly manipulated by the developers and designers of the platform, with little restriction due to the science case being investigated. Although at first glance this might suggest that an 'absolute' design can be achieved, one that ensures the most efficient scientific output and volunteer engagement, the reality could be more complicated. As seen in the previous literature discussed in chapter 3, the impact of certain TWD factors and their interactions can be complex and not always complementary.

AUTONOMY:

The autonomy afforded to the volunteer, i.e. the degree of freedom present to work independently to complete the tasks, can be considerably influenced by the design of the platform. For example, while the science case might require a set number of different tasks to be completed, the design can dictate what order they are completed in, when they are

completed and how many times. Alternatively the volunteer can be allowed free choice over one or a number of these considerations. Greater autonomy has been shown to increase performance (Dubinsky & Skinner, 1984) and also motivation, especially when combined with more complex tasks (Chung-Yan, 2010; see 3.6.2).

SKILL VARIETY:

As the name suggests, skill variety relates to the range of different task types and complexity the user has to undertake in order to complete the work. While to a degree this can be influenced by the science case, i.e. the type of analysis required regarding a particular stimulus of interest, the design of the interface plays its part. For instance, a science case might require 4 differing types of task to be completed, with varying complexity. The TWD of the platform can vary in terms of how these tasks are presented, i.e. does one group of volunteers do task 'A', another group task 'B' etc. and then the results are aggregated? Or do all volunteers carry out all of the tasks, in a directed or random order therefore providing greater variety? While greater variety has been shown to increase motivation (Dubinsky & Skinner, 1984), the timing of the change of task can be important – as changes can both slow down task completion (Lasecki et al., 2014) and also increase engagement rates over longer periods (Rzeszotarski et al., 2013; see 3.1).

TASK TYPE & JUDGEMENT:

The different task types and associated user judgements required to be made in order to complete a classification relates to the categorisations derived by Pelli & Farell (2010) in section 4.2.2. Again, although these can be influenced by the science case, i.e. a crater counting science case requires the completion of an identification task (the user has to, at some point, know what a crater is and identify it on an image), this does not tell the whole story. The user could be asked to complete the identification task through a yes/no judgement (i.e. does the image contain any craters, yes or no?), or a discrimination task through a forced choice judgement (i.e. which image has more craters, left or right?), or a matching task through a threshold judgement (i.e. draw a circle around the visible edge of the crater). All of these methods would result in differing measures regarding

crater numbers in the images, and can be influenced by the design of the platform.

Previous literature suggests that the effect of task type and judgement is complex and dependent on their combination along with the experience of the user involved. While investing detection and identification tasks, Thomas (1985) found no significant difference in accuracy between utilising a rating or forced choice judgement. However, it has been found that when carrying out a detection task, a 2-alternative forced choice (2AFC) judgement is less accurate than a 4-alternative forced judgement (4AFC) when the users are inexperienced, with the opposite being true for a discrimination task (Jäkel & Wichmann, 2006; see 3.2.1).

The Task Workflow Design model outlined in this chapter, along with its constituent factors, is intended to be applicable to the majority of VCS platforms that conform to a general pattern of derivation and development. It is accepted that the model might not be suitable for certain 'special case' platforms, especially in terms of the separation of *science case influenced* and *design influenced* factors. A platform that addresses a very specific science case for instance, might have little leeway in terms of variety and autonomy due to the highly concentrated nature of the task. Alternatively, a platform addressing a very broad science case (i.e. Studying features on the Martian surface) might be able to have more design input regarding the significance factor, as there is a choice of different features and image types to focus on. What is clear is the influence that TWD can have on user satisfaction and performance, and the following postulates were distilled in order to show the veracity of TWD factors.

4.4.3 Postulates from Related Work

The following postulates were derived from the human factors influenced literature review conducted in chapter 3. They provide empirical evidence of the importance of *science case influenced* (P1-P4) and *design influenced* (P5-P11) factors on the effects of task workflow design regarding performance and motivation.

P1: The significance of the work (the science case) is a key motivation for volunteers to visit a VCS platform (Raddick et al., 2010; 2013)

P2: The ability of the volunteer to identify the contribution of the task/s towards achieving the scientific aim increases motivation (Dubinsky & Skinner, 1984)

P3: Feedback improves performance, and its influence increases with volunteer autonomy (Kim & Hamner, 1976; Dodd & Ganster, 1996)

P4: The effect of feedback in influenced by its timing and content (Kluger & DeNisi, 1996; Mao et al., 2013)

P5: Greater autonomy increases volunteer performance (Dubinsky & Skinner, 1984)

P6: The effect of autonomy on motivation is influenced by the complexity of the task (Chung-Yan, 2010)

P7: Greater skill variety increases volunteer motivation (Dubinsky & Skinner, 1984)

P8: The effect of skill variety is influenced by the timing of the change in task (Lasecki et al., 2014; Rzeszotarski et al., 2013)

P9: The effect of user judgement on performance is influenced by the type of task it addresses (Jäkel & Wichmann, 2006)

P10: User judgement influences the efficiency of task completion (Jäkel & Wichmann, 2006)

P11: User judgement and task type influences inter-participant agreement (Hutt et al., 2013)

Whilst there is an abundance of previous empirical research regarding the effect of the task workflow design factors described in the model (chapter 3), it is mostly related to mechanisms that employ a similar workflow paradigm to those of VCS platforms, rather than directly involving citizen science itself. Past studies that do involve citizen science platforms tend to concentrate on the broader concepts of volunteer motivation (related to P1), demographics (Raddick et al., 2009), or general platform functionality (Prestopnik & Crowston, 2012). This thesis takes on the challenge of investigating task workflow design in direct relation to VCS platforms. The effects of adapting TWD factors both in a laboratory setting (chapter 8) and in the wild with a pre-existing volunteer community (chapter 9) are studied using an established citizen science platform.

4.5 FOCUS OF THIS THESIS

With the span of the conceptual model thus developed, it is now time to narrow down the scope to consider the aspects of the model that provide the focus of this thesis. Whereas task workflow design is mediated by the interaction of *all* the factors outlined previously, a detailed experimental study of the impact of each one is considered outside the scope of this thesis.

In order to achieve the required focus for the thesis, the factors that have the most practical applicability to a VCS platform developer (i.e. those that can be influenced by them) have been targeted. In real-world terms, this breaks down to the factors that are not restricted by the science case of the platform, or the data associated with it. The science case of a VCS platform is usually 'set in stone' with little flexibility to adapt or change. This is both because it is normally the key driver for a citizen science platform's creation (due to the lack of alternative methods for the analysis, see 2.3), and also as it has to result in findings that are academically robust (i.e. contribute to the advancement of the discipline, see 2.5). Therefore, while it is important to understand the effect of science case influenced factors, perhaps in order to decide if the VCS approach is suitable in the first place, they have little practical use at the design stage of the platform.

4.5.1 Focus on Design Influenced TWD Factors

This thesis will focus on the design-influenced factors defined in the TWD model, both due to their applicability at the design stage and the lack of previous research into their effect regarding virtual citizen science. From this point forward, task workflow design (TWD) will refer to a simplified version of the model, as depicted in figure 4.2, consisting of only the design influenced factors previously discussed. In relation to the full model depicted in figure 4.1, it is derived from the right side of the figure, the half shown by the green arrow representing design-influenced TWD factors.

These factors consist of (as shown) *Task Type* and *Judgement*; as defined by Pelli & Farell's work regarding psychophysical methods, and *Skill Variety* and *Freedom*; their importance identified by Hackman & Oldham's job characteristic theory work. They are at the forefront of the empirical

section of this thesis; the effects of their manipulation and experimentation are studied, especially in terms of the postulates P5-P11 (see 4.4.3), within a virtual citizen science setting in order to better understand their importance to the scientific outcomes of such platforms and to inform future design.

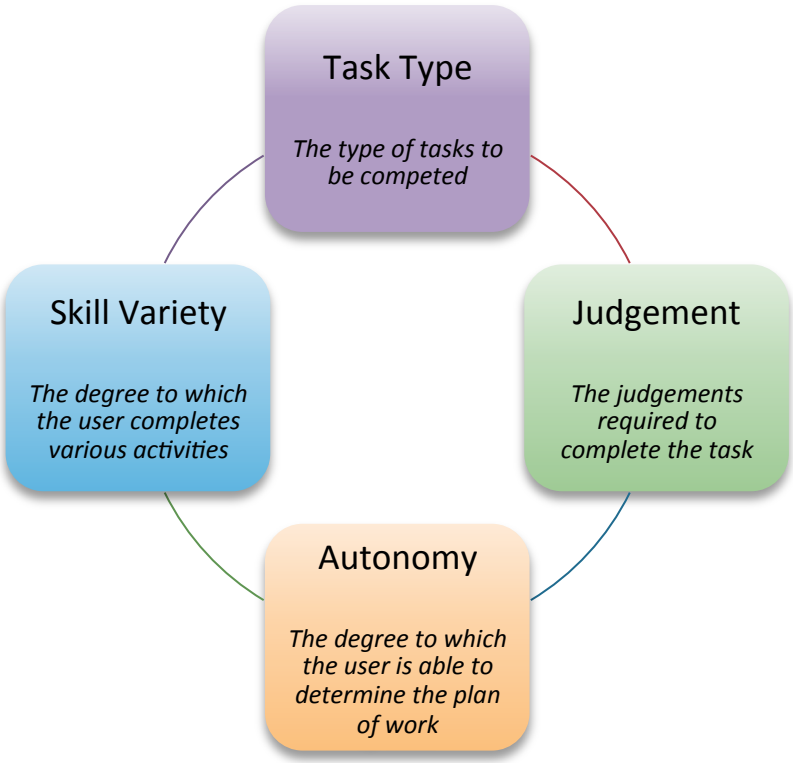


Figure 4.2: Model of Design Influenced TWD Factors

4.5.2 Relation of the Model of Design Influenced Factors to the Empirical Work

While the model of task workflow design factors developed in 4.3 of this chapter seeks to comprehensively frame their importance in the virtual citizen science design process, the empirical work described in part III of this thesis explores the relation of some of these factors through a more systematic approach. The study in chapter 6 firstly reviews existing VCS platforms in terms of their task workflow design through the completion of a hierarchical task analysis (explained in more detail in chapter 5), and investigates its influence on volunteer behaviour. Chapter 7 collects feedback from an existing VCS community regarding their opinions on TWD and other more general interface design factors.

While chapters 6 and 7 concern the current state of play regarding virtual citizen science, chapters 8 and 9 are more experimental in nature. The design-influenced factors (see figure 4.2) provide the independent variables (IVs) that have been manipulated or controlled by experimental design. The experiments test the effects of different manipulations of *task type*, *judgement*, *skill variety* and *autonomy* on the dependent variables (DVs) of scientific outcome. Observations of scientific outcome in the experiments are collected both through participants' feedback and in terms of the resulting scientific data.

4.5.3 Measuring the Scientific Outcomes of a VCS Platform

While the postulates described in 4.4.3 concern the effects of task workflow design factors on measures of motivation and performance, this thesis is interested in their effect on scientific outcome, in order to better understand how to design workflows that ensure a platforms' success. This raises the question: *what is a successful scientific outcome, and how do you measure it?* This section will describe the factors used by this thesis that contribute to a successful scientific outcome for a VCS platform, based on the findings within the literature of chapter 2.

4.5.3.1 Volunteer Motivation & Engagement

One of the key measures of success regarding a VCS platform is its ability to motivate and engage volunteers, i.e. the number of visits and visitors it receives. This impact also relates to the scientific outcomes of the platform, both in order to get through the large of amount of data it was designed to address (Lahav et al., 1995; see 2.3.1) and its ability to retain volunteers, as return users usually complete most of the work (Ponciano et al., 2014; see 2.6.4). Although previous research has considered task workflow design for the casual volunteer (Eveleigh et al., 2014; see 2.6.4), as of yet little has been done regarding the use of TWD to motivate volunteers to return, and perhaps alter the 20-80% split (committed vs. casual).

4.5.3.2 Data Coverage

Some of the most influential VCS platforms are deemed a success due to their creation of large, detailed data catalogues that are much greater in scope than can be achieved by alternative methods (Lintott et al., 2011, see 2.5.1) thus, data coverage is seen as an important scientific outcome. Coverage is also an important factor in reducing the number of 'null' results (classifications with no measure), an issue for scientific teams when reducing and interpreting the platform data.

4.5.3.3 Data Accuracy

In order to provide results and findings that help further the knowledge of science, the data provided by VCS platforms has to be accurate to a level accepted by the discipline involved. Examples of successful VCS platforms have been able to correct and improve existing datasets created by other methods due to their increased accuracy and detail (Hennon et al., 2014; see 2.5.1). Accuracy is also an important scientific outcome in reducing the load on the science team involved, by reducing errors made either by mistake or systematic bias that need consideration at the aggregation stage (Arcanjo et al., 2014; Chambert et al., 2014; see 2.5.2).

Although there are arguably many more facets of scientific outcome that can be used to review the success of a VCS project (outreach and knowledge transfer, see 2.3.3; serendipitous discovery, see 2.3.2), the 3 highlighted will be evaluated as part of this thesis. This is due to their measurability in an experimental context, their direct relation to the core scientific case, and their similarity to the dependant variables of the postulates derived from the existing literature.

4.5.4 Research Questions

The title of this thesis, '*Designing User Task Workflows to Ensure the Best Scientific Outcomes in Citizen Science*' gives rise to the principle research question:

$RQ_{prinicipal}$: What task workflow designs ensure the best scientific outcomes?

As a result of the focus on task workflow design, and specifically its design influenced factors, the effects of different strategies to adapt TWD is investigated in order to understand the ways in which it can be manipulated to effect the scientific outcomes of VCS platforms. The principle question is studied alternatively by addressing three sub-questions RQ_{1-3} in the two study chapters and two experiments (laboratory and field) that comprise the empirical part of this thesis.

In studies 1 and 2, the rationale was to review the current state of play regarding existing virtual citizen science platforms and their task workflow design, and to start to investigate TWDs broad effect on volunteer motivation and behaviour – answering the research question:

RQ_1: What task workflow designs are currently used by citizen science projects?

In the experimental chapters (8 & 9), the aim is to test if manipulating the derived design influenced task workflow factors would influence the scientific outcomes of the platform. The research questions were:

RQ_2: Does task workflow design influence volunteer behaviour?

And:

RQ_3: Does task workflow design influence volunteer results?

Both experiments explored these questions thorough the creation of VCS interfaces that presented different task workflows, and testing their effect. Experiment 1 did this in a laboratory setting using invited participants, while experiment 2 was conducted 'in the wild', promoted through an existing VCS platform and utilising an existing volunteer community.

4.6 SUMMARY

This chapter has developed the conceptual model that underpins the approach of this thesis. Task workflow design has been identified as a key construct that can influence the experience of the volunteer, and the results they produce (see 4.1). The model of task workflow design factors (see 4.3) has been influenced by the JCT work of Hackman & Oldham and the psychophysical methods work of Pelli & Farell (see 4.2), and synthesised from a review of the literature (see 4.4). This also revealed a lack of studies

applying TWD theory specifically to the VCS case, which further motivates the methodology of this thesis.

An analysis of how the factors are influenced (by science case or design) has added focus to the model (see 4.5), and clarified its relation to the design of the experiments in this thesis. The measures of scientific outcome (the dependent variables) used in this thesis have also been defined, as informed by the literature in chapter 2. The principal research question has been stated and subsequently broken down into three sub-questions that drive the studies in chapters 6, 7, 8 and 9, that centre on the manipulation of task workflow design factors to investigate appropriate designs that ensure successful scientific outcomes (see 4.5.4).

The next chapter presents the methodology used to investigate task workflow design specifically regarding the VCS case, both in terms of reviewing existing platforms and experimentally through laboratory and real-world studies.

5

METHODOLOGY OF STUDYING TASK WORKFLOW DESIGN

This chapter takes an in-depth look at the methodology that underlies the empirical approach taken in the presented studies of this thesis (chapters 6 to 9).

After the empirical studies carried out in this work are laid out, the motivation for the choice of Zooniverse VCS platforms as the focus of investigation is discussed. A mixed methods approach is introduced, summarised and justified, including the adaptation of the hierarchical task analysis (HTA) method for the review-type study in chapter 6, and the VCS platform Planet Four for the studies presented in chapters 7, 8 and 9. Furthermore, questionnaires and expert comparison are introduced as supplementing the approach.

A new science case is introduced that has been developed in conjunction with the Zooniverse's Planet Four project specifically for this thesis. The choice to go through the entire process of formulating a new scientific aim, rather than adapt an existing project, is in line with the pragmatic attitude of this thesis – to both develop an understanding of task workflow design that is applicable to the established virtual citizen science community, and also produce a project that could be used beyond the duration of this thesis to address a genuine scientific goal in addition to the experiments' aims. After the merits, limitations and applications of the thesis' approach and science case are discussed, the use of these methods in the empirical work are briefly summarised and contrasted.

5.1 OVERVIEW OF STUDIES AND METHODS EMPLOYED

Table 5.1 describes the empirical studies carried out for this thesis, a description of their purpose, the research questions they address and the methods adopted in the process. A more in-depth description of the methods and procedures undertaken can be found within the chapter each study is described.

Table 5.1: Overview of thesis studies and methods

Study	Purpose	Research question(s)	Methods used	Chapter
Study 1	Review of current citizen science workflows and preliminary analysis of potential influence on volunteer behaviour	What Task Workflow Designs are currently used by Citizen Science Projects?	- Hierarchical Task Analysis (5.3.1)	6
Study 2	To canvas the opinions of an existing citizen scientist community	What are the volunteers' concerns regarding the design of the platform?	- Online questionnaire (5.3.2)	7
Experiment 1	Investigate the effect of Task Workflow Design manipulation in a controlled setting	Which Task Workflow Design configurations do volunteers prefer? – Does Task Workflow Design influence volunteer results?	- Development of science case and platform (5.4.1) - Questionnaire - Expert data comparison (5.4.2)	8
Experiment 2	Investigate the effect of Task Workflow Design manipulation in a real-world environment	Same as experiment 1, and additionally: Does Task Workflow Design effect volunteers' behaviour?	- Development of science case and platform (5.4.1) - Questionnaire - Expert data comparison (5.4.2)	9

5.2 THE ZOONIVERSE SETTING

As explained in chapter 2, over the past decade citizen science has increased in size as a discipline in terms of both its scientific breadth and participation. Virtual citizen science is a major contributor to this growth, and due to the Zooniverse's position at the forefront of this area (see 2.4) its projects have been chosen as the focus of this thesis.

The choice of the Zooniverse ensures that a wide range of scientific disciplines are included in the analysis, while also normalising some of the more broad design and appearance factors associated with website platforms. Additionally, due to the membership of the authors institute within the citizen science alliance, unprecedented access to each platform's statistics has been possible. This has allowed for an initial investigation into understanding the effect of TWD on volunteer behaviour. By studying an established VCS system, it has also been possible to collate the views and opinions of an existing citizen science community (as described in chapter 7), the results of which have been used as a base point for comparison with the findings of chapters 8 and 9.

The pragmatic approach to task workflow design adopted in this thesis makes it necessary to study its effect in a real-world setting. The reason for this is to ensure that the findings of this thesis regarding TWD can be applied and exploited by the VCS community, both in terms of their existing and future design. Also, real-world studies often result in greater ecological validity (Scollon, Prieto, & Diener, 2009), for the real-world context of the studied activities to be considered.

The experiments in this thesis consist of a laboratory and real-world study, the first being a 'trial-run' of the procedure in order to ensure the effectiveness of the latter in the wild. Both raise a set of challenges, which are imposed by the practical difficulties of combining a new scientific aim and data with the development of a new platform that in turn incorporates experimentation in task workflow design. By adapting the Zooniverse system layout and code base (Simpson et al., 2014)[11], it has been possible to create and modify different task workflow design interfaces in a much more timely fashion than would be possible if created from scratch[12]. Also, the standard Zooniverse appearance and layout has been used, negating the need to spend resources on their design. Finally, in terms of the real-

[11] Zooniverse API (Ouroboros) can be found at: https://github.com/zooniverse/Ouroboros. Original Planet Four code base can be found at: https://github.com/zooniverse/planet-four.

[12] Adapted code for Planet Four: Craters (project for this thesis) can be found at: https://github.com/zooniverse/NottMarsZoo.

world study, by launching the platform through the Zooniverse it can be advertised to a large existing community from day one, addressing the issue of identifying and recruiting a large representative audience to take part.

As laid out in chapter 4, the main aim of this thesis is to better understand the effect of task workflow design on the scientific outcomes of VCS platforms, with the focus being on the Zooniverse example. In order to satisfy this requirement, a mixed methods approach is adopted to support the studies in this thesis.

5.3 MIXED METHODS APPROACH

In keeping with the practical stance this thesis adopts, the empirical approach to studying task workflow design in regards to virtual citizen science platforms is characterised by using mixed methods. In particular, the studies' quantitative findings from website analytics, questionnaire agreement ratings and scientific output are contextualised with qualitative insights from 'open-ended' questionnaire responses and platform forum posts. Scientific observations made during the experimental studies (chapters 8 and 9) are accompanied by analysis of system logs that capture volunteer behaviour. The method of collating data from multiple sources to provide corroborating evidence is also known as data triangulation, and can help increase their validity while dealing with quality issues regarding data provided by a single source (Lazar, Feng, & Hochheiser, 2010).

Before moving on to describing the experimental focus of the thesis in terms of the Zooniverse platform Planet Four, the methods incorporated into the review type studies (chapters 6 and 7) are described.

5.3.1 Hierarchical Task Analysis (HTA)

In order to review the different task workflow designs employed by current VCS platforms in the study of chapter 6, a hierarchical task analysis was conducted for each of the Zooniverse projects involved.

The process of HTA is to decompose the main task of a certain procedure into a number of sub-tasks, to the level of detail required. Each sub-task can be defined by a goal, the input conditions that activate it, the actions

required to achieve the goal and the feedback indicating success (Hollnagel, 2003). While originally developed in the 1960's for training process control tasks in the steel and pharmaceutical industries (Annett & Duncan, 1967; Annett et al., 1971), in recent years the approach has been adapted for other disciplines including HCI (Diaper & Stanton, 2003; Stanton, 2006). There is a clear analogy between the decomposition of a procedure carried out within the HTA process and the concept of task workflow set out in section 4.1, and therefore is a suitable method to review the current state of play regarding VCS platforms.

5.3.1.1 Limitations of HTA

Beyond the general limitations of HTA regarding its time expense, rigidity regarding unforeseen events and lack of addressing decision rationale (Jamieson et al., 2007) there are also specific limitations of the method regarding its application to VCS. The advantages regarding identifying the number of tasks required of the volunteer and the autonomy of their completion order are clear, but there is a degree of subjectivity, especially regarding classifying the types of task and judgement required. To address this issue, the thesis has adapted the definitions of Pelli & Farell's psychophysics work (see 4.2.2) and for the majority of cases, this classification was straightforward (i.e. identifying if a crater is present in the image or not is clearly a 'yes/no' judgement), but there have been examples where the author has had to make a subjective decision (with guidance from the supervision team). For instance, is a transcribing task an identification type (identifying which letters/words are displayed based on previous knowledge), a detection task (determining if a word/letter is present in the first place), or both?

5.3.2 Questionnaires

The overview study in chapter 7 and the experiments in chapters 8 and 9 made use of questionnaires in order to review the opinions of VCS platform volunteers. The questionnaires served the important function of contextualising quantitative data gathered through ranked responses with qualitative accounts and descriptions fostered by supplying the opportunity to give 'free text' responses. Whereas some questions were

posed to specifically assess the experimental hypotheses, the qualitative questionnaire data was coded and then collated into themes according to the data-led process of grounded theory (Glaser & Strauss, 1967).

5.3.2.1 Likert Item Type Statements

In order to gather quantitative data regarding volunteer opinion, Likert item type statements have been included within the questionnaires. A Likert item is simply a statement that the respondent is asked to evaluate by assigning a quantitative value to it along a subjective or objective dimension (normally a level of agreement or disagreement). They have been utilised across multiple disciplines involving questionnaire research, including website design and HCI (Braddy et al., 2003; Brave, Nass, & Hutchinson, 2005). For the purpose of this thesis a nine-point scale has been used in order to ensure a more 'fine-grained' response that can detect more subtle differences, to counteract any positive bias especially concerning the responses of the existing volunteer community.

5.3.2.2 NASA Task Load Index

In addition to the more general Likert item type and open-ended 'free text' responses, the questionnaires also included the standardised NASA Task Load Index (TLX)[13] to help gauge the scientific task's overall perceived workload, and the effect of task workflow design on it. The NASA-TLX framework has been used in previous research to measure workload regarding a number of on-screen and HCI type tasks (Adamczyk & Bailey, 2004; Harrower & 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.* For the study in chapter 7, and the experiments in chapters 8 and 9, this served as a measure to check that the manipulation of IV task workflow design factors

[13] Available at http://humansystems.arc.nasa.gov/groups/TLX

were genuinely perceived as different as indicated by significant changes in workload.

5.3.2.3 Limitations of the Questionnaires

There has been a whole host of research exploring the limitations of the Likert format, focusing on the interval levels used and the way in which the resulting data is analysed (S. Jamieson, 2004; Knapp, 1990). For each finding presented however, there are a number of published counter-arguments (Carifio & Perla, 2007), and there seems little consensus regarding any form of 'best practice'. A common theme that does exist in the literature is the importance of the statements themselves, in terms of minimising any type of bias. It is known that statements containing negative words can result in responses with a negative bias and those containing positive words the opposite (Garg, 1996), and so to reduce this effect the questionnaires used in this thesis contain a similar number of both.

Other general issues regarding the use of questionnaires revolve around the timing of their completion. For instance, if the participant completes the questionnaire after the performing the task, they may have forgotten certain usability and workload aspects. To combat this for the experiments in chapters 8 and 9 participants completed the questionnaire immediately after using each interface variation. Regarding NASA-TLX, there has been previous evidence of responses being biased by the participants' perceived performance (Hart, 2006), and therefore the questionnaire results in this thesis have been compared with measures of actual and perceived performance in order to identify any correlation that might exist.

5.4 PLANET FOUR

To investigate the effect of manipulating task workflow design and also separate the effect of other external factors (scientific discipline, media exposure etc.), the study and experiments in chapters 7-9 focus on one particular Zooniverse VCS project, Planet Four.

Launched in January 2013, Planet Four[14] allows volunteers to help planetary scientists identify geological features on the surface of Mars by displaying high-resolution, remotely sensed imagery to analyse. It has been chosen as the focus of this thesis for several situational and logistical reasons. Firstly, due to the author's background regarding planetary science and GIS analysis, both the subject matter and the specific scientific measurements collated are of obvious interest and act as an important motivation. Secondly, the timing of launch at the beginning of the thesis, and its subsequent parallel 'live' period across several years of the work has its own benefits. It ensures that the thesis studies involve the full spectrum of volunteers, from the one-time visitors drawn by initial promotion to more dedicated volunteers who have taken part regularly over the projects' lifetime. A final practical reason is due to the platforms' appearance on the BBC's 'Stargazing Live' transmission. This unique opportunity for exposure has an obvious knock-on effect in terms of visitor numbers to the site (at least initially) and subsequently can only help the thesis' studies in terms of sample size.

5.4.1 The Crater Counting Science Case

Although Planet Four is the backdrop for a large part of the empirical process in this thesis, rather than using the original science case on seasonal 'fans'[15], a new science case has been developed. This has been done both for the reason of understanding the complete 'process' of developing a VCS platform (see the beginning of this chapter), and also to minimise any disruption to an existing scientific team. For instance, manipulating the task workflow design of a 'live' platform could have a profound effect on the established volunteer community, and the results they produce, meaning that any such change would require the agreement of the science team involved. By having a dedicated science case for this

[14] The planet four VCS platform is available at: http://www.planetfour.org

[15] Details on the original Planet Four science case are available at: http://www.planetfour.org/#/about

thesis, there has been the freedom to make modifications in a timely fashion without causing disruption to any external parties' research.

For the thesis, the dedicated science case involves the marking of craters on the surface of Mars. It has been developed in collaboration with geophysicists based at the University of Bristol, who can use the results to help build on their research regarding the age of the surface (Taylor, Teanby, & Wookey, 2013). Ultimately, it is hoped the data can be used to inform future missions to the planet, and a more detailed explanation of the science case can be found in the empirical part of this thesis.

5.4.2 Volunteer-Expert Data Comparison

An advantage of having a dedicated science case for the thesis, and an associated science team engaged with the process, is that the data collated can be compared and contrasted with previously collected expert data and findings. This technique has been used before in order to evaluate the success of VCS platforms, and also directly in a crater-counting, planetary science setting (Robbins et al., 2014). As such, the level of agreement between expert and volunteer data has been used to evaluate the scientific outcome of the experiments in this thesis, as outlined in section 4.5.3. Although there have been issues identified regarding the use of expert results as a 'ground truth', particularly in terms of the variability of agreement between different 'experts' using different analyst tools (see the Robbins work above), it is envisaged that the science teams' specific knowledge regarding the science case developed and the geographical area chosen ensures a reasonable ground truth comparison for this work.

5.4.3 Participant Issues

As explained by Reips (2000), self-selection bias is a potential issue for online experiments. Who is willing to take part? Who provides the most data? Who is asked to participate? The answers to these questions could mean that certain types of individual are over or under-represented in the data. The problem can be further confounded by individual motivation and demographic factors such as education and exposure to technology.

However, it is hoped that this is less of an issue for the review study in chapter 7, as the Planet Four platform has been studied over a significant section of its life from launch to ensure all types of volunteer are represented. Likewise for the experimental studies of this thesis the within-subjects design should mitigate individual differences, as the unit of analysis is the individual. Fluctuations in within-subject measurements are compared to each other per individual when investigating volunteer experience and data contribution while aggregated results are used only when appropriate, i.e. when evaluating volunteer agreement or average marking position.

5.4.4 Ethical Issues

For studies involving human participants (chapters 7, 8 and 9), both the Zooniverse consortium and the Nottingham Geospatial Institute of the University of Nottingham, where the laboratory study and 'live' study were supervised, have guidelines for appropriate ethical conduct. In all cases, the University of Nottingham Engineering faculty's ethical committee approved the studies' aims, methods and procedures, especially in terms of guaranteeing anonymisation of data and prior informed consent of the participants.

Beyond the more formal ethical issues that are standard practice to address, the discipline of citizen science has some specific ethical issues regarding the treatment of its volunteers (Bonney et al., 2014) – with a key principle identified as 'not wasting peoples' time' (Kintisch, 2011). Although defining what does or does not constitute a waste of peoples' time is deemed beyond the focus of this thesis, by developing a specific science case for the experimental research involving an invested science team, it ensures that there is at least the possibility that the project could continue beyond the duration of the thesis to address a legitimate science case. Likewise, the aims of the project in terms of contributing knowledge to VCS platform design have been clearly communicated to the volunteer community to ensure nobody has felt misled.

5.5 STUDIES IN THIS THESIS

While the research questions that link the two review studies and two experimental studies were stated in 4.5.4, the methods of the studies are now briefly contrasted and summarised.

The two review-type studies reported in chapters 6 and 7 are conceptually similar in that they both consider the current landscape of VCS platforms, the first adapting the technique of hierarchical task analysis to review existing task workflow designs, the second through questionnaires involving Likert-type items and free-text responses to review the thoughts and opinions of existing volunteers. Also, both these studies take place in a real world setting, reviewing current 'live' Zooniverse platforms and involving a current 'live' platform active community.

Building on the knowledge gained regarding TWD and volunteer behaviour in the review studies, the experimental studies reported in chapters 8 and 9 move forward to directly investigate the manipulation of TWD factors. Through the development of a new science case for the Zooniverse platform Planet Four, they test aspects of scientific outcome through three different TWD interfaces in a hypothesis-driven fashion informed by the postulates stated in section 4.4.3. This is done through the derivation of controlled and counter-balanced independent variables as part of a repeated-measurements, within subject design. Scientific analysis and behavioural data is gathered in situ for later data analysis, while post-hoc questionnaires serve to contextualise and enrich the data qualitatively.

In experiment 1 (chapter 8) a laboratory setting is used, involving 30 participants taking part in controlled conditions. This is done not only for the research focus reasons mentioned, but also as a check for more general usability and error-catching issues regarding the website design. Experiment 2 (chapter 9) launches the platform 'into the wild', advertising it to the wider volunteer community through Zooniverse channels. Rather than seeing this experiment as disparate from the previous lab study, it can be seen as a natural progression and an evaluation of the developed task workflow design interfaces in a real-world application.

Bringing the research questions posed in 4.5.4 back into focus, the overarching theme of the studies and experiments is the exploration of the

ways scientific outcome can be affected by different strategies regarding task workflow design. While the studies in chapters 6 and 7 concentrate on existing TWD strategies and their effect on volunteer behaviour, the experiments in chapters 8 and 9 explicitly compare different task workflow designs in terms of *task type*, *judgement, autonomy* and *variety*, using factors of scientific outcome as dependent measures.

5.6 SUMMARY

This concludes Part II of the thesis. In chapter 4, the approach to studying task workflow design was developed, with the conceptual model of task workflow design factors, either science case and design influenced, at its centre. The argument of practical applicability served to define the focus of this thesis on investigating the manipulation of design-influenced TWD factors as the key adaptation of a VCS platform. The definition of this focus in turn resulted in the principal research question: "what task workflow designs ensure the best scientific outcomes?" and three component sub-questions that are at the heart of the empirical studies.

While chapter 4 in essence detailed the '*what*' of this thesis, this chapter concentrates on the '*how*'. Specifically, how does the thesis study the effect of task workflow design on the scientific outcomes of VCS platforms? The chapter firstly described the focus of the empirical work on the Zooniverse (see 5.1), both due to its importance to the discipline and its links with the author's institution. The methodology underlying the studies is then outlined, with a focus on the mixed methods approach (see 5.2). Specifically, the hierarchical task analysis technique and questionnaire design are explained in order to gather both quantitative and qualitative observations. With an emphasis on experimentation 'in the wild' the Planet Four platform is introduced, along with a specially derived scientific case, as the focus for the experimental studies of the thesis (see 5.3). Finally, the application of the methodology to the empirical work was contrasted and summarised (see 5.4).

The following four chapters of Part III together comprise the account of the empirical work of this thesis.

PART III
EMPIRICAL WORK

6

THE TASKS OF CITIZEN SCIENCE PLATFORMS

This chapter presents a study to identify the different task workflow designs currently employed by virtual citizen science platforms, and the factors that influence this design. Additionally, it investigates the effects of task workflow design factors on volunteer behaviour. A hierarchical task analysis including measures of volunteer freedom shows that task workflow design is independent of the scientific discipline. Statistical analysis using website analytics shows that certain task workflow factors have a significant effect on volunteer behaviour regarding the total person hours spent on the platform, the percentage of return visits, and the average time spent performing the science task.

6.1 INTRODUCTION

The absence of consideration of task workflow design factors – such as task type, judgement and freedom – in most research defining VCS platform structures has been noted previously (see chapter 4). Current research defines such systems by the functionality (Prestopnik & Crowston, 2012) and associated usability (Reed et al., 2012) of the site as a whole, rather than considering the specific scientific task required of the volunteer. One study of note that does consider the types of user task required (Wiggins & Crowston, 2012) investigates *Observer* type 'in the field' citizen science projects, rather than the *Classifier* type VCS platforms that are the focus of this thesis. Additionally, the absence of task-focussed design is reflected in the experimental methods of research regarding volunteer behaviour and performance. Instead, the effects of motivation (Raddick et al., 2010; 2013; Rotman et al., 2012) and involvement in community forums (Luczak-Roesch et al., 2014) form the basis of the majority of this research.

The lack of formal research regarding task design however doesn't necessarily indicate that developers of citizen science projects are failing to think in a design-led fashion. Recent work has begun to derive implications for design practice based on the experience of developing a project (Simpson et al., 2014) and the engagement patterns of their community (Ponciano et al., 2014). This, coupled with the author's informal conversations with the Zooniverse development team, demonstrates a hunger in the citizen science community to better understand how task design can influence project performance.

Although task workflow design factors are acknowledged as important characteristics that affect user motivation, satisfaction and performance (see 3.2 and 3.6), this knowledge has yet to be applied directly to virtual citizen science. The one study that does mention the importance of task design suggests that tasks should be designed to fit the volunteer profile and encourage 'one more click' from the majority user group who visit the site only once or twice (Eveleigh et al., 2014). This thesis however chooses to take a different approach, investigating how task workflow design can affect user motivation and performance, and ultimately even alter the user profile of VCS platforms.

If it is accepted that existing VCS platforms have not been informed by research into the effect of task workflow design on motivation and performance, the factors that might have influenced their design need to be understood. Therefore, this study firstly explores the different task workflow designs that currently exist regarding VCS platforms, and attempts to identify any pattern in this design and associated reasoning and/or restrictions affecting this choice.

Secondly, in the reviewed literature no study was found that experimentally researched the effect of task workflow design factors on VCS volunteer behaviour. Therefore, this study will investigate if such an effect exists by comparing task workflow design variations against different website analytic measures of visitor traffic.

6.2 RESEARCH QUESTIONS

We now proceed to study current virtual citizen science platforms with regards to their task workflow design. The experiment defines task workflow design by its constituent parts outlined in chapter 4, namely: *User Task Type, Judgement, Task Variety* and *Autonomy.* These factors are compared separately and in unison across a range of scientific disciplines, in an attempt to answer the following research question, introduced in chapter 4:

RQ_1: What Task Workflow Designs are currently used by Citizen Science Projects?

Furthermore, in understanding the task workflow designs that are currently employed, and accepting the hypothesis that they are not informed by volunteer behaviour considerations, this study attempts to answer:

RQ_{1a}: Why are certain Task Workflow Designs used?

Finally, by comparing the different designs currently employed by VCS platforms with associated web analytics regarding visitor metrics, this study attempts to answer the question combining the two:

RQ_2: Does Task Workflow Design have an effect on volunteer behaviour?

While at first glance the connection between this question and the overarching title of the thesis may not be obvious, understanding the relationship between task workflow and volunteer behaviour will give insight into its effect on the volunteer profile of VCS platforms, which directly relates to volunteer experience and retention rates. These factors have been shown to influence the scientific output of such projects (Ponciano et al., 2014, see section 2.6.4; Curran et al., 2009, see section 3.4), and so their relevance is clear.

6.3 EXPERIMENTAL DESIGN

6.3.1 Current Virtual Citizen Science Platforms

For this study, 12 Zooniverse VCS projects have been chosen to represent the current environment regarding task workflow design, as outlined in table 6.1. They cover a broad range of scientific disciplines, requiring volunteers to analyse an array of differing data formats. As Zooniverse platforms are created using a similar design template by the same development team, any effects due to broader usability issues (visuals, general website navigation etc.) or organisational norms should be minimised.

Table 6.1: Zooniverse Virtual Citizen Science platforms

Zooniverse Platform	URL	Scientific Discipline	Date of Launch
Moon Zoo	moonzoo.org	Astronomy	Feb 2009
Solar Storm Watch	solarstormwatch.com	Astronomy	Dec 2009
Planet Hunters	planethunters.org	Astronomy	Dec 2010
Milky Way Project	milkywayproject.org	Astronomy	Dec 2010
Planet Four	planetfour.org	Astronomy	Jan 2013
Old Weather	oldweather.org	Climate	Oct 2010
Cyclone Center	cyclonecenter.org	Climate	Sept 2012
Ancient Lives	ancientlives.org	Humanities	Jul 2011
Whale FM	whale.fm	Nature	Nov 2011
Seafloor Explorer	seafloorexplorer.org	Nature	Sept 2012
Bat Detective	batdetective.org	Nature	Oct 2012
Cell Slider	cellslider.net	Biology	Oct 2012

As explained in section 2.4, another compelling argument for reviewing Zooniverse VCS platforms is due to their position within the discipline. As one of the earliest coupled with varying forms of media attention, the Zooniverse consortium has amassed a large following through its projects, with over a million registered users. This has cemented its position as one of the most popular, experienced and influential VCS organisations. Access to data regarding platform performance in terms of visitor retention and behaviour has been established through the University of Nottingham's

membership of the Citizen Science Alliance, the founders of the Zooniverse, providing another, more pragmatic, reason for the platform choice.

6.3.2 Hierarchical Task Analysis (HTA)

Each of the 12 selected Zooniverse VCS projects was analysed by way of a hierarchical task analysis, as described in chapter 4. The overall scientific task required of the volunteer for each project was decomposed into the set of subtasks that had to be completed in order to achieve this goal. Flow diagrams were generated describing the tasks carried out by users in order to complete a classification of a subject, as required by the specific science case of the platform. These classifications included *"marking the position and size of a crater"*, *"identifying the type of a galaxy"* or *"transcribing a ships' logbook entry"*. Extraneous tasks relating to general website use and navigation (viewing developer credits, feature examples etc.) were not included in the analysis. Figure 6.1 shows an example of one of the flow diagrams, with the volunteer tasks to complete a classification on the Seafloor Explorer citizen science project displayed.

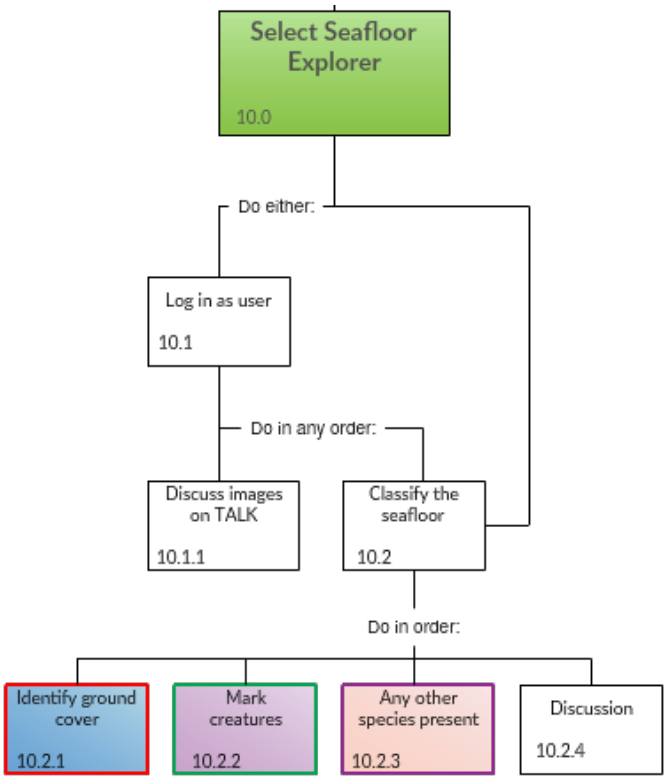


Figure 6.1: HTA flow diagram for Seafloor Explorer website

Additionally, the hierarchical task analysis of each project reviewed the order and number of tasks involved for a classification, and initially classified each task involved as using a 'tool type' (illustrated by the different box colours in figure 6.1) requiring a participant judgement (the border colour). User freedom was also recorded regarding each project, i.e. does the platform require tasks to be completed in a certain order by the volunteer, or can each volunteer choose their own preferred approach?

The goal of the HTA was to review the different task workflow design models currently being used by virtual citizen science projects, specifically in terms of the four factors outlined in section 6.2. Therefore the tasks for each platform were further categorised in terms of their type and required volunteer judgement, based on the definitions outlined by Pelli & Farell (2010, see 3.2). Table 6.2 gives an overview of the task type and volunteer judgement categories, along with their definitions.

Table 6.2: Task type & volunteer judgement categories

Task Types	Volunteer Judgements
Detection *The goal of the volunteer is to determine if a certain stimulus is present or not, usually through making a yes/no or threshold judgement.*	**Forced Choice** *The user is forced to categorise the presented stimuli using one of the allowed response alternatives. Usually required during identification or discrimination tasks.*
Direct Scaling *The goal of the volunteer is to provide a scale of some perceptual magnitude (brightness, rockiness etc.) as a function of the corresponding physical dimension.*	**Matching** *The user is asked to adjust an annotation to match the stimulus based on a particular metric, i.e. "mark the circumference of the crater". It could be argued this also involves a threshold judgement when deciding the annotation endpoints.*
Discrimination *The goal of the volunteer is to indicate the difference between two stimuli based on a physical dimension, i.e. "which image has most rocks in it". Usually requires a forced choice or yes/no judgement.*	**Rating Scale** *Either a number of discrete ratings, or a continuous scale. The endpoints of the rating scale are definitive i.e. "the image is definitely blank" and "the image is definitely not blank", with intermediate degrees of confidence in between.*
Identification *The goal of the volunteer is to identify the stimulus type based on a pre-defined classification, i.e. "what type of galaxy is present, spiral or smooth?"*	**Threshold** *The user has to adjust a stimulus contrast measure to the point where it is "just barely detectable", i.e. "paused the video when the solar flare first appears".*
Matching *The goal of the volunteer is to physically annotate the image to match the stimulus, i.e. "mark the length of each fish".*	**Yes/No** *The user is only allowed two response alternatives, yes or no. Usually required during detection or discrimination tasks.*

While Pelli & Farell's original use for this categorisation was to provide a general user-guide in the use of psychophysical methods when answering practical questions related to perception and performance, there is significant cross-over with the types of tasks and judgements required of VCS volunteers. As such, and in the absence of a pre-existing formal categorisation focussed on virtual citizen science, it has been used as the basis for this study.

6.3.3 Measuring Volunteer Behaviour through Website Analytics

Website analytics of the 12 platforms were studied in order to better understand their volunteer behaviour, provided by the Google Analytics platform and courtesy of the Zooniverse development team. This allowed not only the recording of general website statistics such as *total number of visitors* and *total person hours* spent on each site, but also the analysis of more specific metrics such as *average time spent classifying per visit* (volunteers attempting the main scientific tasks) and *percentage of return visits* (an indication of the volunteer retention success of the site). The statistics reviewed were:

- ***Number of Return Visits*** – number of visits by volunteers who have visited the platform for the second or greater time
- ***Number of New Visits*** – number of visits by volunteers who have visited the platform for the first time
- ***Total number of Visits*** – the sum of the previous two statistics
- ***Average Time Classifying per Visit*** – the average time each volunteer spent on the 'classify' page of the platform per visit
- ***Total Person Hours on Platform*** – the total hours spent using the platform by all the volunteers combined

Platform statistics were analysed over three different time periods: from official launch date to 30th June 2013, an arbitrary date when the figures were collated for the study; the first 30 days since launch; and the previous 30 days counting back from 30th June 2013. This was in order to better understand volunteer behaviour over three distinctly different phases of a VCS project: its long-term life, immediately after launch when media attention and promotion is at its peak, and a more settled period when the site becomes 'self-sustaining' (i.e. little external promotion or traffic drive).

Finally, the collated statistics for all three time-spans were compared with the task workflow design models defined for each platform during the hierarchical task analysis of the study. Specifically, the factors that make up task workflow design were considered, in order to investigate any effect that they have on volunteer behaviour.

6.4 ANALYSIS AND RESULTS

6.4.1 Hierarchical Task Analysis Findings

The HTA revealed that the 12 studied platforms had a total number of tasks ranging between 2 and 9 (mean of 3.83, median of 3, SD = 1.90) that needed to be completed in order for a volunteer to make a scientific classification. When considering task and judgement variety, this ranged from 1 to 4 different types of task and judgement required (mean of 2.33, median of 2, SD = 0.90; and mean of 2.50, median of 2.50, SD = 0.90 respectively). In terms of user freedom, 8 of the 12 platforms required the volunteer to perform the tasks in a prescribed order, with 4 of the platforms allowing the user to take their own approach. Table 6.3 shows a full breakdown of the task workflow design factors for each of the 12 platforms.

When considering the effect of the scientific discipline involved on task workflow factors (see table 6.4), no significant difference was found regarding task and judgement variety. For all of the five disciplines, the average number of different task types and judgements ranged between 2 and 3 (mean and median). Likewise little effect was found regarding the total number of tasks to be completed to make a classification, with 4 of the 5 disciplines averaging between 4 and 5 tasks (mean, median ranged between 3 and 5). The one exception was the *Nature* discipline, with a mean of 2.67 tasks, SD = 0.58 (median of 3). An effect is noticeable when considering user freedom however. For 3 disciplines (*Climate, Nature* and *Biology*) all projects forced the volunteer to complete the tasks in a predefined order, while for the *Humanities* discipline, the opposite was true. The *Astronomy* discipline was the only one considered that had projects in the study that either allowed the user to take their own approach (60%), or forced them to follow a set order (40%).

Table 6.3: Hierarchical Task Analysis results

Zooniverse Platform	No. of Tasks	Task Types Used		User Judgements Required		User Freedom
Moon Zoo	3	Discrimination Direct Scaling Matching	1 1 1	Forced Choice Rating Scale Matching	1 1 1	Any order
Solar Storm Watch	9	Detection Matching	8 1	Forced Choice Matching Threshold	1 1 7	In order
Planet Hunters	4	Detection Direct Scaling Identification Matching	1 1 1 1	Forced Choice Matching Rating Scale Yes/No	1 1 1 1	In order
Milky Way Project	3	Identification Matching	1 2	Forced Choice Matching	1 2	Any order
Planet Four	2	Matching	2	Matching	2	Any order
Old Weather	3	Detection Identification	2 1	Forced Choice Threshold	1 2	In order
Cyclone Center	5	Discrimination Identification Matching	3 1 1	Forced Choice Matching	4 1	In order
Ancient Lives	5	Detection Identification	3 2	Forced Choice Threshold Yes/No	2 2 1	Any order
Whale FM	2	Discrimination	2	Forced Choice Yes/No	1 1	In order
Seafloor Explorer	3	Detection Identification Matching	1 1 1	Forced Choice Matching Yes/No	1 1 1	In order
Bat Detective	3	Discrimination Detection	2 1	Forced Choice Threshold Yes/No	1 1 1	In order
Cell Slider	4	Discrimination Direct Scaling Identification	1 2 1	Forced Choice Rating Scale	2 2	In order

Table 6.4: Task Workflow Factors by discipline

Scientific Discipline	No. of Tasks	No. of Task Types Used	No. of User Judgements	User Freedom
Astronomy	Mean 4.20 Median 3 (SD 2.77)	Mean 2.40 Median 2 (SD 1.14)	Mean 2.60 Median 3 (SD 1.14)	Set Order 2 User Choice 3
Climate	Mean 4.00 Median 4 (SD 1.41)	Mean 2.50 Median 2.5 (SD 0.71)	Mean 2.00 Median 2 (SD 0.00)	Set Order 2 User Choice 0
Humanities	Mean 5.00 Median 5 (SD 0.00)	Mean 2.00 Median 2 (SD 0.00)	Mean 3.00 Median 3 (SD 0.00)	Set Order 0 User Choice 1
Nature	Mean 2.67 Median 3 (SD 0.58)	Mean 2.00 Median 2 (SD 1.00)	Mean 2.67 Median 3 (SD 0.58)	Set Order 3 User Choice 0
Biology	Mean 4.00 Median 4 (SD 0.00)	Mean 3.00 Median 3 (SD 0.00)	Mean 2.00 Median 2 (SD 0.00)	Set Order 1 User Choice 0

In terms of the specific types of task used by each discipline as defined by Pelli & Farell in table 6.2, the HTA revealed an absence of preference, with 3 of the 5 disciplines (*Astronomy*, *Climate* and *Nature*) each using to some degree at least 4 of the 5 different defined types (see table 6.5). For the two disciplines that did not (*Humanities* using 2 types and *Biology* 3), only one platform from each was included in the analysis, and so this lack of variety agrees with the findings in table 5.4. Regarding the prevalence of each task type per platform in each discipline, again no preference is clear. While it could be argued, for instance, *Astronomy* platforms involve more Matching tasks (mean of 47.55%) compared to Discrimination tasks (mean 6.67%) the relatively high SD value for each (35.90 and 14.91 respectively) indicates a large variation depending on the platform. This trend is true of all of the disciplines where more than one platform has been studied (see table 6.5).

Table 6.5: Mean average of each task type per discipline

Scientific Discipline	% of Detection	% of Direct Scaling	% of Discrim.	% of ID	% of Matching
Astronomy	22.78 (SD 38.51)	11.67 (SD 16.24)	6.67 (SD 14.91)	11.67 (SD 16.24)	47.22 (SD 35.90)
Climate	33.33 (SD 47.14)	0.00 (SD 0.00)	30.00 (SD 42.43)	26.67 (SD 9.43)	10.00 (SD 14.14)
Humanities	60.00 (SD 0.00)	0.00 (SD 0.00)	0.00 (SD 0.00)	40.00 (SD 0.00)	0.00 (SD 0.00)
Nature	22.22 (SD 19.24)	0.00 (SD 0.00)	55.55 (SD 50.92)	11.11 (SD 19.24)	11.11 (SD 19.24)
Biology	0.00 (SD 0.00)	50.00 (SD 0.00)	25.00 (SD 0.00)	25.00 (SD 0.00)	0.00 (SD 0.00)

The HTA likewise showed little preference regarding the associated user judgements required by each discipline. 4 of the 5 (*Astronomy*, *Climate*, *Humanities* and *Nature*) all involved at least 3 different judgement types, while *Biology* incorporated 2 (with only one platform included in the study). Again, regarding the prevalence of each user judgement per platform, no preference is clear with high SD values indicating a large variation depending on the platform (see table 6.6).

Table 6.6: Mean average of each judgement type per discipline

Scientific Discipline	% of Forced Choice	% of Matching	% of Rating Scale	% of Threshold	% of Yes/No
Astronomy	20.55 (SD 14.64)	47.22 (SD 35.90)	11.67 (SD 16.24)	15.55 (SD 34.78)	5.00 (SD 11.18)
Climate	56.67 (SD 33.00)	10.00 (SD 14.14)	0.00 (SD 0.00)	33.33 (SD 47.14)	0.00 (SD 0.00)
Humanities	40.00 (SD 0.00)	0.00 (SD 0.00)	0.00 (SD 0.00)	40.00 (SD 0.00)	20.00 (SD 0.00)
Nature	38.89 (SD 9.62)	11.11 (SD 19.24)	0.00 (SD 0.00)	11.11 (SD 19.24)	38.89 (SD 9.62)
Biology	50.00 (SD 0.00)	0.00 (SD 0.00)	50.00 (SD 0.00)	0.00 (SD 0.00)	0.00 (SD 0.00)

6.4.2 Website Analytics

Table 6.7 shows the website statistics for each of the 12 Zooniverse platforms, as recorded by the Google Analytics service and defined in section 5.3.3, for the three different time periods described. For each of the projects, considerably more visits from volunteers occurred during the first 30 days since launch rather than the last 30 days up to the 30th June 2013, in some cases the ratio being above 10:1. Highlighting this early period in more detail, figure 6.2 shows that each platform receives most of its visits within the first few days since introduction, and then rapidly falls to a lower "long tail" level of sustained activity – independent of the initial visit 'peak'.

The average time a volunteer spends classifying on each of the platforms ranges from a minimum average of 59.24s (*Planet Hunters*) to a maximum of 604.50s (*Milky Way Project*) over the full time period studied. While no measures of variance across volunteers regarding classifying time is freely available from the Google Analytics service, Its effect can be seen regarding the total person hours spent on each site. For instance, although *Planet Hunters* has had over three times the total number of visits than *Old Weather* (1,461,268 verses 411,685), the total person hours committed to using the site is less (24,045.98 verses 28,807.66) due to *Old Weather's* superior average classifying time.

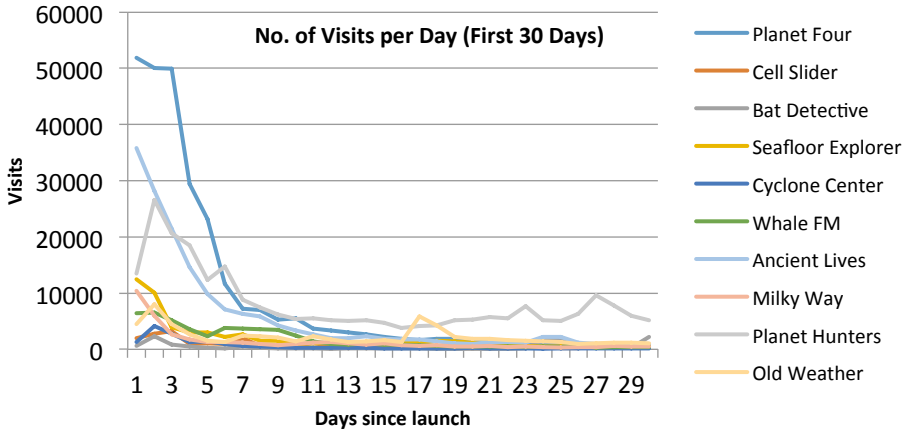


Figure 6.2: Attrition of visits in VCS projects

Table 6.7: Website statistics for each Zooniverse platform

Zooniverse Platform	Total No. Of Visits	No. Of New Visits	No. Of Return Visits	Average Time Classifying (Seconds)	Total Person Hours
Moon Zoo	663491	480130	183361	190	35018
First 30 days	123616	96145	27471	163	5597
Last 30 days	6354	4962	1392	215	379
Solar Storm Watch	441394	291554	149840	497	60937
	67606	48506	19100	592	11117
	4156	2724	1432	348	402
Planet Hunters	1461268	850423	610845	59	24046
	246297	142859	103438	58	3975
	18850	12180	6670	61	318
Milky Way Project	351814	269411	82403	605	59075
	40203	27797	12406	219	2451
	7427	6903	524	944	1948
Planet Four	337045	219307	117738	312	29201
	275454	189004	86450	310	23748
	8383	4788	3595	245	570
Old Weather	411685	188038	223647	252	28808
	67669	39518	28151	271	5091
	4343	2214	2129	163	197
Cyclone Center	38366	26876	11490	562	5992
	17425	13672	3753	473	2291
	1097	538	559	736	224
Ancient Lives	394321	264021	130300	181	19772
	168118	123358	44760	189	8848
	18486	14842	3644	149	765
Whale FM	129999	97496	32503	196	7079
	53647	42206	11441	207	3090
	3625	2973	652	154	155
Seafloor Explorer	111093	63541	47552	230	7088
	58963	38077	20886	204	3337
	522	253	269	267	39
Bat Detective	25434	19540	5894	142	1002
	14651	12241	2410	167	682
	1461	1046	415	115	47
Cell Slider	251313	211829	39484	342	23889
	23507	18698	4809	249	1624
	9328	4355	4973	364	943

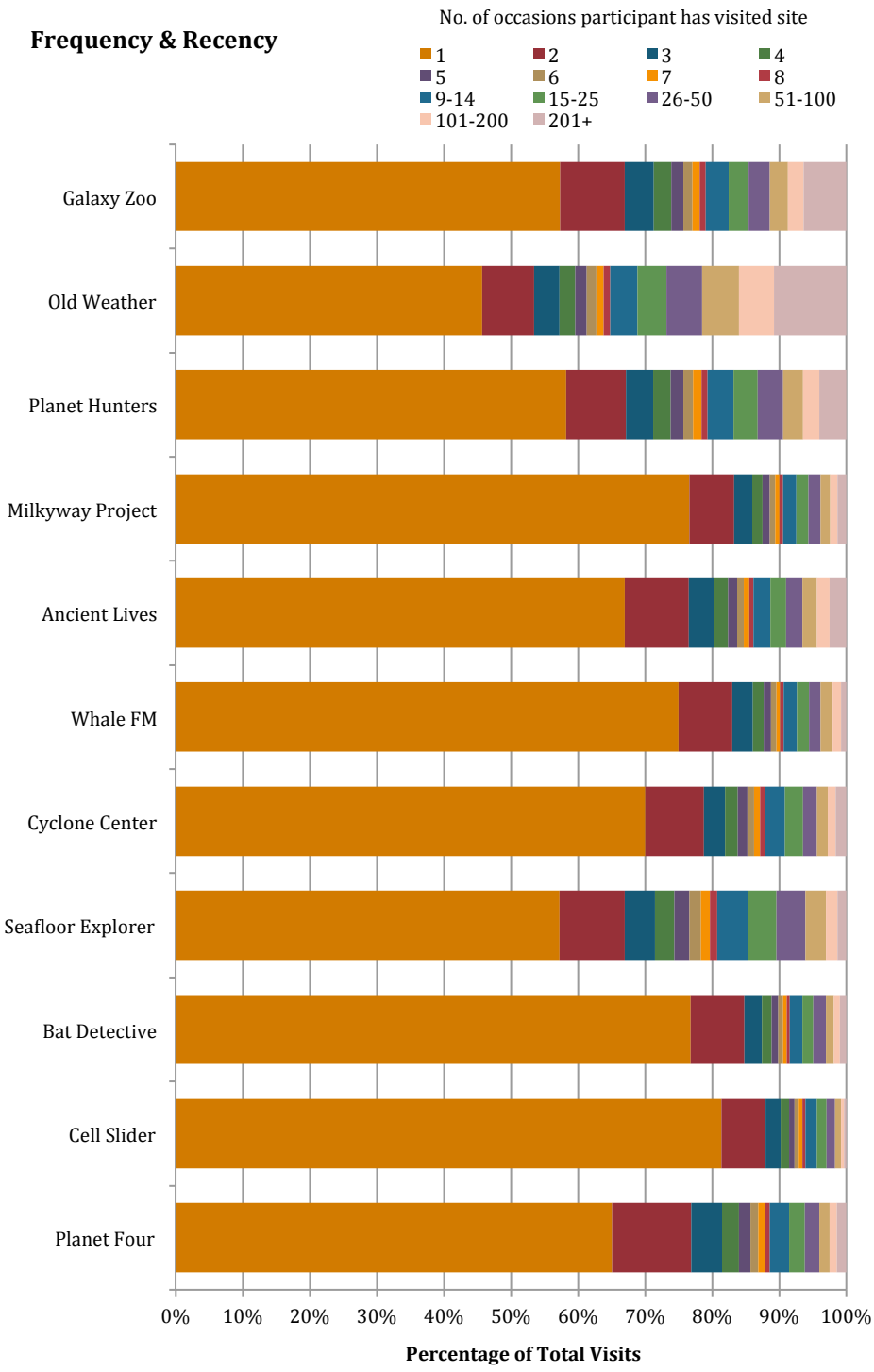


Figure 6.3: Visit frequency of volunteers

When considering the number of new visits versus the number of return visits by volunteers for the 12 platforms, the study results broadly agree with the findings of Ponciano et al. (2014, see section 2.6.4), where the majority of volunteers only visit once. Figure 6.3 shows a breakdown of the number of occasions a volunteer has visited each site, as a percentage of

the total number of visits. For 11 of the 12 platforms (data for *Solar Storm Watch* was unavailable), over 50% of visits were by first time volunteers with this figure increasing to between 60-80% for 8 of the projects. The highest percentage of first time volunteers belonged to the *Cell Slider* site, with new visits making up 81.36% of the total. Bucking the trend (and Ponciano et al.'s findings), the *Old Weather* platform's total visits is made up of less than 50% of new volunteers, with over 10% of visits coming from volunteers returning for the 201st time or greater.

6.4.3 The Effect of Task Workflow Design on Volunteer Behaviour

In order to analyse whether the effect of different task workflow designs for each platform yielded significantly different measurements of volunteer behaviour, firstly scatterplots were derived of the independent variables (IV) – the task workflow design factors, verses the dependent variables (DV) – the website statistics. For those that appeared to display a monotonic relationship between the two variables (i.e. they increase together, or as one increases the other decreases), a Spearman's Rank-Order Correlation test was carried out in order to measure its strength and direction.

6.4.3.1 Task Type

Figure 6.4 shows scatterplots of the percentage of discrimination tasks per platform (IV) verses total person hours (DV), and the percentage of detection tasks per platform (IV) verses the percentage of return visits to the site (DV) respectively.

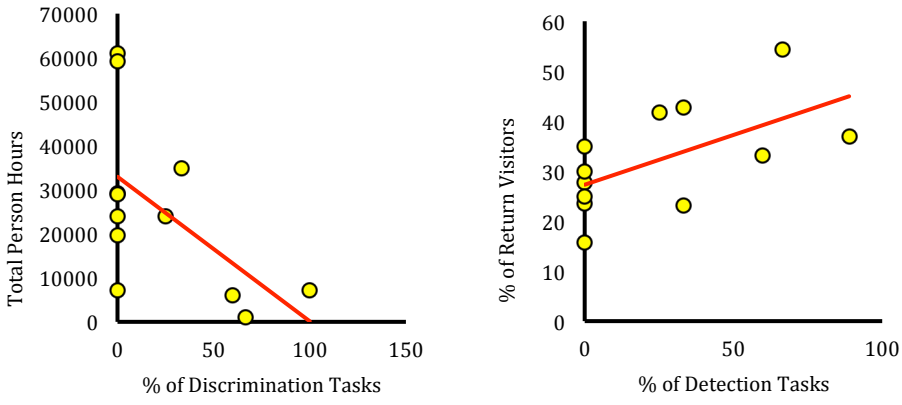


Figure 6.4: Task type verses volunteer behaviour

The Spearman's test showed that there was a significant negative correlation between the percentage of discrimination tasks and the total person hours spent on the site ($\rho(12) = -.608$, $p = .036$), and a significant positive correlation between the percentage of detection tasks and the percentage of return visits to the site ($\rho(12) = .583$, $p = .046$).

6.4.3.2 User Judgement

Figure 6.5 shows scatterplots of the percentage of yes/no judgements per platform (IV) and the percentage of forced choice judgements per platform (IV) verses the total person hours (DV) respectively.

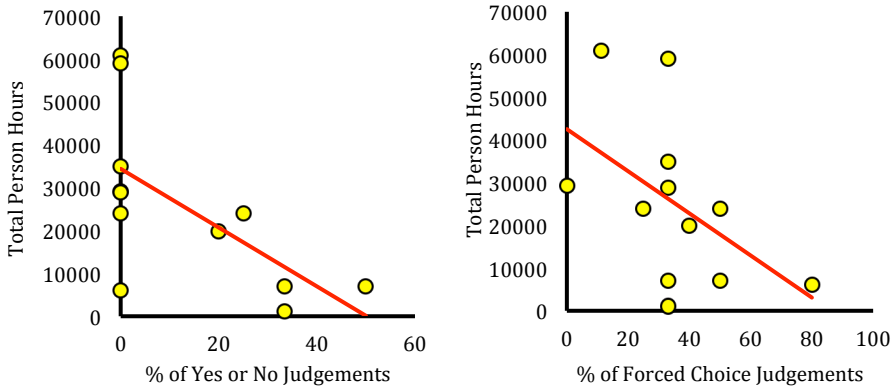


Figure 6.5: User judgement verses volunteer behaviour

The Spearman's test showed that there was a significant negative correlation both between the percentage of yes/no judgements and the percentage of forced choice judgements and the total person hours spent on each site ($\rho(12) = -.653$, $p = .021$; and $\rho(12) = -.621$, $p = .031$ respectively).

6.4.3.3 Judgement Variety and Freedom

Figure 6.6 shows a scatterplot of the number of different user judgements required per platform (IV) verses the average time spent classifying per visit (DV), and a boxplot of the total person hours (DV) depending on user freedom (IV).

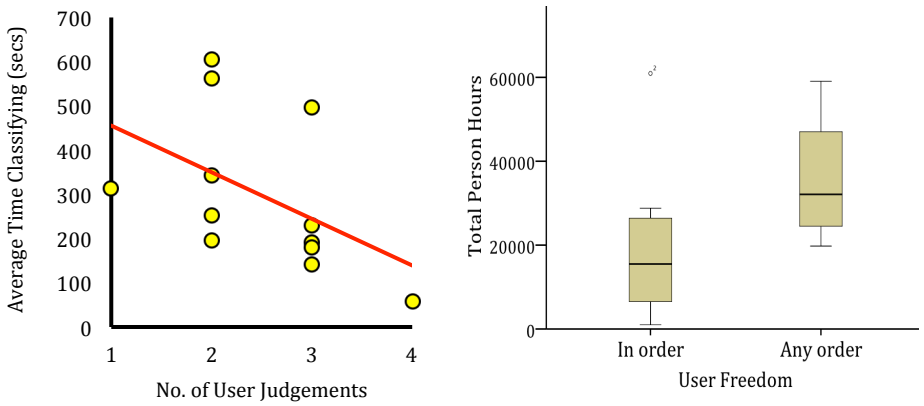


Figure 6.6: Variety and freedom verses volunteer behaviour

The Spearman's test showed that there was a significant negative correlation between the number of different user judgements required per platform and the average time spent classifying per visit ($\rho(12) = -.648$, $p = .023$). Considering platforms grouped in terms of freedom, platforms that required the volunteer to complete tasks in a set order had significantly less total person hours spent on them (13986.32 ± 11156.00 hours) compared to those that allow the volunteer to take their own approach (35766.53 ± 16761.04 hours), $t(12) = -2.615$, $p = 0.028$.

6.5 DISCUSSION

6.5.1 Limitations

The method of using a sample of virtual citizen science platforms to represent the current landscape of the discipline in terms of task workflow design, along with selecting a set time period to study volunteer behaviour by way of website analytics, clearly has its challenges and limitations.

6.5.1.1 Platform Selection Bias

The advantages of using 12 Zooniverse sites for the study, in terms of separating out any effects due to other design issues (appearance, general functionality etc.), have been explained in the experimental design section of this chapter. However, it could be argued that bias might be introduced through developer habit and company routine i.e. task workflow designs being favoured as they have been used before and 'have done the job'. It is

hoped that due to the Zooniverse's wide range of platforms, the use of a team of developers who each have differing involvement, and the various science teams that are part of the process of creating each site, such bias through routine has been kept to a minimum.

6.5.1.2 Volunteer Behaviour Bias

While the study has considered volunteer behaviour (by way of website analytics) over three time periods to get a fuller picture of its evolution, certain external influences that might bias the statistics are unknown. For instance, any downtime experienced by any of the projects (i.e. periods when volunteers cannot access the site) that lasted less than a day has not been recorded by the Google Analytics service, and therefore cannot be accounted for in this study. No instances of day or longer outage occurred during the study period. It is hoped that by using projects that are part of a long-standing, experienced platform, and by considering volunteer behaviour over a long time period (in some cases several years), the effect of these issues are minimalized.

6.5.2 Hierarchical Task Analysis

In attempting to better understand the task workflow designs incorporated into current VCS projects, the HTA has revealed a contrasting picture of conformity and dissimilarity depending on the factor considered.

While each of the projects generally require volunteers to complete a similar variety of different task types and judgements (around 2 or 3), the total number of tasks that need to be completed in order to make a classification varies considerably. In terms of the specific types of task and judgement used by each project as defined by Pelli & Farell no preference was found, with each of the categories defined for both featuring to some extent. This lack of preference continues when considering the scientific discipline involved.

When considering the volunteer's freedom to complete the tasks how they choose, there does appear to be a preference for constricting the volunteer to follow a set path, with two thirds of the projects studied designed with a set task completion order. Although this freedom or lack of it appears at

first glance to be dependent on the scientific discipline involved (with 4 of the 5 disciplines being polarised one way or the other), the significance of this result is clearly reduced due to the limited number of projects considered for some of the disciplines.

6.5.3 Website Analytics

In using website analytics in order to investigate volunteer behaviour, it has been generally found that in terms of number of visits there is agreement with the previous literature (Ponciano et al., 2014 - see 2.6.4), with each of the sites taking on a 'long-tail' profile independent of their initial launch peak. This does not tell the whole story however, with large variations in average classifying times per visit present – that has directly impacted on the total person hours spent on each site.

A prevalent comparison to make is that between the *Planet Hunters* and *Old Weather* projects. The Planet Hunters site follows the Pareto distribution, with a massive peak of visits in the first few days that falls away with the majority of users only visiting once. In contrast, the Old Weather site had a much smaller peak (due to considerably less media attention at launch) that again takes a long-tail profile but it's tail has fallen away less steeply, with over 50% of its visits coming from return volunteers, and over 10% coming from volunteers returning for over the 200th time. This suggests that although at face value Planet Hunters has been the more successful in terms of number of visits, Old Weather has been able to create a much larger community of committed volunteers that return to the site on a regular basis.

6.5.4 Task Workflow Design vs. Volunteer Behaviour

In comparing the task workflow design factors of current VCS platforms with the measures of volunteer behaviour recorded, various notable findings have been made.

In terms of task type, it has been found that the percentage of discrimination tasks involved in a project has a significant negative effect on the total person hours spent on the site, while the percentage of detection tasks has a significant positive effect on the percentage of return

users to a site. It could be argued that this is in agreement with the literature as described in section 3.6.2, where the less complex discrimination task (which is bigger/brighter etc.) has resulted in reduced volunteer satisfaction while the more complex, involved detection task (when/where is something present) has resulted in more volunteers returning to the site (Chung-Yan, 2010; Ghani & Deshpande, 1994).

Regarding the user judgements involved, it was found that both the percentage of yes/no and forced choice judgements involved in a project has a significant negative effect on the total person hours spent on the site. Again this is in agreement with previous literature where tasks that involve little autonomy (i.e. you have to pick from a restricted choice of inputs) have a negative effect on user satisfaction (Dubinsky & Skinner, 1984).

Finally considering the task workflow design factors variety and freedom, it was found that increasing the variety of user judgements required in order to make a classification had a significant negative effect on the average time classifying per visit. This is in agreement with Lasecki et al.'s (2014) findings regarding the importance of continuity regarding task success. Platforms that allow the volunteer to complete the required tasks in an order of their own choosing have a significantly greater amount of total person hours spent on them, again adding weight to the importance of autonomy previously outlined by Dubinsky & Skinner (1984).

6.6 CONCLUSIONS

This study represents a first attempt at using task workflow design to capture and study the types of work carried out by volunteers of current VCS platforms. The results of the hierarchical task analysis have revealed a variety of different TWD configurations in terms of the factors outlined in the model (see figure 4.2), whilst the website analytics studied have shown differences in volunteer behaviour between projects.

Thesis RQ_1: What task workflow designs are currently used by citizen science projects?

The results from the HTA show a large variation between projects, especially in terms of the types of task and judgements required. Furthermore, the results imply that there seems little consideration of task

workflow design at the development stage of current VCS platforms, or at least the variety observed suggests a lack of 'best practice' procedures in place. The study has not found any relationship between the scientific discipline involved and the platforms' task workflow design, or any other compelling reason or pattern for their derivation. This suggests a more experimental approach has been taken along the lines of a '*if it works, don't change it*' mentality, i.e. if the tasks and judgements etc. involved are returning the scientific results expected at a 'good enough' quality, there is no need to risk change.

Thesis RQ_2: Does task workflow design influence volunteer behaviour?

In considering volunteer behaviour through the analysis of website analytics, the study shows that while all of the VCS platforms follow the 'long-tail' profile over time in terms of the total number of visits, there are important differences in behaviour revealed when considering other factors such as time spent classifying or percentage of return users. This raises questions regarding the type of user behaviour that is right for a project, and therefore will result in the best possible scientific outcomes.

For instance, would a site be more successful with a higher number of total visits, that might well analyse the dataset in a shorter timespan, or would it benefit more by generating a greater number of return visits from volunteers who perhaps understand the tasks required more thoroughly?

The results from the study suggest that TWD factors can have a significant effect upon volunteer behaviour, and also the effect can vary and be seen as either positive or negative depending on the outcomes sought. Related to the question raised earlier, this will greatly depend on the volunteer behaviour required to make the site a success.

A project may require the creation of an established volunteer community that regularly visits the site, perhaps as the science case is more involved and requires a longer time period for volunteers' to reach the required performance level, or the outreach benefits explained in section 2.3.3 achieved. In this case, the results of the study suggest a task workflow design that incorporates tasks and judgements with greater variety and complexity, in a structure with greater freedom will achieve the desired effect. However a potential downside is that such a structure could also

alienate the more casual volunteer, with the increased variety and complexity resulting in first-time visitors classifying for a shorter period. Rather than there being an ideal 'best practice' for task workflow design that fits all VCS platforms, it is more likely that a trade-off is going to have to be made which will depend on the goals of the project.

Nevertheless, the results found suggest that task workflow design does appear to influence volunteer behaviour and therefore should be considered by the developers and science teams of future VCS platforms. Additionally, the type of volunteer behaviour that best fits achieving the projects' goals needs defining to understand the appropriate design. This could include the 'live fast and die young' approach that encourages the casual user to make just one more click (Eveleigh et al., 2014) or the 'slow burner' that has a smaller, but dedicated community that continuously comes back for more.

6.7 SUMMARY AND OUTLOOK

In the broader context of this thesis, this is an exploratory study to first better understand the current landscape of VCS platforms in terms of task workflow design, and also investigate the application of lessons learnt from other related fields regarding the subject. It highlights the importance of task workflow design factors, while further analysis also showed their differing effect on various volunteer behaviours.

Methodologically however, while this experiment has potential in comparing the task workflows and volunteer behaviour in a broader context, a more platform specific analysis will be required to understand the complex relationships between task workflow design, volunteer behaviour and ultimately scientific output. This approach will not only allow a more direct appraisal of the effect of task workflow design on scientific results, but will also separate out other factors that effect volunteer behaviour, such as media attention and the initial motivation to take part.

Another limitation of the study is that while it has identified the influence of task workflow design on volunteer behaviour, it has not established whether this is a passive or conscious effect, i.e. is the volunteer aware of task workflow design factors, and their impact on their own behaviour? It

also does not consider the intrinsic nature of the task to the science case, or variables regarding the communities involved. The study in chapter 7 responds to this issue, by considering the opinions and thoughts of volunteers regarding the Zooniverse VCS platform Planet Four.

7

THE VOLUNTEER'S POINT OF VIEW

This chapter presents a study that investigates whether the volunteers of a VCS platform have any issues or concerns regarding the design and usability of the site, and if any can be related to the task workflow design factors derived in chapter 4. Findings from an online questionnaire completed by members of an existing citizen science community (Planet Four) show that while opinions regarding the general design and appearance of the site are positive, volunteers do have concerns regarding the tasks and tools used, the imagery presented to them and the discussion and feedback mechanisms provided. Additionally, a large proportion of volunteer concerns directly mention, or can be related to, task workflow design factors, both those that are science case influenced and those that can be influenced predominantly by the design of the platform. 'Free response' comments suggest that more consideration should be given to task workflow design, in order to address some of the issues raised.

7.1 INTRODUCTION

Clearly the success of any crowdsourcing project is related to its ability to attract and engage participants. Previous research has shown that citizen science projects are no different (see 2.5), relying on large participation numbers to get through the data analysis, whilst the level of participant engagement can have an effect on overall performance (Prather et al., 2013). The study in chapter 6 has shown that task workflow design and the factors associated with it can have an effect on volunteer behaviour and engagement, specifically in terms of their loyalty (how often they return to the platform), the time they spend performing the scientific tasks, the total number of visitors that visit the project and the total person hours they commit.

As reviewed in section 2.6.3, the body of research that has directly surveyed the volunteer has concentrated on investigating their motivations for visiting the citizen science platform (Raddick et al., 2010; 2013; Mankowski et al., 2011), rather than sourcing their views regarding the experience of using the platform itself. Other research directly concerned with volunteer behaviour (see 2.6.4) tends to be more quantitative in nature, involving website analytics and performance measures (Ponciano et al., 2014; Eveleigh et al., 2014) without using qualitative findings to add context and reasoning to any observed trends.

The study presented in this chapter attempts to fill this gap, by collating the views and opinions of volunteers regarding the design and usability of a specific VCS platform. In doing this, the study explores whether volunteer behaviour can be directly related to concerns regarding platform design. Specifically, it investigates if concerns regarding task workflow design factors exist in the volunteer community, as suggested in chapter 6, and if other platform design factors have an influence. This knowledge can be used to help understand if factors that can be influenced at the design stage of the platform are being implemented at their optimum, resulting in the benefits to the scientific outcomes of the project as outlined in section 4.5.3.1.

7.2 RESEARCH QUESTIONS

To better understand the volunteer's point of view regarding the general design and usability of a VCS platform, the community of a specific project (Planet Four) was surveyed through an online questionnaire, asking volunteers their opinions across several design topics. The principal research question of this study is:

RQ: What are the volunteer concerns (if any) regarding the design of the platform?

In addressing this question regarding the general design of a VCS platform, the study also answers the sub-question:

RQ_a: Do volunteers have any concerns regarding the task workflow design of the platform?

The study in chapter 6 has shown that task workflow design factors can have an effect on volunteer behaviour in terms of quantitative, web analytic measures. This study adds context to these findings by identifying the specific concerns of volunteers regarding task workflow design, and how they might affect their behaviour and engagement with the platform. In terms of the overall aims of this thesis, this study's contribution helps to address RQ_2 (see 4.5.4), by revealing if and how task workflow design influences volunteer behaviour when using a current VCS platform.

7.3 EXPERIMENT DESIGN

To determine the views and concerns of volunteers regarding an existing VCS project, an online questionnaire was developed and disseminated to registered users of the Zooniverse's Planet Four platform. Launched in January 2013 to study 'fans' – seasonal features that appear on the surface, Planet Four is the focus of the empirical work of this thesis for the reasons explained in section 5.3.

7.3.1 Questionnaire Design

An online questionnaire[16] was created using the survey tool Qualtrics[17] that contained four sections regarding the infrastructure of the site: *design & usability, tasks & tools, imagery* and *discussion & feedback*. Each of the sections consisted of a number of Likert-type statements as explained in section 5.2.2.1, in conjunction with a 'free-response' type box where participants could record any additional opinions or comments that were not adequately covered through the Likert responses. Prior to these four sections, participants completed a general demographics section regarding their background and experience.

[16] An example of the questionnaire layout and content can be found in appendix B of the thesis.

[17] The output for this chapter was generated using Qualtrics software, Copyright © 2015 Qualtrics. http://www.qualtrics.com.

7.3.1.1 Design & Usability

This section contained statements regarding the general design, usability and appearance of the site as a whole. This was intended to determine the opinions of volunteers regarding general website navigation, logic, ease of use and its overall aims separate from the more specific scientific task required of the site.

7.3.1.2 Tasks & Tools

This section contained statements more directly involving the specific tools used and tasks required to achieve the scientific classification or goal of the site. When using Planet Four, volunteers have two main tools to use in order to mark fans on the imagery. The 'fan' tool allows the user to mark a general fan shape, while the 'blotch' tool allows the user to mark a more generic rounded blob when the direction of the fan is hard to distinguish. This section intended to determine the users opinions regarding the task workflow design of the site (as defined in 4.5.1), and if the tools provided were suitable to achieve the scientific goal. NASA-TLX type statements were also included as described in section 5.2.2.2, in order to measure the volunteers' perceived workload when carrying out the tasks.

7.3.1.3 Imagery

In order to detect and mark fans on the surface of Mars, volunteers were shown colourised, HiRISE imagery of the planet surface (see figure 7.1). The imagery section of the questionnaire contained statements regarding both the quality and content of the images displayed, intended to determine the degree to which they assisted or hindered the completion of the scientific task.



Figure 7.1: Planet Four Interface

7.3.1.4 Discussion & Feedback

The final section of the questionnaire contained statements regarding the discussion tools and feedback provided by the platform. As with most Zooniverse projects, Planet Four includes a forum on which volunteers can share information, request assistance and communicate with the scientific experts involved. This section was designed to determine the volunteers' opinions and usage of the forum, along with more general views regarding feedback about their contribution and its end use.

7.3.2 Role

Carried out with full agreement of the Citizen Science Alliance board and Planet Four science team, the author had full responsibility and in agreement with supervision performed all the activities necessary to conduct this experiment. This ranged from experiment design, orchestration, set up and support. Participant recruitment was performed in collaboration with the Zooniverse and Planet Four teams, who advertised the study through their regular 'newsletter' emails and social media posts.

7.3.3 Participants

178 participants completed the online questionnaire between 2nd April and the 15th September 2013. 98 were male (55.1%), 76 female (42.7%) with 4 participants declining to respond. The participants were between 11 and 75 years old (mean = 46, median = 52) with the majority being resident in the UK (65.7%), while 12.4% were from the USA and a total of 22 countries were represented. The importance of media attention is demonstrated, with over half of respondents (53.9%) discovering the Planet Four site through the BBC's Stargazing Live broadcast. Likewise, over half of the participants have taken part in other Zooniverse citizen science projects as well as Planet Four, with the most popular sites involving a similar discipline (Planet Hunters – 36.0% and Galaxy Zoo – 31.5%).

7.4 ANALYSIS AND RESULTS

The results of this chapter's study are presented for each questionnaire section in turn, as described in 7.3.1, in terms of participant responses to the Likert-type statements in conjunction with the 'free-response' qualitative data to add context to the findings. Finally, the questionnaire results are further investigated in relation to the participants' age and familiarity with the Zooniverse platform.

7.4.1 Design & Usability Results

Responses from the design & usability section of the questionnaire suggest that participants have no strong issues or concerns, with the majority agreeing with the positive statements regarding ease of use, visual attractiveness, scientific aims and quickness of learning. Figure 7.2 shows the participants' level of agreement with the four statements, with 9 (right of graph) representing 'strongly agree' and 1 representing 'strongly disagree'.

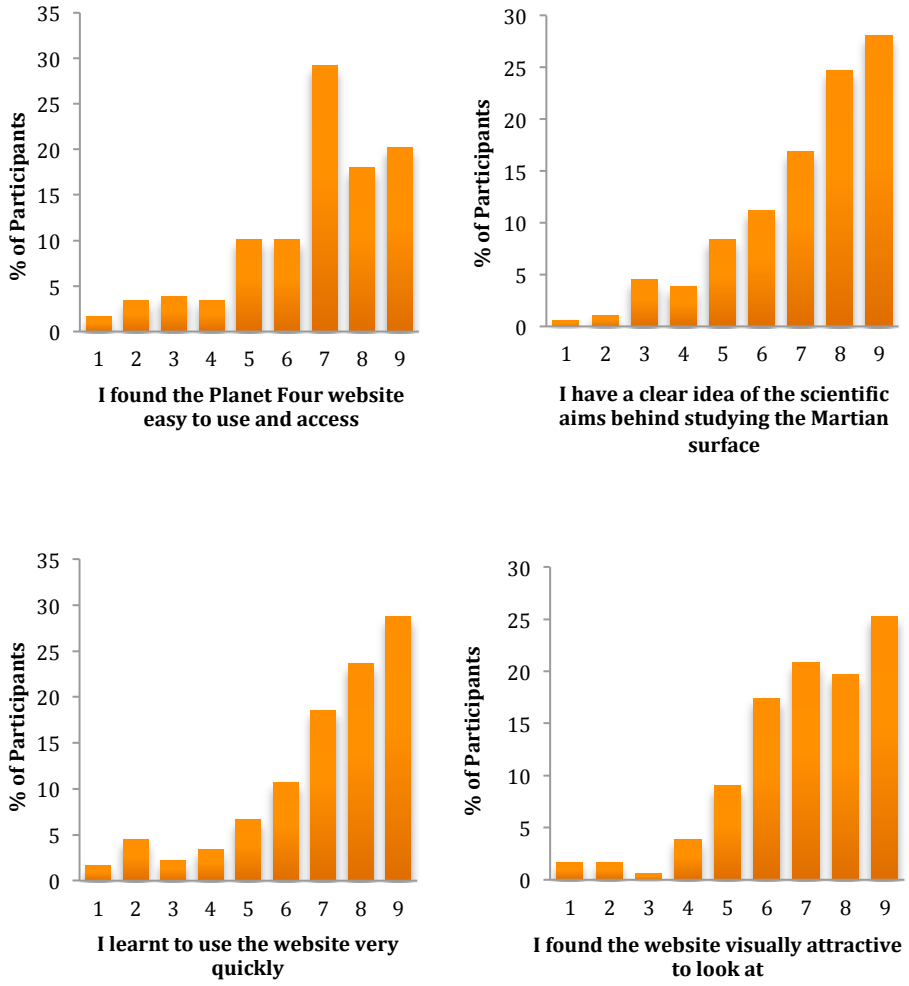


Figure 7.2: Participant responses to Design & Usability section

A positive skew is clearly visible, with over 65% of participants responding with a score of 7 or above for each of the statements. Additionally, 38 out of the 178 participants also provided 'free response' replies. These responses have been coded under the categories *ease of use, appearance, accessibility* and *technical issues*, as shown in table 7.1. The categories *ease of use* and *appearance* are self-explanatory and relate directly to the statements presented to the participants regarding design & usability. They are predominantly positive in nature.

Table 7.1: 'Free' responses to design & usability section

Response Category	Number of Responses
Ease of Use	9
Appearance	2
Accessibility	7
Technical Issues	17
Total:	35

The *accessibility* category represents statements from the participants regarding the method and technology they used to access the site and its suitability, as stated by participant P95:

> *"It would be great if the site was compatible with iPhone 5"*

Finally, the *technical issues* category represents statements from the participants regarding more general technical problems they encountered when using the site. These issues were mostly regarding the site lagging or crashing at certain times, for example (P88):

> *"I found that the site regularly crashed during prolonged sessions of mapping"*

While the responses in these two sections could be of use to the Zooniverse developers in order to inform technical aspects of the sites' infrastructure, they are considered beyond the scope of the more human factors orientated themes of this thesis. The final three responses (taking the total to 38) are general positive exclamations regarding the site, that don't address any particular issue or theme (P134):

> *"This is a really great program!"* and P167: *"good website!"*

7.4.2 Task & Tools Results

The participant responses to the task & tools section of the questionnaire are much less clear-cut compared to the previous section, showing less agreement overall across the sample. Figure 7.3 shows the participant responses in regards to both the level of mental and physical demand of the 'marking fans' task. In assessing *'how mentally demanding was the task of marking fans?'* 20% responded with a score of 3, where 1 represents 'very low' and 9 represents 'very high'. However, over 25% of participants gave a

score of 6 or over (5 being the mid-point), suggesting they found the task mentally demanding. Participant responses suggest that the task is slightly less physically demanding, with 25% of participants giving the lowest score of 1, however as with the mental demand statement 25% of participants gave a score of 6 or over.

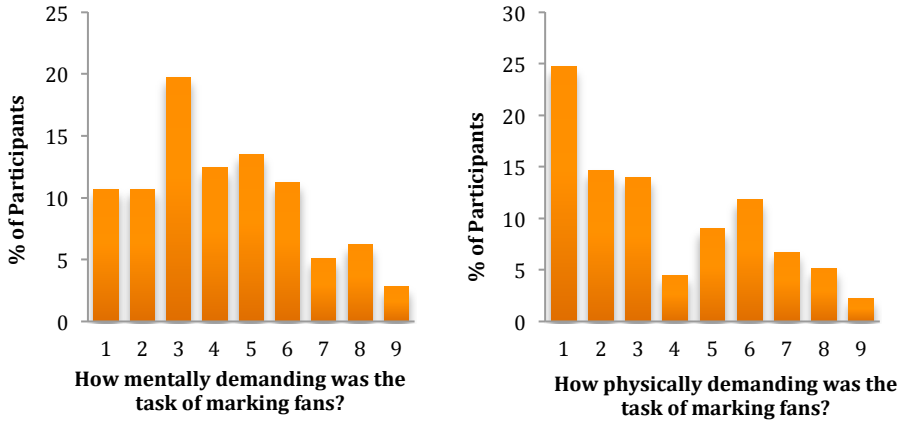


Figure 7.3: Participant responses to task mental and physical demand

Figure 7.4 shows participant responses to the next four statements of the section, regarding task pace, success effort and frustration. In assessing *'how hurried or rushed was the pace of marking fans?'* 21% of participants responded with the lowest score of 1 (very low), 3% gave the highest score of 9 (very high) while 19% of participants gave the central score of 5. No absolute time limit is present when performing the 'marking fan' task on Planet Four, suggesting that time pressures are being caused by more subtle factors.

In assessing *'how successful were you in accomplishing the task of marking fans?'* 28% of participants responded with a score of 3, 5% gave the lowest score of 1 (perfect) and 2% gave the highest score of 9 (failure). Over 17% of participants responded with a score of 6 or higher, suggesting doubts about their performance in completing the task. Considering the question *'how hard did you have to work to accomplish your level of performance?'* there is a large spread of responses, with the majority of participants giving the central score of 5 (19%). Over 5 times the number of participants (6%) gave the lowest score of 1 (very low) compared to the highest score of 9 (very high), but overall there is no clear indication that the task of marking fans is perceived as either easy or hard work.

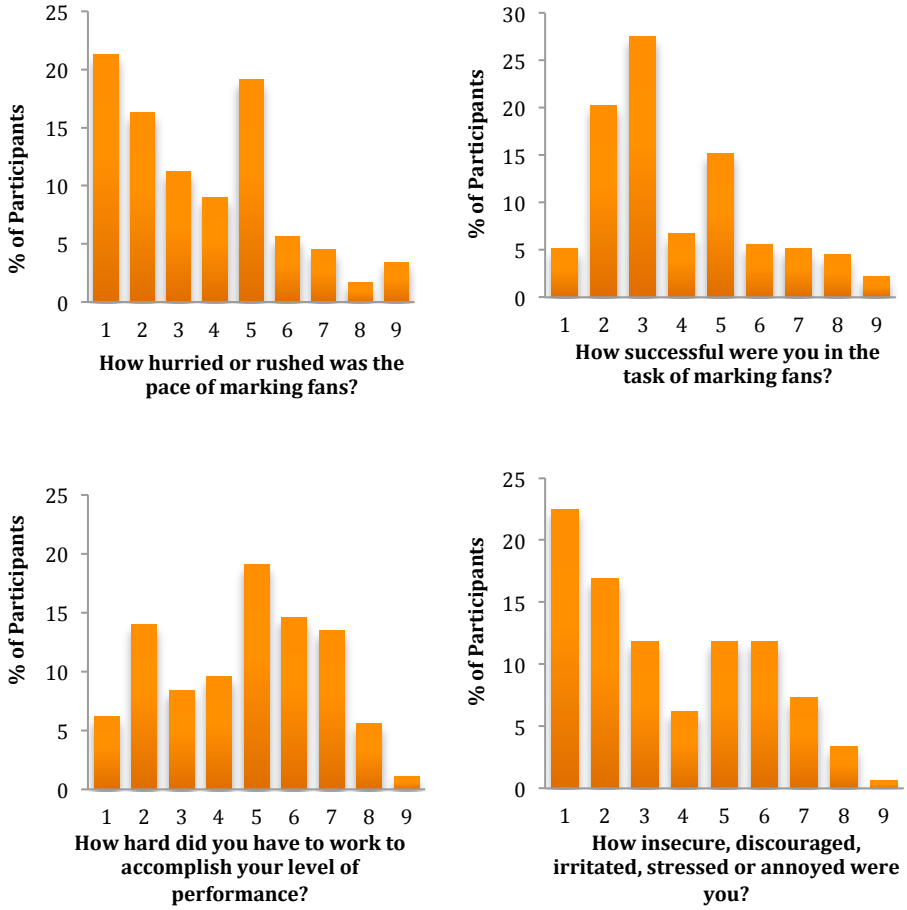


Figure 7.4: Participant responses to pace, success, effort & frustration

In assessing *'how insecure, discouraged, irritated, stressed or annoyed were you?'* 23% of participants responded with the lowest score of 1 (very low) and 1% gave the highest score of 9 (very high). Although suggesting a low level of stress and irritation among participants during the task, it should be noted that over 35% of participants gave a score of 5 or above.

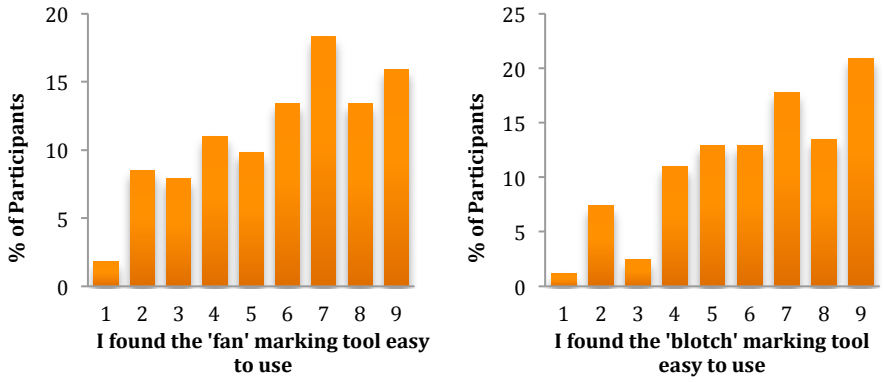


Figure 7.5: Participant responses to 'fan' and 'blotch' tool use

Figure 7.5 shows participant responses to statements regarding the ease of use of the two marking tools used on Planet Four, 'fan' and 'blotch'. In assessing their agreement with the statement *'I found the fan marking tool easy to use'* 18% of participants responded with a score of 7, 2% gave the lowest score of 1 (strongly disagree) and 16% gave the highest score of 9 (strongly agree). Although suggesting most participants found the tool easy to use, over 35% responded with a score of 5 or below, with a large spread apparent across each score. Finally, considering the statement *'I found the blotch marking tool easy to use'* 1% gave the lowest score of 1 (strongly disagree) and 21% gave the highest score of 9 (strongly agree). Although suggesting that participants found the blotch tool slightly easier to use than the fan tool, there is again a large spread present across the scores.

In terms of 'free response' replies for the task & tools section, 53 of the 178 participants gave a response, which have subsequently been coded into 6 categories shown in table 7.2: *extra tools, grouping & duplicate, accuracy, zoom & scale, tool mechanics* and *technical issues.*

Table 7.2: 'Free' responses to tasks & tools section

Response Category	Number of Responses
Extra Tools	12
Grouping & Duplicate	7
Accuracy	9
Zoom & Scale	9
Tool Mechanics	12
Technical Issues	4
Total:	53

As with table 7.1, the *technical issues* category refers to statements beyond the human factors scope of this thesis. The *extra tools* category refers to responses where participants have indicated their desire for a new or different tool in order to carry out the task, as stated by participant P63:

> *"A tool to mark fissures that have multiple eruptions. Some maps have lots of apparent dust but no obvious fans or blotches. Perhaps a stamp tool for these maps…"*

The *grouping & duplicate* category represents responses where the participant has indicated the need for a tool or system to replicate previous markings, when needed (P51):

> *"... A repeat marking tool, especially fan tool when image has 'hundreds' of fans"*

The *accuracy* category refers to issues participants have reported regarding their perceived accuracy when using the tools (P102):

> *"Make them easier to control so I can mark more accurately."*

The *zoom & scale* category represents responses where the participant has indicated the need for a specific type of tool that allows for 'zooming' in and out of the image in order to mark less clear features. It might also include a scale indicator for reference, for instance (P36):

> *"A zoom feature & an easy indicator of scale would be helpful..."*

Finally, the *tool mechanics* category refers to instances where the participant has suggested alterations/improvements to the existing tools in terms of their appearance and design (P5):

> *"[The] ability to change colour of tools to assist with different coloured backgrounds..."*

7.4.3 Imagery Results

The participant responses to the imagery section of the questionnaire reveal a mixed-picture across the four statements included. Figure 7.6 shows the responses to statements regarding feature visibility, image quality and complexity. In assessing their agreement with the statement *'I found it hard to 'pick out' fans and features'* 25% of participants responded with the score of 2, 12% gave the lowest score of 1 (strongly disagree) and 1% gave the highest score of 9 (strongly agree). Over 65% of participants recorded a score below the central mark of 5, suggesting they could pick out the fan feature without too much difficulty. In considering the statement *'I thought the image quality was consistent'* 19% of participants responded with a score of 7, 3% gave the lowest score of 1 (strongly disagree) and 18% gave the highest score of 9 (strongly agree).

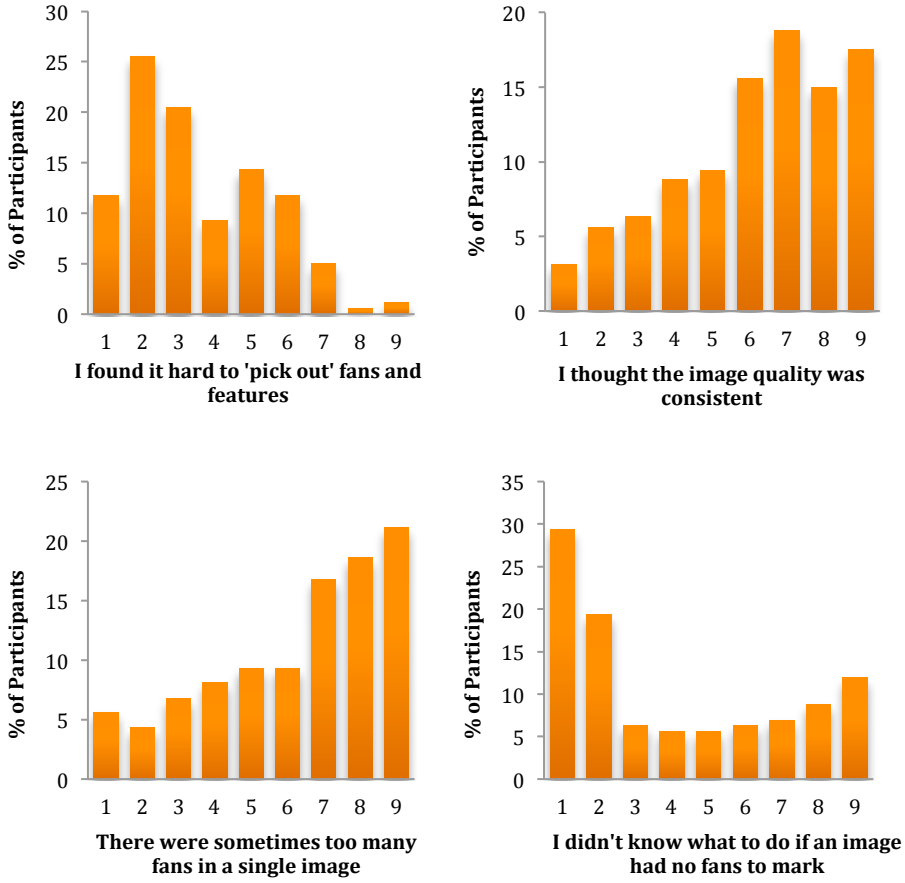


Figure 7.6: Participant responses to imagery statements

In assessing their agreement with the statement *'there were sometimes too many fans in a single image'* 6% of participants responded with the lowest score of 1 (strongly disagree) and 21% gave the highest score of 9 (strongly agree). Over 55% of participants recorded a score of 7 or above, suggesting that the number of fans in an image can sometimes be an issue. Finally, in considering the statement *'I didn't know what to do if an image had no fans to mark'* 29% of participants responded with the lowest score of 1 (strongly disagree) and 12% gave the highest score of 9 (strongly agree). This seems to suggest evidence of a split in opinion, with only 6% responding with the central score of 5.

Considering the 'free response' replies of participants for the imagery section, 27 of the 178 participants gave a response. Their answers have been coded into five categories: *feature identification, image quality, feature abundance, response ambiguity* and *information* as shown in table 7.3.

Table 7.3: 'Free' responses to imagery section

Response Category	Number of Responses
Feature Identification	4
Image Quality	6
Feature Abundance	5
Response Ambiguity	6
Information	6
Total:	27

The *feature identification* category refers to responses participants have made regarding their difficulties in identifying features in the imagery, for example (P110):

> *"Sometimes it was difficult to identify if something was a fan or not."*

Statements included in the *image quality* category specifically refer to variations in image appearance and clarity (P106):

> *"The images usually appear well reworked and focused, easy to watch."*

The *feature abundance* category refers to statements participants have made regarding the number of features present in one image, and the issues that this can cause (P140):

> *"Some of the images have so many features on them that I 'run out of steam' when highlighting them…"*

The *response ambiguity* category refers to participant responses regarding concerns of what to do when no features are present in the image, and usually indicate a lack of understanding in how to deal with the situation (P143):

> *"I've said it already - it needs to be far clearer what to do if no fans or blotches are visible."*

Finally, the *information* category refers to statements participants have made indicating that they would prefer more information about the imagery to be displayed (P58):

> *"A little more information regarding height of the satellite that is photographing the surface of the planet would help to understand the*

scale of what we see on the planets surface. (Am I looking at pebble or a mountain?)"

7.4.4 Discussion & Feedback Results

Responses to the discussion & feedback section of the questionnaire reveal that the majority of participants (87%) have 'signed up' to the Planet Four platform (i.e. provided a username and password). In terms of the Likert-type statement responses figure 7.7 shows participant scores regarding their Planet Four profile, discussion pages and Planet Four associate blog.

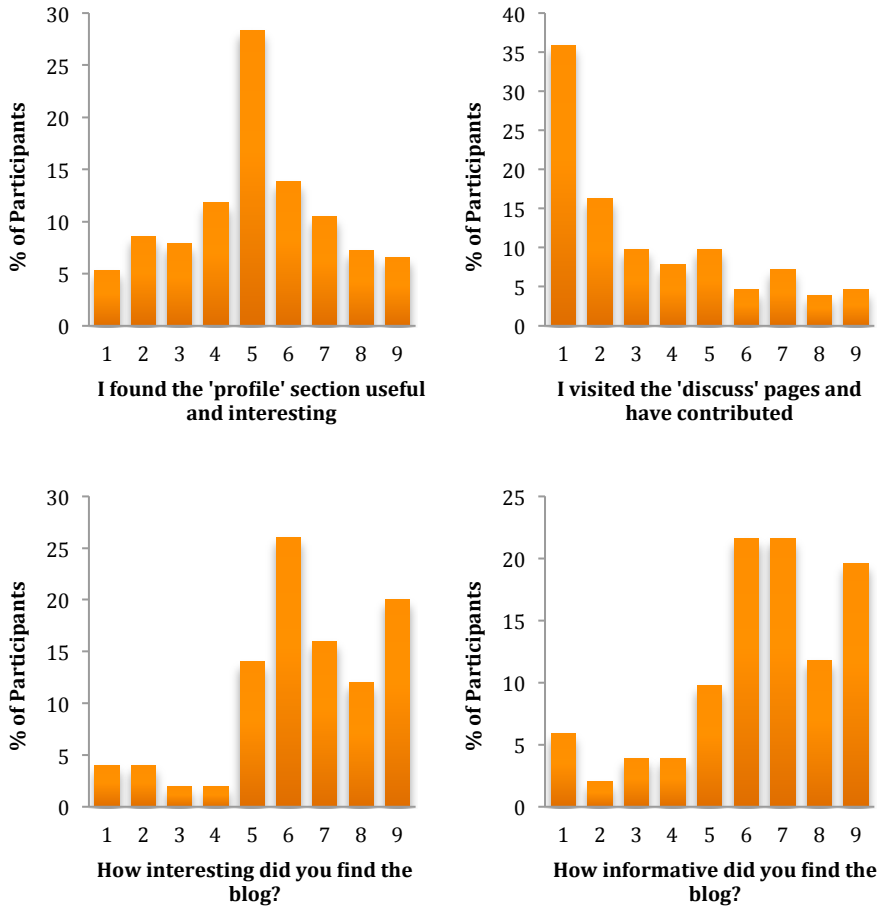


Figure 7.7: Participant responses regarding profile, discussion & blog

In assessing their agreement with the statement *'I found the 'profile' section useful and interesting'* 28% of participants responded with the central score of 5, 5% gave the lowest score of 1 (strongly disagree) and 7% gave the highest score of 9 (strongly agree). There is no strong opinion either way regarding the usefulness/interestingness of the profile section. In considering their agreement with the statement *'I visited the 'discuss' pages*

and have contributed' 36% gave the lowest score of 1 (strongly disagree) and 5% gave the highest score of 9 (strongly agree). Over twice as many participants responded with the lowest score than any other, suggesting low engagement with the discussion forum of Planet Four.

The majority of participants have never visited the Planet Four blog (68%). In terms of those who have, when considering the question *'how interesting did you find the blog?'* 26% responded with the score of 6, 4% gave the score of 1 (very boring) and 20% gave the highest score of 9 (very interesting). There is a noticeable spread of differing degrees of interest (scores between 5 and 9), with less than 12% of participants giving a score lower than 5. In assessing the question *'how informative did you find the blog?'* 22% of participants responded with the highest score of 9 (very informative). Overall, there is a similar spread present as was found with the blog interest statement towards the higher scores.

Figure 7.8 shows participant response to the final two statements of the discussion & feedback section, and the questionnaire itself. They concern the sites' contribution towards the participants' understanding of Mars, and in the reverse the participants' feelings of contribution towards the science case.

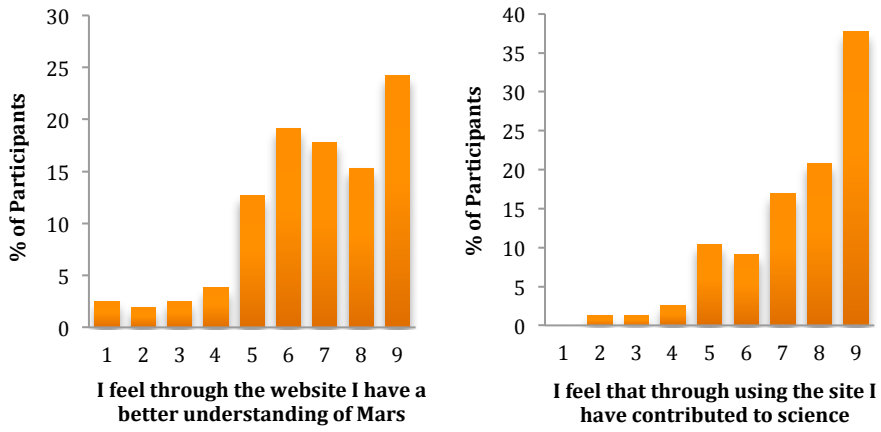


Figure 7.8: Participant responses regarding understanding & contribution

When considering their agreement with the statement *'I feel through the website I have a better understand of Mars'* 3% of participants responded with the lowest score of 1 (strongly disagree) and 24% gave the highest score of 9 (strongly agree). Again, there is a large spread of differing

agreement between the scores of 5 and 9, with less than 12% of participants giving a score of below 5. Finally, in assessing their agreement with the statement *'I feel that through using the site I have contributed to science'* no participants responded with the lowest score of 1 (strongly disagree) and 38% gave the highest score of 9 (strongly agree). The participants clearly feel they have contributed to the discipline, with over 58% responding with a score of 8 or 9.

In terms of the 'free response' replies given by participants for the discussion & feedback section, 23 of the 178 participants provided a response. These answers have been subsequently coded under 5 categories: *awareness, scientist feedback, contribution, forum design* and *sign in issue,* as shown in table 7.4*.*

Table 7.4: 'Free' responses to discussion & feedback section

Response Category	Number of Responses
Awareness	5
Scientist Feedback	9
Contribution	4
Forum Design	3
Sign In Issues	2
Total:	23

The *awareness* category refers to comments participants have made indicating they were unaware of the existence or how to use the discussion tools provided i.e. (P125):

> *"What's the profile? Never used it…"*

Responses included in the *scientist feedback* category specifically refer to participants' requesting a greater amount of communication and feedback from the experts involved in the project (P9):

> *"It would be good to have more regular blogs, if possible, and maybe a bit more feedback on progress of the project…"*

The *contribution* category concerns statements participants have made requesting extra information about the contribution they are making to science when taking part (P112):

> *"I think this is important? Why bother? Throw me a bone here people…"*

The *forum design* category refers to comments participants have made indicating problems they have encountered when using the discussion forums and community tools (P63):

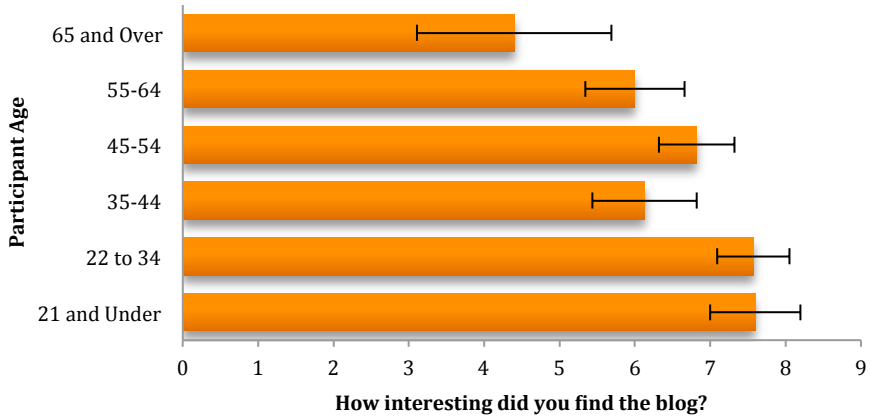
> *"I reiterate that I find it difficult to discuss the [image] I am on as I am taken to a what appears to be a general discussion board."*

Finally, the *sign in issues* category refers to responses participants have given regarding issues they have had signing in to the site – a prerequisite of using the discussion tools (P59):

> *"I find the Zooniverse login policies inconsistent… very confusing and frustrating. Sometimes I have to log in 3 or 4 times before I can start classifying…"*

7.4.5 Questionnaire Responses by Age Group

The following section addresses the effect the participants' age has on their responses to the Likert-type statements within the questionnaire. When considering responses to the *design & usability, tasks & tools* and *imagery* sections, participant age does not have a significant effect, with scores following the overall patterns previously discussed in this chapter. However, when considering the *discussion & feedback* section responses across participant age groups an effect is evident, as shown in figure 7.9.



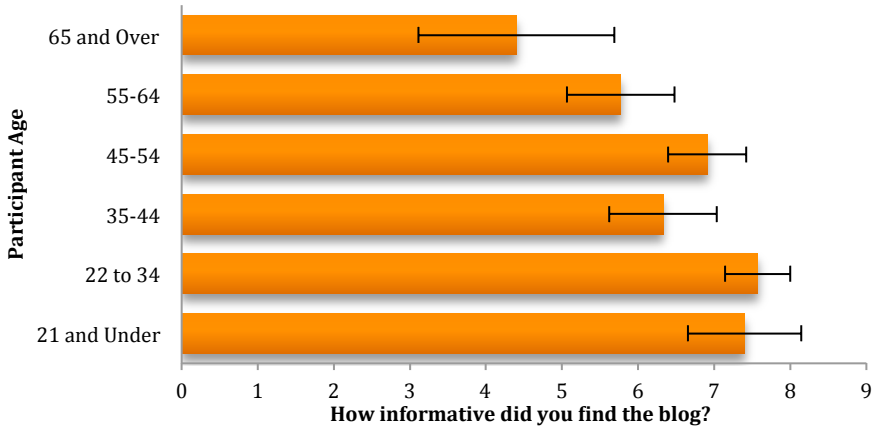


Figure 7.9: Average response to blog questions per age group

Specifically when responding to questions regarding the Planet Four blog, a relationship can be seen between participant age group and how interesting and informative they found it. Spearman's rank-order correlation tests confirm a strong, negative correlation between participant age and both how interesting and informative they found the blog, which are statistically significant ($r_s(8) = -0.387$, $p = 0.005$ and $r_s(8) = -0.380$, $p = 0.006$ respectively).

7.4.6 Questionnaire Responses by Zooniverse Experience

Finally, the effect of if the participant is new to the Zooniverse-type VCS platform or has been a Zooniverse volunteer previously is analysed. When considering responses to the *design & usability* and *imagery* sections, whether the participant is a new or existing Zooniverse volunteer had no significant effect on their responses, with scores agreeing with the overall patterns found before. However there are differences evident when considering the ease-of-use of the tools and contributions to the discussion pages, as shown in figure 7.10.

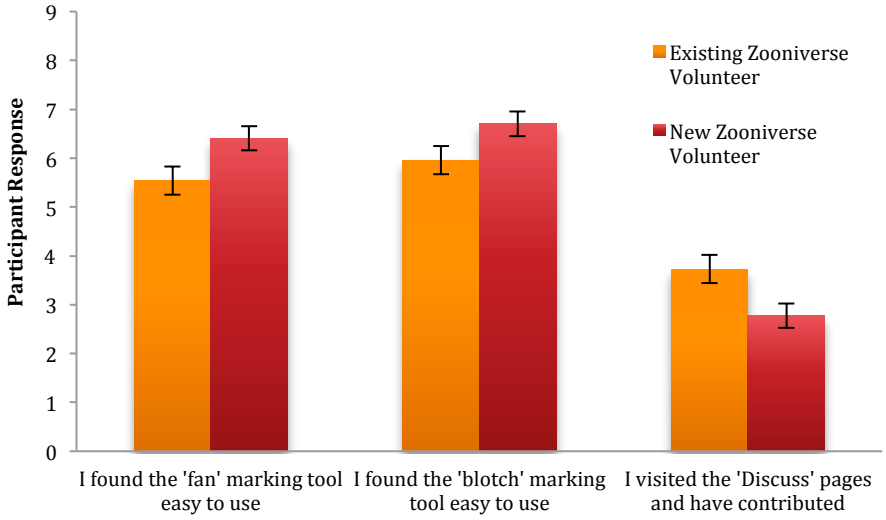


Figure 7.10: Participant responses to tool 'ease of use' and discussion page statements by Zooniverse experience

Participants that are new to the Zooniverse platform indicated a greater agreement with the statements that both the 'fan' and 'blotch' marking tools are easy to use (average score of 6.41 ± 2.26 and 6.70 ± 2.01 respectively) compared to participants who have taken part in other Zooniverse projects previously (5.54 ± 2.25 and 5.96 ± 2.30 respectively). A Mann-Whitney test showed that these differences in perceived tool 'ease of use' are significant, with $U = 2583$, $p = 0.012$ and $U = 2674.5$, $p = 0.037$ respectively. The responses to the statement concerning contribution to the 'discuss' pages indicated that existing Zooniverse volunteers are more in agreement (average score of 3.73 ± 2.69) compared to participants new to the platform (2.77 ± 2.15). It should be noted however that both groups' average score was considerably nearer the 'strongly disagree' end of the scale. Again a Mann-Whitney test showed that this difference is significant, with $U = 2354$, $p = 0.036$.

7.5 DISCUSSION

In this section the findings of the study are discussed by contextualising the quantitative Likert results with more in-depth qualitative descriptions from the 'free response' replies. Participants' quotes are used to highlight their point of view and the issues raised are related back to broader concerns regarding task workflow design and other VCS platform factors.

7.5.1 Volunteer Concerns Regarding Design & Usability

The overriding positive skew of participant responses to statements regarding the general design and usability of the site suggested that volunteers found Planet Four easy to use and navigate, quick to learn, visually attractive and they understood the scientific aims of the project, i.e. what the site is for. Furthermore, this pattern of response is independent of the participants' age or whether they are new or existing volunteers to the Zooniverse platform.

The 'free response' replies of participants supported this view. While the majority of responses given concerned more technical issues (that perhaps are unavoidable considering the broad range of methods and technology used to access the site), those that were about design & usability were overwhelmingly positive (P150):

> *"Planet 4 website is easy to use, and I have a clear understanding of what we're looking for…"*

Other participant responses were equally praising when discussing the appearance and look of the site (P42):

> *"Looks fine to me!"* and (P18): *"Looks great TBH…"*

Overall, both the quantitative Likert responses and the qualitative 'free response' comments suggested that participants had no strongly held issues or concerns regarding the general design and usability of the Planet Four site.

7.5.2 Volunteer Concerns Regarding Tasks & Tools

Unlike the previous section, the participant responses to statements regarding the tasks and tools of the site were inconclusive, with a wide spread of responses given. Although at first glance the data suggested participants overall did not find the task of marking fans demanding, difficult or stressful, over 25% of responses given were nearer the highest score of 9 (very high) than the lowest of 1 (very low) when considering mental and physical demand, how hard they had to work to achieve the task and how stressed etc. it made them. Additionally, over 17% of responses given were nearer the score of 9 (failure) than 1 (success) when

indicating how successful they felt they were at achieving the task. In terms of the ease of use of the two tools, while new participants to the Zooniverse system indicated that they found them slightly easier to use than existing volunteers, again over 25% of responses were nearer the lowest score of 1 (strongly disagree) than the highest score of 9. Overall the responses indicated that volunteers did have some issues and concerns regarding the tasks and tools of the Planet Four site.

The 'free response' replies of participants supported this conclusion, with participants highlighting a number of issues and changes that can be related to the task workflow design factors explained in chapter 4 of this thesis. A number of participants indicated the need for extra tools to use (P149):

> *"Extra tools to mark ice features etc."*

This suggests they wanted to perform tasks beyond those prescribed to them, thus demonstrating the importance of autonomy in ensuring user satisfaction (Chung-Yan, 2010). Other participants mentioned a specific tool that would be useful for making duplicate markings (P40):

> *"Some images have literally dozens of fans on them, and this quite monotonous marking all of them. If they are all in the same region is there any point in marking them all?"*

This demonstrates volunteer concerns regarding what is perceived to be repetitive, monotonous work, showing a preference for more variety in the tasks they undertake (Dubinsky & Skinner, 1984). Other participant comments concerned accuracy (P32):

> *"The tools were a bit clunky to use, especially when the features were very small. It was hard to get a precise outline of the tiny images…"*

Zoom & scale tools (P25):

> *"Some of the fans and blotches are too small to mark properly with the tools available, if it was possible to zoom in on the image without increasing the size of the tools, that would be great."*

And tool mechanics (P162):

> *"Being able to mark fans by clicking once at the source where the fan originates, then clicking once where the streak begins to fade away instead of having to drag the mouse across the screen."*

Such comments can all be related to concerns regarding the type of task they are required to complete, and the associated judgement they have to make to do so (Denis G. Pelli & Farell, 2010), with the range of differing opinions given suggesting it is a complex issue.

7.5.3 Volunteer Concerns Regarding Imagery

When considering image quality, the overall positive skew of scores suggested that participants had no issues or concerns regarding the quality and clarity of the images presented to them. Although the results are perhaps less 'clear-cut', again the positive skew of scores suggested that most participants also had no problem in identifying 'fan' features within the imagery. In terms of the number of fans that are found in the imagery, the clear negative skew of scores suggested that volunteers had concerns regarding there being 'too many' fans present. Responses regarding 'null' imagery (those with no fans present) suggested a split, where although the majority of participants indicated they knew how to deal with the situation 12% of participants gave a score of 9 (strongly agree) indicating they did not know what to do.

The 'free response' replies of participants raised a number of concerns and issues regarding the imagery, a number of which can again be attributed to task workflow design factors. Comments regarding the image quality were predominantly positive in nature, explaining how beautiful and clear participants found their display (in agreement with the Likert-type scores). Other participants indicated issues when identifying fans in the imagery (P168):

> *"Because I'm unsure of some features I often wonder if my input is correct or not…"*

It appears that volunteers again had issues with the type of task they were asked to perform, and the judgement they had to make in deciding what features to mark. A number of participants raised concerns regarding the abundance of fans in the imagery (P63):

> "*On some maps there are lots of fans and/or blotches. Perhaps a special fan (say a red coloured one) could be drawn on one to denote fans in excess of 20.*"

This can be related to the comments in the task and tools section regarding the need for a duplication tool. This again alludes to participant concerns regarding the repetitive nature of the work, and the perceived importance of task variety. Finally, several participants requested more information, some in general terms regarding the imagery (P64):

> *"It might be nice to have a couple of zoom boxes to show where each is on the planet so that you feel you're covering known space."*

Others commented more specifically regarding what to do with null imagery (P143):

> *"I've said it already - it needs to be far clearer what to do if no fans or blotches are visible."*

The former request identifies the importance of intrinsic feedback (as explained in 3.6.1) and its effect on participant satisfaction.

7.5.4 Volunteer Concerns Regarding Discussion & Feedback

Participant Likert scores regarding contributions to the discussion forum of the Planet Four site suggested an agreement with previous literature, where a committed few are responsible for the majority of contributions (Luczak-Roesch et al., 2014; see section 2.6.4). This pattern also holds true when considering the Planet Four blog, where almost 70% of participants have never visited it. This perhaps represents a missed opportunity, as the positive skew of scores regarding the blog suggested that those who have visited it mostly found it both interesting and informative. It is worth noting however that this positive outlook is inversely related to the age of the participant, which in itself could be an issue if previous research regarding the average age of citizen scientists holds true (Raddick et al., 2013). The general positive skew of participant responses regarding the scientific aims of the project and contribution suggested that participants clearly knew the reasons behind the Planet Four site, and understood the contribution they were making in taking part.

The 'free response' replies of participants for the discussion and feedback section of the questionnaire provided explanations for the apparently low take up of the various communication tools available, and also raised a number of other task workflow design-related concerns. Several participants drew attention to what could be described as generational differences when considering methods of communication (P168):

> *"Being older and not so great with computers I don't know what a blog is or how to use it. (This might seem stupid to the younger users)…"*

This Agrees with the findings of 7.4.6, and also alludes to the fact that maybe 'social-media' centric methods of communication are not the most apt considering the older demographic of VCS volunteers. Other participants indicated their lack of awareness that such tools existed (P125):

> *"What's the profile or blog? I have never heard of them…"*

This again seems to suggest that not enough information is present to inform volunteers of this functionality, a missed opportunity as those who are aware generally find the tools informative and interesting. A group of participant responses made reference to the content of discussion and blog communications, and several requested more feedback regarding their contribution (P109):

> *"I think that there should be more information about mars and more explanations about how we help…"*

And (P74):

> *"I love getting results back from the science team and knowing how I contributed. Keep pointing out our usefulness!"*

These comments can clearly be related to the task workflow design factors of feedback and task significance, and provide evidence to support the assertions of Hackman & Oldham (1975) regarding the importance of volunteers knowing the work is being undertaken for a reason, and receiving clear, actionable information (see section 4.2.1).

7.6 CONCLUSIONS

Using an online questionnaire to investigate the opinions and points of view of an existing VCS platform volunteer community, it was found that participants had a number of issues and concerns about the design and usability of the Zooniverse's Planet Four site. Although some of these concerns can be attributed to more general technical issues regarding access and compatibility, others are directly related to the task workflow design factors described in this thesis' conceptual model (see 4.3). Addressing the research question '*what are the volunteer concerns regarding the design of the platform?*' – A number of points were raised.

With regards to the general design of the site in terms of navigation and appearance, participant comments supported their corresponding Likert scores and suggested an overall positive outlook. This position is further supported by more general participant comments that express their appreciation of the site. In terms of relating volunteer concerns to task workflow design (RQ_a of this study), the task & tools section of the questionnaire produced a much larger range of Likert responses suggesting a range of opinions. Participant comments showed that issues exist regarding the mechanics of the tools used and the preference for extra tools, that can all be related to factors regarding autonomy, task variety, task type and judgements required. Likewise, responses to the imagery section also raised a number of participant concerns. However, rather than being related to the quality or clarity of the images themselves they referred to the content of the imagery and pointed to issues regarding the repetitiveness of the task (underlining the importance of variety) and difficulties in feature identification (related to the task type and judgement). Finally, the discussion and feedback responses questioned the suitability of blogs and forums as communication tools considering the demography of Planet Four users. It also raised concerns regarding the amount of feedback received and information regarding contribution, both of which directly relate to task workflow design factors.

Overall, the results of this study have revealed a number of concerns and issues that exist from the volunteer's point of view, many of which directly involve task workflow design factors. In relation to the overall thesis research question RQ_2: *Does task workflow design influence volunteer*

behaviour? – The self-reported opinions of current participants suggest that TWD can influence their motivation to take part, and therefore their behaviour.

Although a number of these factors are arguably science case influenced (feedback and significance for instance, see 4.4.1) others, such as task type, volunteer judgement, variety and autonomy can be influenced at the design stage of a VCS platform without too many caveats stifling the process. Participants were positive about the general design and appearance of the site, but it could be argued that more consideration should be given to task workflow design factors when VCS platforms are developed, especially to address some of the issues and concerns that exist from the volunteer's point of view.

7.7 SUMMARY AND OUTLOOK

The findings, and in particular the revealing of several volunteer issues and concerns, furthers the understanding of the effect of task workflow design factors in this thesis. The previous study (chapter 6) showed the effect that design-influenced task workflow factors can have on volunteer behaviour in more quantitative terms regarding website analytical measures. This study has added context to these findings, providing conscious volunteer issues and concerns that directly relate both to design-influenced and science-influenced task workflow design factors.

However, while the previous chapter and this study explore volunteer behaviour and opinion regarding task workflow design factors, what are the potential effects of these factors on scientific output? Previous literature has shown the relationship between volunteer experience and data accuracy (Prather et al., 2013) and the effect of task design on volunteer behaviour (Eveleigh et al., 2014), but none as yet has directly considered the relationship between task workflow design and scientific output. Methodologically, while the employed online questionnaire allowed participants to provide opinions and concerns regarding perceived accuracy and success at performing the task, such evidence is clearly anecdotal in nature. The following laboratory study in chapter 8 addresses this shortcoming by directly manipulating task workflow design factors associated with a purposely designed VCS platform and scientific case,

focussing on the effect such variations have on volunteer behaviour, opinion and scientific output through both quantitative and qualitative measures.

8

MANIPULATING TASK WORKFLOW DESIGN

This chapter presents a study that investigates the effects of manipulating task workflow design upon volunteer motivation, engagement and performance. Participants' self-reported feedback through questionnaire responses indicated a preference for interfaces involving greater autonomy and variety, with free-text responses suggesting that autonomy was the more important. This did not translate to improved performance however, with the Full interface (most autonomy) not resulting in significantly better performance in terms of coverage, agreement or efficiency compared to other less autonomous interfaces. Participants also reported frustration raising issues regarding the significantly less efficient Stepped interface, in terms of having to revisit previously marked craters several times to perform additional tasks. In terms of performance, the interface involving the least complex task, less variety and less autonomy (Ramped Position) resulted in the greatest data coverage. The interfaces with less autonomy and variety resulted in significantly better productivity (Ramped Position and Ramped Mark) in terms of the amount of time spent classifying each image. Performance in terms of agreement, both between participants and with expert judgement, was significantly improved when using the interfaces that included tasks that directly measured the required metric (Stepped and Ramped Position for crater position, Full and Ramped Mark for crater diameter).

8.1 INTRODUCTION

Due to the growing size and participation in citizen science and virtual citizen science projects, the discipline itself is becoming established as a distinct field of enquiry (Jordan et al., 2015). Building on this fact, a growing library of research exists that has studied facets of existing 'real

world' citizen science projects with established volunteer communities beyond the original scientific goals they were designed to achieve. To date, this research has predominantly concerned the motivations of volunteers, their contribution and how both evolve over time (as outlined in sections 2.6.3 and 2.6.4), rather than considered design and interface factors and their effect on volunteer performance and experience.

The studies discussed in the previous chapters 6 and 7 have shown that task workflow design factors such as autonomy, variety, task type and judgement can have an effect on volunteer behaviour, and that volunteers have a conscious opinion regarding issues that relate to these factors. Although some recent research has considered the effect of task and judgement on volunteer performance (Hutt et al., 2013), and considered how they should be designed dependant on volunteer commitment (Eveleigh et al., 2014 - see the Pareto distribution of volunteer behaviour in section 2.6.4), to our knowledge none has directly experimented with the manipulation of task workflow factors in order to investigate their effect on volunteer behaviour, experience and scientific output.

The experiment presented in this chapter attempts to fill this lacuna, by investigating the effect that different interface designs of a purpose-built version of the Zooniverse's Planet Four platform have on participant behaviour, experience and performance. In doing this, the experiment explores whether task workflow design factors can directly influence such measures of scientific output, and therefore if some interface designs are better suited to address the scientific question than others. Specifically, it investigates not only if the relationship between task workflow design and volunteer behaviour and experience exists as found in the studies of chapters 6 and 7 when directly manipulated concerning a specific platform, but extends this to consider the effect of task workflow design on volunteer performance and data output.

8.2 ESTIMATING THE AGE OF THE MARTIAN SURFACE

As briefly explained in section 5.3.1, a new science case has been derived on which to investigate the manipulation of task workflow design factors, to nullify any restrictions that might exist when working with an established platform. Developed in collaboration with the University of

Bristol's School of Earth Sciences as a sister platform to the Zooniverse's existing Planet Four, the primary scientific aim was to estimate the age of the Cerberus Fossae[18] region of the Martian surface. The reason for this choice of science case and target is that the region is in the vicinity of the landing zone of NASA's InSight[19] mission (scheduled launch in 2018), the first mission to the planet that will measure the seismic activity deep below the surface. In order to maximise the scientific success of the mission, regions of greater seismic activity need to be identified and the level of such activity estimated so that instrumentation can be calibrated accordingly. By knowing the age of the surface (younger surface equates to greater seismic activity), estimates of seismic activity can be made which in turn can inform NASA's InSight and future missions.

An established method for ageing the surface of a planet is through crater density analysis. In basic terms, this involves the counting and size measurement of craters in a certain area with the theory being that a new surface has no impact craters, and as time passes craters occur at a rate dependant on the size of the impact. The impact rate has previously been calibrated through the dating of samples taken from the lunar surface (McGill, 1977), and although similar samples as of yet are not available from Mars following work has scaled this measurement accordingly for Martian impactor rates (Hartmann, 1999; 2005). Figure 8.1 shows a graphical illustration of how crater density analysis relates to the age of the surface.

[18] Information on the Cerberus Fossae available at:
http://blog.planetfour.org/2015/05/19/the-cerberus-fossae/

[19] Information on the NASA InSight mission available at:
http://insight.jpl.nasa.gov/home.cfm

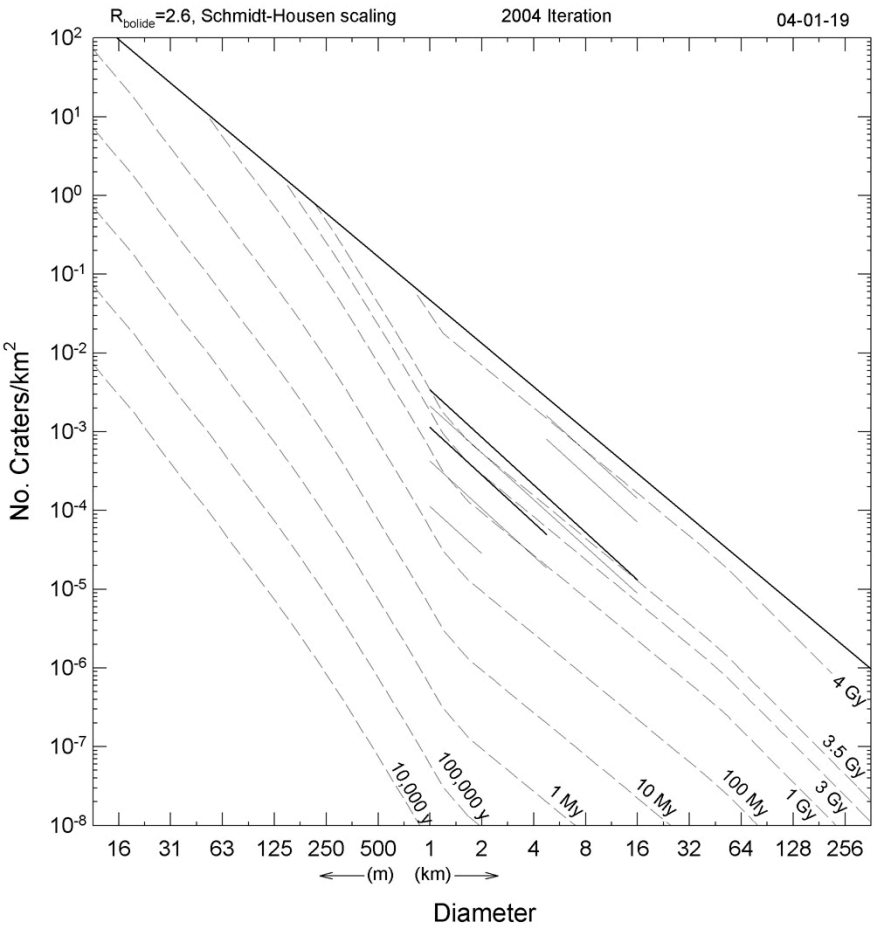


Figure 8.1: Isochrons for Martian Crater Populations[20]

The process of crater marking and size measurement required for crater density analysis is a time-consuming and arduous task, with sometimes several hundred craters needing to be identified in a single mission image, making it an ideal task for a citizen science platform and community. Previous studies have utilised the crater density analysis technique to estimate the age of areas surrounding the Cerberus Fossae region, giving bounds of between 5-20Ma (Murray et al., 2005; Berman & Hartmann, 2002), but were limited to considering craters >1km in diameter. The scientific aim of the VCS platform designed for this thesis was to validate and refine this estimated age range, by utilising citizen science volunteers to mark and measure craters over a greater size range specifically within the Cerberus Fossae region.

[20] Hartmann W. K. 2004, available at:
https://www.psi.edu/epo/isochrons/chron04b.html

Beyond the primary scientific impact the science case will potentially have, a second advantage of utilising a crater-counting citizen science project is that it is not the first occasion such a task has been presented to the public. In recent times several VCS platforms have involved volunteers marking craters on remotely sensed planetary body imagery, including the Zooniverse's Moon Zoo site (Joy et al., 2011). Existing research regarding these projects follows the trend of considering volunteer motivation (Brown, Gay, & Daus, 2011), dealing with inherent erroneous results (Antonenko et al., 2013) or considers the viability of resulting data through expert-volunteer comparison (Robbins et al., 2014). The platform used in the following experiments of this thesis builds on this research by extending it to consider not only these issues but also the design of the task itself and how it is presented to the volunteer.

8.3 IMAGE PROCESSING AND PREPARATION

With a science question and target area identified, the next step is to source and prepare the imagery to be displayed to volunteers for them to analyse. Over the past 50 years NASA has performed at least 16 missions to Mars (Sprinks et al., 2015 - see section 2.4) each of which have returned imagery data of varying quality and detail. While the obvious route might be to use the best resolution imagery available (currently HiRISE at ~0.5m/pixel) issues regarding the extent and coverage of the imagery make this decision more complex. For instance, while HiRISE imagery might result in the most detail, the greater image extent and surface coverage of an instrument like MRO's context (CTX) camera (~1% vs. ~78%, Sidiropoulos & Muller, 2014) might make it a more suitable choice for certain feature types and regions of interest. For the purposes of crater counting previous research (Daubar et al., 2013; Fassett et al., 2014) has tended to focus on CTX imagery, due to its resolution (~6m/pixel) being detailed enough to detect features the size of craters and its 'field of view' wide enough to include the size range that craters can be (several metres to kilometres in diameter). Accordingly, for the experiments in this thesis CTX imagery was used, specifically for the

study in this chapter the image G05_020119_1895_XN_09N198W[21] due to its location in the centre of the Cerberus Fossae region.

NASA has openly made available a range of products to the public through their Planetary Data System (PDS) library, and as such the required CTX image was downloaded through this system. However, CTX imagery is provided in a raw unprocessed format that has not been optimised for visual display. To address this, the USGS software ISIS 3 was used to calibrate, clean and finally convert the image into a universal format suitable for online use (in this case JPEG)[22]. Finally, the resulting image was cut up into 'manageable' slices of a size, both visually and digitally, that can be quickly served and displayed on a website platform. Raw NASA imagery is often several gigabytes in size, and even when converted into a universal graphical format can be several hundred megabytes. Generally, CTX images are tens of thousands of pixels in width and length, representing a geographical area of up to 30 x 160 kilometres in total. Clearly, these large figures can cause issues both in terms of visual appearance on screen, and in terms of processing, where images can take up large amounts of computer memory. This process is in line with those used by several existing Zooniverse platforms (for instance Planet Four) that also rely on PDS remotely sensed imagery.

[21] An explanation of the CTX specification and product can be found at: http://pds-imaging.jpl.nasa.gov/documentation/MRO_ctx_pds_sis.pdf

[22] A full description of this process and the workflows involved available at: http://isis.astrogeology.usgs.gov/IsisWorkshop/index.php/Working_with_Mars_Reconnaissance_Orbiter_CTX_Data

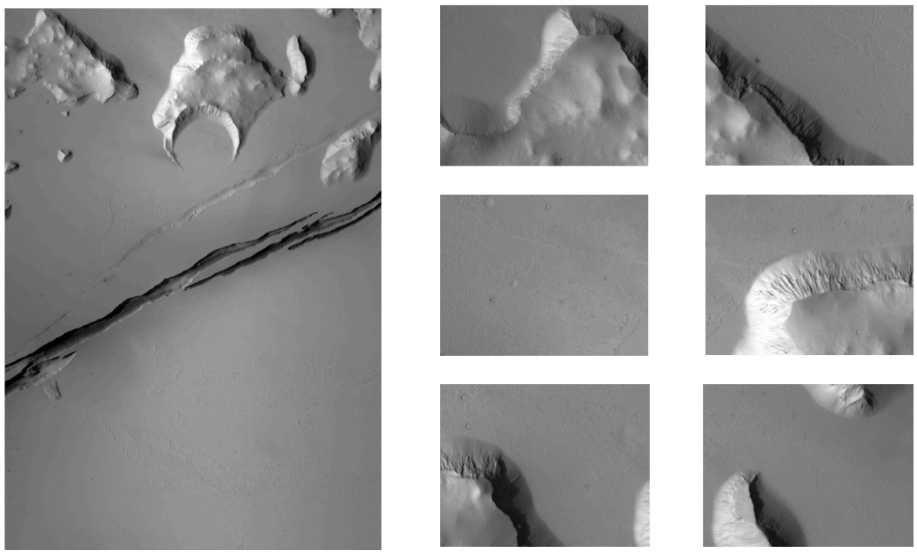


Figure 8.2: Original G05_020119_1895_XN_09N198W image (left) and first row of slices (right)

A total of 78 smaller image 'slices' were created using the standard Zooniverse workflow and software, measuring 840 x 648 pixels (the standard size used for Zooniverse projects) with an included overlap of 100 pixels to ensure features on the edges of each slice were adequately displayed. Figure 8.2 shows the original CTX image along with the first row of image slices.

8.4 RESEARCH QUESTIONS AND HYPOTHESES

Task workflow design factors are studied in a laboratory setting where participants' environment, equipment and workload are standardised throughout. To test the effects of task workflow design, the design-influenced factors of autonomy, variety, task type and user judgement (independent variables 1-4) were operationalized through the creation of three separate platform interfaces. This experiment attempts to answer the question:

RQ: Which task workflow design configurations do volunteers prefer?

And in relation to the research questions introduced in 4.5.4:

RQ_3: Does task workflow design influence volunteer results?

In doing this, the experiment firstly attempts to underpin the results found in the studies of chapters 6 and 7 by investigating if the opinions shared by volunteers about task workflow design factors are consistent when considering a fixed scientific discipline and case with the same image set and stimulus. Secondly, it builds on earlier findings by considering the effect of task workflow design factors on the volunteer scientific data collated, and its accuracy when compared with the expert equivalent. In addressing these two questions regarding the effect of task workflow design, the study also answers the sub-question:

RQ_a: What is the relationship between these effects?

For example, does a task workflow design configuration that results in greater analysis efficiency correspond to a more satisfying volunteer experience? Conversely, does a better analysis coverage (i.e. more crater markings) have an adverse effect on the accuracy of the markings?

Inspired by the related human factors work summarised in chapter 3, that subsequently resulted in the design-influenced postulates described in section 4.4.3 (P5-11), testable hypotheses were formulated for a mix of qualitative and quantitative dependant measures regarding volunteer behaviour, experience and performance:

H1: Volunteers using an interface with more autonomy perform better.

H2: Volunteers using an interface involving more task variety have more motivation.

H3: Volunteers performing a less 'involved' task and judgement produce greater data coverage.

H4: The type of task performed and judgement made by the volunteer influences volunteer agreement (in terms of the data produced).

8.5 EXPERIMENT DESIGN

In a within-subjects design, the task workflow design 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-TLX type statements to assess volunteer opinion and perceived workload. As

with the study in chapter 7, 'free-response' replies were again sought in order to add context to the findings. Figure 8.3 gives an overview of the experimental design in relation to the model of task workflow design factors derived in section 4.3 and the measures of scientific outcome described in section 4.5.3.

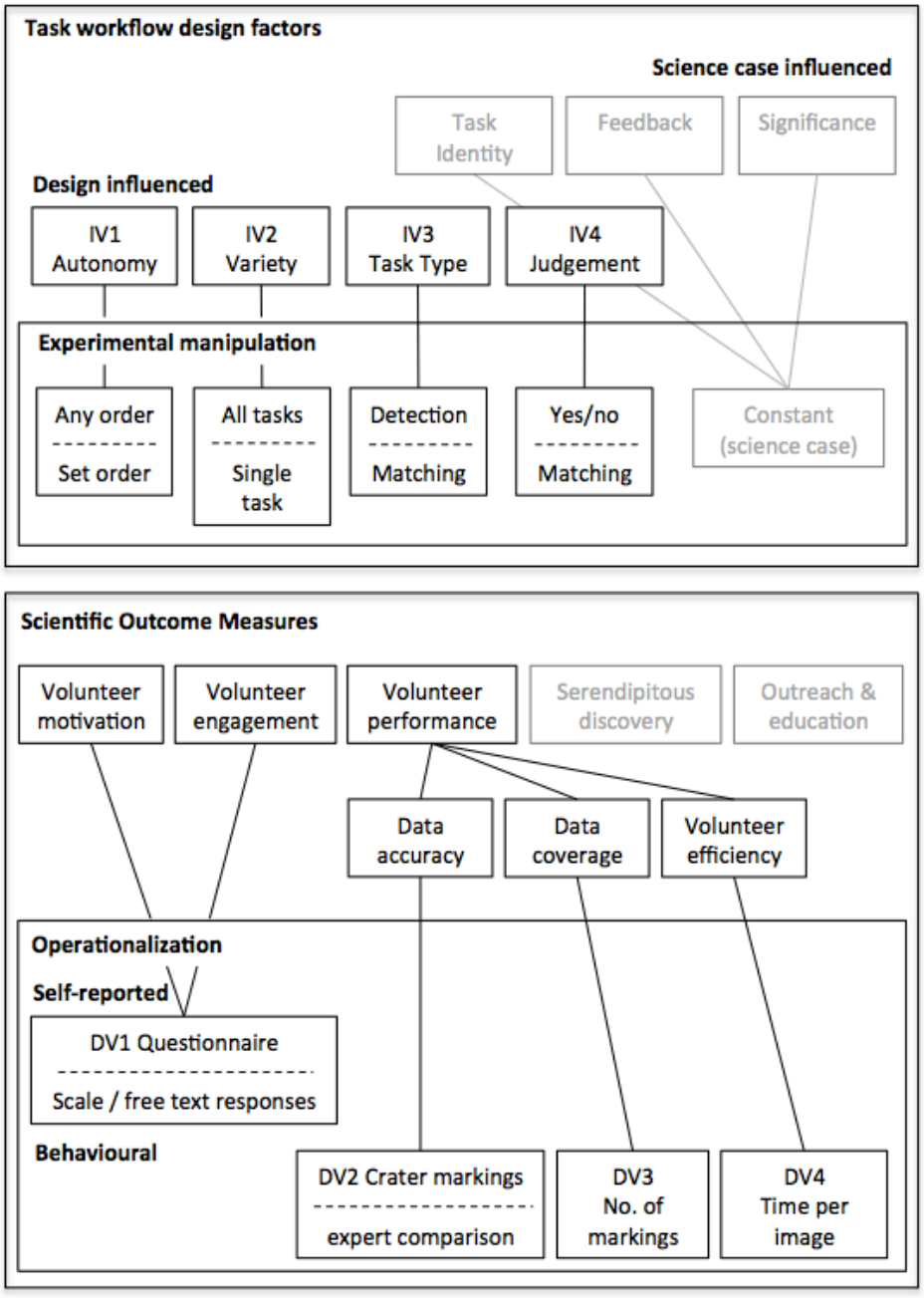


Figure 8.3: Overview of the experimental design in relation to the model of Task Workflow Design factors

The four design-influenced TWD factors (IV1-4) are experimentally manipulated through the design of three VCS classification interfaces, and the impact of this manipulation on the operationalisation of scientific

outcome is then measured. This is done by participants' self-reports giving their opinions on motivation, engagement and experience (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). Other science case influenced TWD factors are held constant by experimental design, i.e. by addressing the same science case across each condition (interface).

8.5.1 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 in order to allow participants to mark craters, and the types of task and judgement they involve. Figure 8.4 shows how these tools are presented to the participant and the types of annotation they produce.

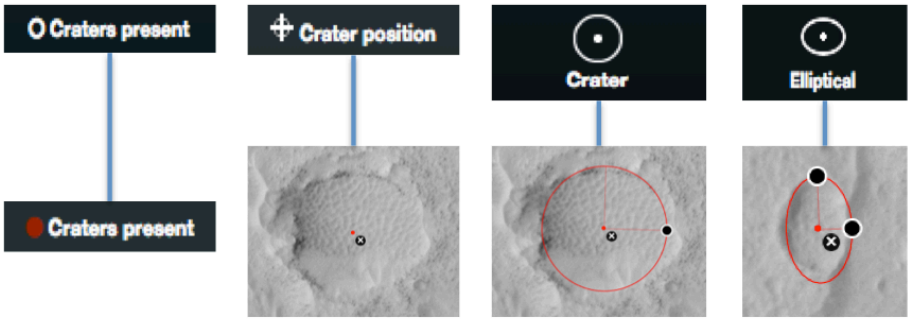


Figure 8.4: Platform tools (above) and annotations (below)

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 indicating 'yes'). In essence, it involves a detection task through making a forced choice (yes/no) judgement.

Crater Position tool: This tool allows the user to mark the centre 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 centre of the crater) through making a matching judgement.

Crater tool: This tool allows the user to mark a circle around the edge of each crater, by clicking in the centre and dragging the cursor to the edge. It also allows the user to resize the circle to 'fine tune' its final position. This involves again a detection task and 2 matching tasks for each crater (the centre & the edge) through making matching judgements.

Ellipse tool: This tool allows the user to mark an elliptical shape around the edge of craters that appear less uniform. Although considerably more rare than circular craters, they are important to identify as they are believed to be caused by secondary impacts (i.e. not directly from a meteor, but from ejecta caused by a separate event) and so need to be discounted from crater counting analysis calculations (Michael & Neukum, 2010). The user clicks in the centre of the crater and drags the cursor to the furthest edge. Two resize tools then allow the user to fit the longest and shortest radii to the shape of the crater edge. This tool involves a detection task and 3 matching tasks for each crater (the centre, furthest edge and nearest edge) through making matching judgements.

8.5.2 Classification Interface Design

In order to experiment with the manipulation of task workflow design factors, the platform has been developed to include three different classification interfaces for marking craters. By varying the way that some or all of the tools outlined in the previous section are presented to the user the interfaces differ in terms of task type, judgement, task variety and autonomy. Figure 8.5 shows the tools available to each user and when for each interface, namely stepped, ramped and full:

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, and moves on to the next once they have indicted 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 required).

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 (in the case of the laboratory study 15) the tool changes i.e. the participant presses or depresses the 'craters present' button for 15 images,

then marks the centre of the craters with the 'crater position' tool for 15 images etc. Each tool change represents a step up in complexity.

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. This includes the decision on how many of the tools to use for each image, for instance if an image has a large number of craters present the participant may choose to just press the 'craters present' button and move on, without physically marking any of them with the other tools provided. The full interface bears the most resemblance to the design of the current live Planet Four platform that studies fans.



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.

Figure 8.5: Flow diagram of tools available to the user for each interface

In terms of the independent variables 1-4 laid out in figure 8.3, the interface configurations are as shown in table 8.1 while figure 8.6 shows screenshots of the three interfaces as they appear on the final platform:

Table 8.1: Task workflow design configuration of each interface

Interface	Autonomy	Variety	Task Type	Judgement
Stepped	Set order	All tasks	Detection & matching	Yes/no & matching
Ramped	Set order	Single task	Detection or Matching	Yes/no or matching
Full	Any order	All tasks	Detection & matching	Yes/no & matching

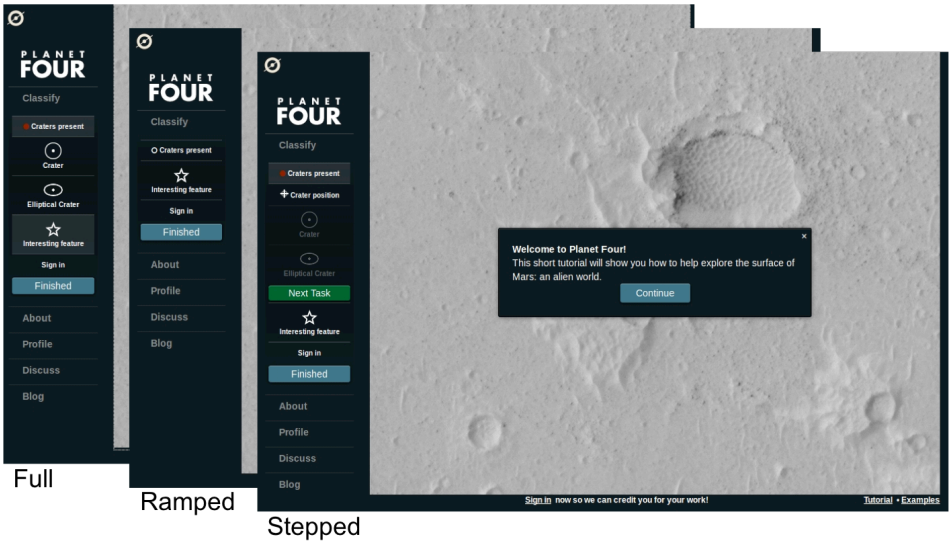


Figure 8.6: Screenshots of Planet Four interfaces

8.5.3 Other Platform Pages

While the focus of the experiment in this chapter is the manipulation of the classification interface, the platform also consisted of a number of other pages that fulfilled a more general information and 'sign posting' role[23]. These pages were standardised across each classification interface, and were constructed based on the general format and design of the original Zooniverse Planet Four site. They were included both to make the laboratory experiment as authentic as possible in terms of the guidance

[23] A hierarchical diagram of their relationship can be found in Appendix C of this thesis

given to the participant, and also to aid the transition to a 'live' VCS platform for the experiment in chapter 9. The pages are as described:

Home page: The landing page of the site that each participant first sees, and from where each other page including the classification interface can be accessed.

Classify page: The page containing the three different interfaces as previously described, where participants carry out the crater marking tasks.

About page: Page describing the science case of the site (i.e. the reason volunteers are needed), the source and format of the images displayed and the reason for the three different interfaces provided.

Classification examples: A page explaining examples of different types of craters that might be seen and what they look like, for reference.

FAQs: A list of frequently asked questions regarding craters and how to mark them, with explanations and answers. The content of this page was heavily influenced by the Zooniverse's Moon Zoo platform that requires volunteers to carry out a similar task but on images of the lunar surface.

Education: A page explaining the different ways Zooniverse projects can be used in the classroom, including links to external NASA created educational resources.

The scientists: Biographies of the scientists involved in the platform, both in terms of the science case (based at the University of Bristol's school of Earth Sciences) and in terms of the aims of this thesis (the author and supervision team).

Profile page: A link to the volunteers' Zooniverse profile, giving information about their contribution and any 'favourite' images tagged. For the purpose of this experiment, each participant was provided with a new account login making this page effectively redundant.

Planet Four Blog: A link to the original Planet Four blog, giving information and updates regarding the progress of the project. Again, as the information is generally regarding the original science case involving seasonal fans, participants were not encouraged to visit this page.

8.5.4 Questionnaire Design

In order to collate participant views and opinions, each participant completed a questionnaire[24] after using each interface. The design of the questionnaire followed that of the one used for the study in chapter 7, containing Likert-type statements and 'free-text' responses under the categories *design & usability*, *tasks & tools* and *imagery* as described in section 7.3.1. The *discussion & feedback* section has been omitted as this was a 'self contained' study without an established community or 'live' feed to the expert community afforded by a fully launched platform. As before, prior to these three sections, participants completed a general demographics section regarding their background and experience. Unlike the study in chapter 7 however, where participants completed the questionnaire online at a time of their choosing, on this occasion the questionnaires were completed via hard copy (i.e. pen and paper) immediately after using each interface. This was in order to minimise the interruption to the participants' platform experience, removing the need to 'switch' between different screens associated with services external to the VCS platform.

8.5.5 Participants

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). 8 of the participants had actively taken part in the original Planet Four project previously, while 10 participants had never heard of the site. Regarding experience with the Zooniverse, only 4 participants had visited other sites and they were predominantly of a similar space theme (Galaxy Zoo (4), Planet Hunters (1), Ancient Lives (1)). Participation in the experiment was compensated with a £10 gift voucher.

[24] An example of the questionnaire can be found in Appendix C of this thesis.

8.5.6 Design and Procedure

The study took place in the same room and was carried out on the same laptop for each participant to ensure factors such as lighting conditions and screen setup remained constant, as differences could otherwise influence an image analysis task (see section 3.3). After a short introduction by the author outlining the procedure and gaining their informed consent, participants sat in front of the laptop with the home page of the platform displayed. After familiarising themselves with the site, when ready they used each of the interfaces in turn to mark craters on a set number of the image slices, with the interface order differed (same number of participants doing each order) to account for bias caused by learning of the system or fatigue. 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.).

Before using each interface, each participant completed an online tutorial learning how to use the tools, marking craters on a separate example image (as shown in figure 8.6). The marking and behavioural data collected regarding the example image was not included in the data analysis later on in this chapter. After using each interface each participant completed a questionnaire as explained in 8.5.4 asking them to rate their agreement, via a nine-point Likert scale, with a number of different statements across the categories described. At the end of each section, 'free text' boxes were available for participants to give additional comments and opinions to add context to any findings. Overall the study was designed to last no longer than 2 hours per participant.

8.5.7 Role

The experiment was carried out under supervision at the Nottingham Geospatial Institute, University of Nottingham. The author had full responsibility in agreement with the supervision team to perform all activities necessary to conduct the experiment. This included experimental design, orchestration, participant recruitment, set up and support, briefing, and debriefing. Additionally, the technical infrastructure that supported the

experiment was designed and implemented by the author, namely the classification interfaces and additional page layout and functionality. This was done with the support of the Zooniverse development team based at Adler Planetarium, Chicago, who kindly gave access to the original Planet Four platform 'front end' code.

8.6 RESULTS AND ANALYSIS

As outlined in figure 8.3, dependent variable measures have been recorded both through participant self-reporting and more behaviourally through crater marking performance. The following section presents the analysis and results for each method in terms of their relation to the independent variables regarding task workflow design.

8.6.1 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 significant difference between each interface. This is perhaps as expected as many of the design features, tasks performed and tools used along with the image format are constant throughout the experiment.

When considering the visual attractiveness of the site and image quality, participant responses were positively skewed. Over 80% of all raw scores across each interface lie between 6 and 9 (very attractive/high quality). Similarly, when considering the NASA-TLX statements regarding perceived task workload, over 70% of all scores lie between 1 and 4 (low demand) with no significant difference existing between each interface (figure 8.7).

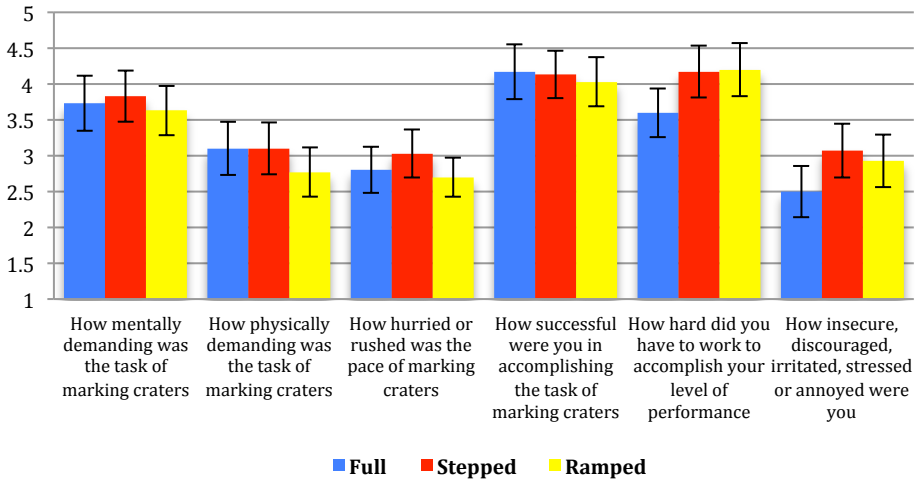


Figure 8.7: Participant responses to NASA TLX statements (with standard error shown)

In terms of responses to the statement *'I found the crater marking tool easy to use'*, again there is no significant difference in scores between each interface. Participant responses showed no clear preference, with the Full, Stepped and Ramped interfaces having mean scores of 6.00 ± 2.45, 5.70 ± 1.69 and 5.77 ± 2.14 respectively (1 being strongly disagree and 9 being strongly agree). This is also true when considering the elliptical tool, with mean scores of 5.61 ± 2.02, 5.91 ± 1.56 and 5.57 ± 1.85 respectively.

Regarding the statement *'I would like more tools to mark different shapes and features'* there is no significant difference in participant response between each interface, with mean scores of 4.13 ± 2.62, 4.07 ± 2.68 and 4.50 ± 2.66 for the Full, Stepped and Ramped interfaces respectively. In terms of the consistency of image quality, participant responses were positively skewed across each interface, with over 90% of responses being between 7 and 9 indicating strong agreement. Finally, responses to the statement *'I didn't know what to do if an image had no craters to mark'* again were positively skewed across each interface, with over 90% of responses lying between 1 and 3 indicating strong disagreement. Figure 8.8 shows mean participant score per interface for each of these statements.

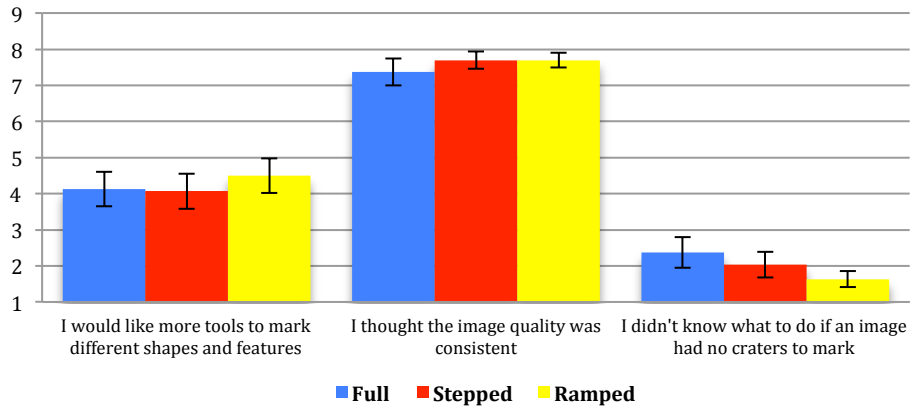


Figure 8.8: Participant responses per interface (with SE shown)

However, differences between each interface do exist when considering how quickly participants 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 indicated that they learnt the Full interface more quickly than the Stepped (mean score of 7.90 ± 1.54 vs. 6.87 ± 2.06, $p = .001$). Likewise when considering how easy to use the participants found each interface, a statistically significant difference is shown ($F(1.817, 52.684) = 4.957$, $p = .013$). Participants scored the Full interface easier to use than the Stepped (mean score of 6.93 ± 1.68 vs. 6.07 ± 2.18, $p = .011$) and slightly easier than the Ramped (6.70 ± 1.80), although the difference is not statistically significant ($p = .118$). Participants also scored the Ramped interface easier to use than the Stepped ($p = .039$), and overall participant responses suggested that the Stepped interface is the least easy to use and access out of the three.

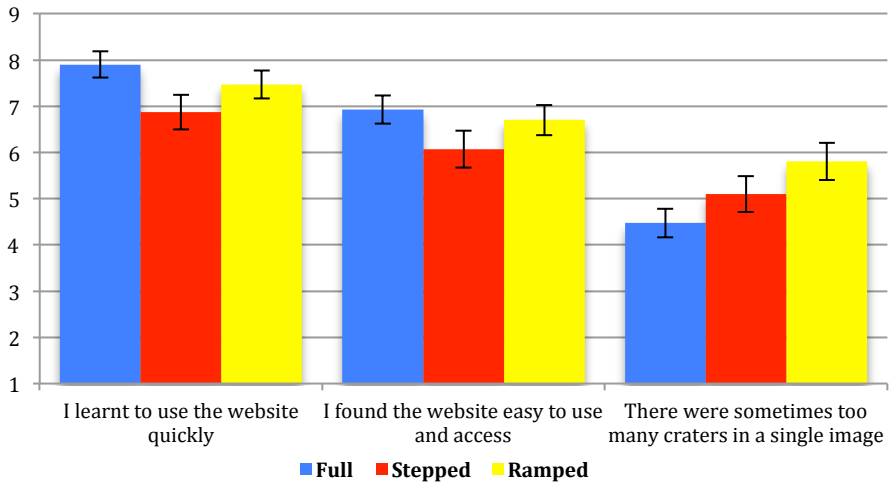


Figure 8.9: Response differences per interface (with SE shown)

In terms of the number of craters in each image, there is also a difference in participant responses across the interfaces ($F(1.811, 52.507) = 5.184$, $p = .011$). Participants felt that there were sometimes too many craters in an image with the Ramped interface compared to the Full (mean score of 5.80 ± 2.20 vs. 4.47 ± 2.05, $p = .004$) and slightly more than the Stepped (5.10 ± 2.16), although the difference is not statistically significant ($p = .053$). Figure 8.9 shows the average participant level of agreement 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 the scores. The responses for the three sections of the questionnaire have been collated for each interface and subsequently coded under 4 categories: *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 were specifically about the tools provided to mark craters; and imagery comments discussed the remotely sensed imagery displayed. Table 8.2 shows a breakdown of the number of comments made by participants regarding each interface, sorted into the four different categories.

Table 8.2: Number of responses by category and interface

Topic	Full Interface	Stepped Interface	Ramped 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 were predominantly regarding the autonomy allowed by each interface, in terms of the tools available and order of use (participant S19):

> *"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 (R4):

> *"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 involved the marking of smaller craters, this time directly attributing the problem to tool use and design (R19):

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

Other comments in this category indicated the benefit a 'zoom' tool would have for navigating around the image, as with the findings in chapter 7 (see 7.4), presumably to aid the marking of smaller, less clear craters (F11):

> *"A zoom function would be nice…"*

Finally, comments in the *imagery* category either concerned the content of the imagery, i.e. the number of craters found in an image (S11):

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

Or were more general comments regarding the quality of the imagery (F7):

"Too bad the quality is low…"

Other participant comments that do not fall into the categories described were either general praise regarding the overall appearance of the website or requests for more information about the overall scientific goal, all of which were independent of the interface being used.

8.6.2 Crater Marking Results

This section summarises the results of the behavioural measures described in figure 8.3, i.e. the participants' crater marking data (DV2). Firstly, general behavioural statistics are compared across each interface (DV3 & DV4), followed by a between-participant comparison in terms of marking agreement. Finally, participant performance when using each interface is measured through comparison with expert data.

8.6.2.1 Crater Marking Behaviour

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 on each image. As explained in the experiment design section of this chapter, the Ramped interface requires participants to use one tool over a number of images, and then another tool etc. therefore Ramped (position) represents results where participants only mark the centre of craters and Ramped (mark) represents results where participants mark the shape. The comparatively large values of standard deviation can be explained by image variation, with some images having no craters present and others having several dozen (a common occurrence for VCS platforms involving planetary data).

When considering the percentage of participants that marked at least one crater per image, a Friedman test revealed a statistically significant difference between each interface ($X^2(3) = 6.1$, $p = .05$). Post hoc analysis using the Wilcoxon signed-rank test with a Bonferroni correction applied revealed that participants were more likely to identify that an image has at least one crater and mark it using the Ramped (position) interface compared to the Stepped (63.24 ± 33.76% vs. 53.67 ± 33.21% of

participants 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 Ramped (mark) are 58.24 ± 35.45% and 57.94 ± 35.02% respectively.

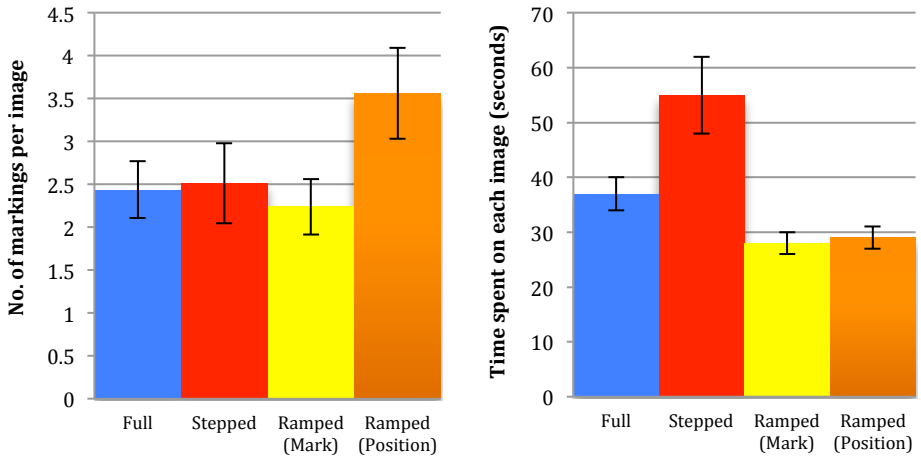


Figure 8.10: Crater marking results (with SE shown)

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 Ramped (position) interface resulted in a greater number of markings (3.61 ± 4.67) compared to the Full (2.46 ± 2.93, p = .001), Stepped (2.55 ± 4.17, p = .003) and Ramped (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 Stepped interface compared to the Ramped interfaces (55 ± 64 seconds vs. 28 ± 16 seconds, p = .001) and more time compared to the Full (37 ± 27 seconds, p = .01). Figure 8.10 shows the average number of crater markings and the average time spent on each image using each interface, along with standard error.

8.6.2.2 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 participants 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[25]. The function takes a nearest neighbour clustering approach, performed using x and y position values and diameter. The maximum linking length (i.e. the furthest apart two markings can be to be considered within the same cluster) was set as the diameter field, ensuring markings belonging to the same cluster at least have some overlap. After performing the function on each interface dataset, the clustering results were reviewed by the author in order to check for any obvious omissions or incorrectly amalgamated results. Whilst 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, Batistakis, & Vazirgiannis, 2001), and often inspection by eye is the best solution.

The minimum number of markings of the same feature required to form a cluster was set at two. By doing this, all single markings of a feature are removed from the analysis – which should remove the majority of malicious / error markings (assuming the likelihood that two participants would replicate the same malicious mark or error is small). Previous research has shown that increasing the number of markings that contribute to a cluster can improve crater-marking performance when compared to an expert (Bugiolacchi et al., 2016), and often weight clusters based on marking contribution. It could be argued that including clusters consisting of only two markings will reduce overall performance, and is not a direct representation of the current analysis methods used in crater marking citizen science. However, it does allow for each interface to be compared in terms of this effect (for instance, does the same number of contributions to a cluster on each interface produce the same level of performance?), and

[25] A description of the function can be found at: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/grouping-analysis.htm

also for an appraisal of the trade-off between data volume and performance to be reviewed.

Table 8.3 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 centre position and the average standard deviation of the cluster crater diameter in terms of screen pixels (i.e. the level of agreement between participants).

Table 8.3: Crater marking agreement for each interface

	Full Interface	**Stepped Interface**	**Ramped (Mark)**	**Ramped (Position)**
No. of crater clusters	182	185	163	297
SD of position	2.64 ± 0.35	2.41 ± 0.46	2.67 ± 0.40	2.05 ± 0.07
SD of diameter	6.52 ± 0.57	7.70 ± 0.57	5.10 ± 0.44	-

Although participants using the Full and Stepped interfaces have identified a similar number of crater clusters (182 and 185 respectively), it is clear that the Ramped (Position) interface has resulted in more crater clusters being marked (298, ~ 61% greater). Conversely, the Ramped (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 clusters marked, 86 were identified across all four interfaces, and so can be compared like-for-like in terms of participant agreement. 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 Stepped interface varied significantly less (greater agreement) than those made using the Full and Ramped (Mark) (standard deviation of 2.00 ± 0.90 vs. 2.53 ± 1.15, $p = .001$ and vs. 2.57 ± 1.51, $p = .002$ respectively). Likewise, position markings made using the Ramped (Position) interface showed significantly greater agreement than those made using the Full (2.20 ± 0.91 vs. 2.53 ± 1.15, $p = .033$) and greater agreement than those made using the Ramped (Mark) though the difference is not significant ($p = .065$).

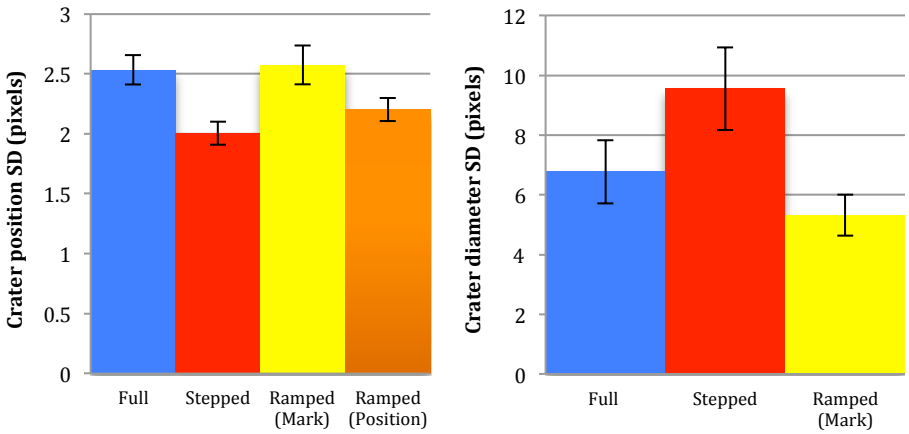


Figure 8.11: Comparison of marking agreement per interface

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 Stepped interface (standard deviation of 9.55 ± 13.26) showed significantly less agreement than those made using both the Full (6.77 ± 10.10, $p = .001$) and Ramped (Mark) (5.32 ± 6.57, $p = .001$). The diameter of markings made using the Ramped (Mark) interface also showed significantly more agreement than those using the Full ($p = .007$).

Breaking down participant agreement in terms of crater identification (i.e. clusters) further, figure 8.12 shows the number of crater clusters marked verses the minimum number of participants that marked them.

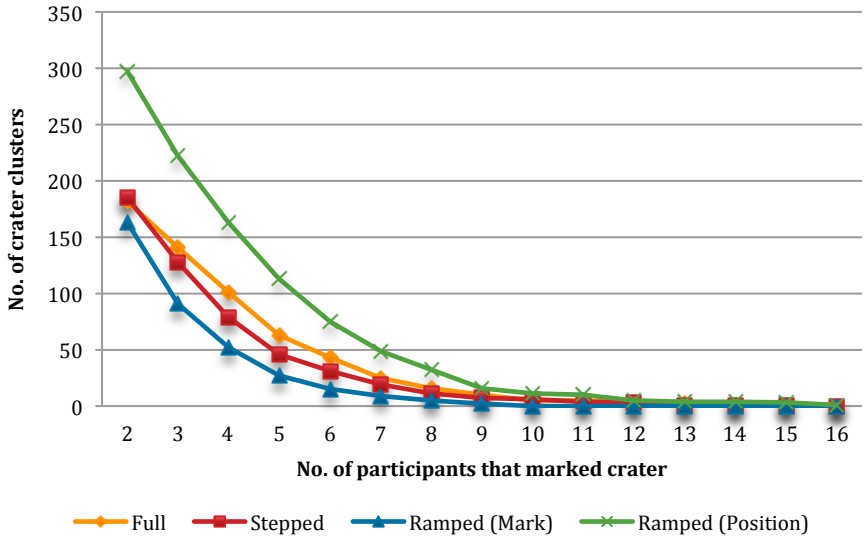


Figure 8.12: Participant crater identification agreement

As can be seen, the pattern of a greater number of crater clusters identified using the Ramped (Position), and fewer identified using the Ramped (Mark), continues when considering craters identified by up to 9 different participants. There are only a few occasions when more than 9 participants have agreed on the identification of a crater, independent of the interface used. Although participants using the Full and Stepped interfaces have identified a similar number of crater clusters (182 and 185 respectively), the Full interface shows slightly better agreement, with more craters being identified by between 3 and 8 participants.

Finally, table 8.4 breaks down the number of crater clusters identified using each interface in terms of their size. As explained in section 8.2, and illustrated in figure 8.1, the crater counting analysis method for aging the surface is highly dependent on the size distribution of the craters as different sizes are thought to occur at differing rates. The size 'bins' shown have been chosen to match those previously used by the expert team (J. Taylor et al., 2013).

Table 8.4: Number of crater clusters per size bin

Size Bin (Diameter)	Full Interface	Stepped Interface	Ramped (Mark)
>0-10 Pixels	17	21	25
>10-20 Pixels	137	110	113
>20-40 Pixels	24	51	24
>40 Pixels	4	3	1
Total:	182	185	163

Although participants using both the Full and Stepped interface have identified a similar number of crater clusters, the Full interface has resulted in a greater number of clusters between 10 and 20 pixels in diameter (75% vs. 59% of the total) while the Stepped resulted in more between 20 and 40 (28% vs. 13%). Despite the fewest number of clusters, the Ramped (Mark) interface has resulted in proportionally more clusters of a diameter less than 10 pixels (15% vs. 9% (Full) and 11% (Stepped) of the total).

8.6.2.3 Participant – Expert Comparison

One of the advantages of developing a purpose-built science case and having a dedicated science team is the access to previous research and expertise this affords (as explained in section 5.3.2). Previous work carried out by experts from the University of Bristol has identified 365 separate craters on the CTX image used in this study, 280 of which have been identified as a crater cluster (marked by more than one participant) on at least one of the interfaces. Although, as suggested by Robbins et al. (2014), it is not possible to be certain that the experts have captured every crater, due to the science teams' specific interest in a relatively small area (1 CTX image) and their consideration of craters down to the size of a few pixels in diameter (towards the limit of existing marking tools) their analysis is considered a suitable ground truth for the purpose of this study. Table 8.5 contrasts the crater marking clusters made using each interface against the expert benchmark regarding crater identification.

Table 8.5: Crater identification compared to expert

Interface	True Positive	False Positive	False Negative
Full	168	14	197
Stepped	170	15	195
Ramped (Mark)	148	15	217
Ramped (Position)	247	50	118

The performance (assuming the expert data as a ground truth) of participants using the Full and Stepped interfaces was very similar, with a comparable number of identified clusters also confirmed by the expert (~92% of clusters), a similar number misidentified by participants (~8% of clusters) and therefore a similar number of craters identified by the expert but not marked as a cluster by participants (~54% of expert markings). The lower number of crater clusters identified by participants using the Ramped (Mark) interface equates to a lower number confirmed by the expert (~90%) and a greater number of craters identified by the expert but missed by participants (~59%). Finally, although the Ramped (Position) interface resulted in a greater total number of clusters marked that were

confirmed by the expert, proportionally more clusters were misidentified as craters (~20% of the total). The greater number of clusters made however has resulted in fewer of the expert markings being missed by participants (~32%).

Out of the 365 expert crater markings made, 84 of them were subsequently correctly identified as a crater cluster on all four of the different interfaces, and so can be directly compared in terms of their variation from the expert equivalent. Figure 8.13 shows a 'slice' of the study image, with the average markings made using each interface along with those made by the expert.

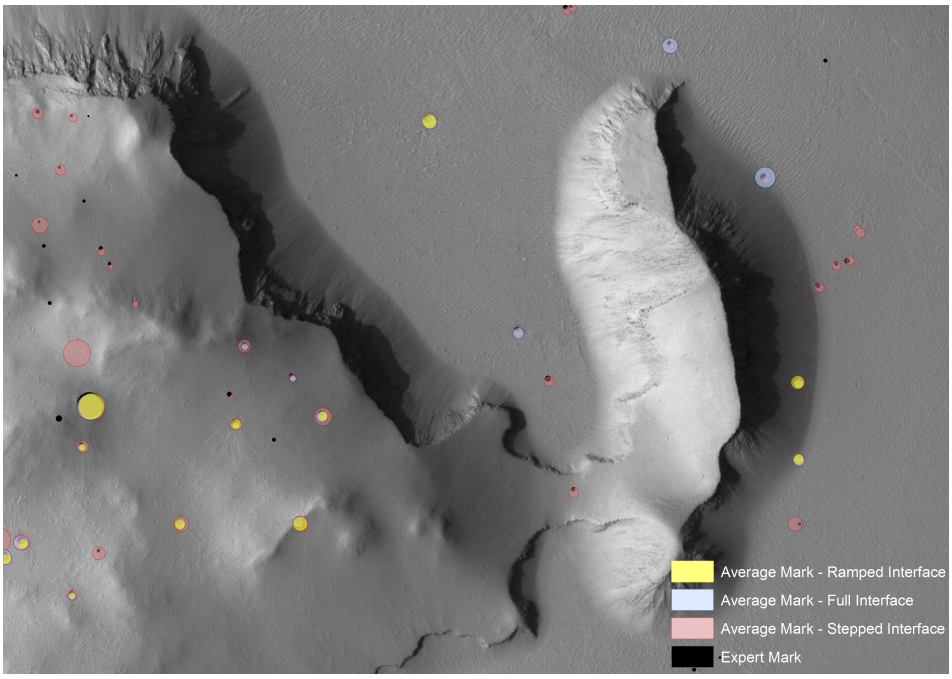


Figure 8.13: Expert and participant average crater markings per interface

A repeated measures ANOVA test with a Greenhouse-Geisser correction showed a significant difference between each interface ($F(2.619, 217.379) = 2.075$, $p = .05$) when considering the average difference between participant crater position and expert crater position. Post hoc tests using the Bonferroni correction revealed that the position of markings made using the Ramped (Position) interface were significantly closer to the expert equivalent than those made using the Full and Ramped (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 Stepped interface (4.21 ± 1.13) were also closer in position to the expert

than the Full and Ramped (Mark) interface, though the difference was not significant.

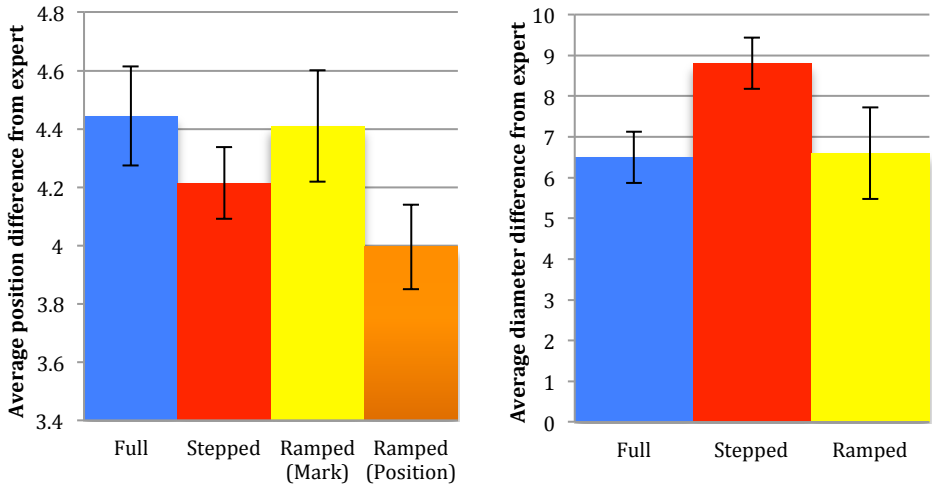


Figure 8.14: Position and diameter difference of participant markings compared to expert

Regarding the average difference between participant crater diameter and the expert equivalent, a repeated measure ANOVA again showed a significant difference between each interface ($F(1.499, 130.415) = 5.439$, $p = .011$). Post hoc tests using the Bonferroni correction revealed that the diameter of crater markings made using the Stepped interface were significantly more different than the expert equivalent compared to those made using both the Full and Ramped (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 8.14 shows the average difference in position and diameter between the expert markings and those made by participants using each interface.

Breaking down participant crater identification further, table 8.6 compares the expert data with the crater marking clusters made using each interface but this time across the different size bins as explained at the end of section 8.6.2.2. Participants had difficulty correctly identifying craters >0 – 10 pixels in diameter on all of the interfaces, with between 58-64% of craters not marked at all. Of those that were correctly identified the majority were falsely marked as larger craters, although participants using the Ramped (Mark) interface marked a greater number of craters within the correct size range (7.9% vs. 4.0% (Full) and 5.0% (Stepped)). Participant performance was comparable across each interface when considering

craters of diameter >10 – 20 pixels, and an improvement on the smaller size. Between 63-69% of craters were correctly identified, and of that number between 59-62% were also marked within the correct size bin.

Table 8.6: Participant crater marking sizes compared to expert

Size Bin (Diameter)		Full Interface	Stepped Interface	Ramped (Mark)
>0 – 10 Pixels	True positive:	12	15	24
	False positive:	112	112	86
	False negative:	179	176	193
>10 – 20 Pixels	True positive:	23	23	20
	False positive:	14	16	14
	False negative:	17	15	20
>20 – 40 Pixels	True positive:	4	2	3
	False positive:	0	0	0
	False negative:	1	3	2
>40 Pixels	True positive:	3	2	1
	False positive:	0	0	0
	False negative:	0	1	2

When considering larger craters (>20 pixels in diameter) all of the craters correctly identified by participants were also correctly marked in terms of size range across each interface. The full interface resulted in a greater coverage, with 87.5% of craters correctly identified compared to 50% when participants used the Stepped and Ramped (Mark) interface.

Finally, figure 8.15 shows the number of participants that contributed to a cluster verses the percentage that are true positives, when compared with the expert data. In effect, it shows the number of participant markings 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 of the clusters of 4 or more participant contributions also recognised by an expert equivalent. This was followed closely by those made on the Full interface where 5 or more participant markings were needed. Of the clusters identified using the Stepped interface, those made up of 7 or more participant contributions were also recognised by the expert as a crater. The Ramped (Position) interface required the highest number of participant markings, as only clusters made up of 8 or more resulted in a 100% expert agreement, twice as many markings as was required using the Ramped (Mark) interface.

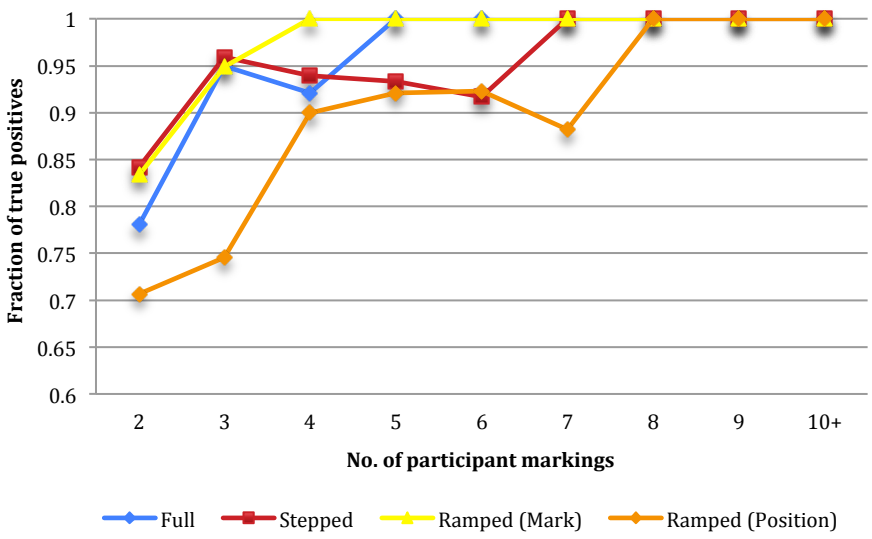


Figure 8.15: Cluster participant contribution verses fraction of true positives

It should be noted that while the fraction of true positives raises steadily with the number of participant markings per cluster made on the Ramped (Mark) interface, there is a slight dip of ~2% before raising 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 have the appearance of craters, at least in the eyes of the majority of participants.

8.6.3 Summary and Hypotheses

- Hypothesis H1 (greater autonomy = better performance) is not supported by the findings regarding measures of performance, with the Full interface not resulting in a significant improvement.
- Hypothesis H2 (greater variety = more motivation) 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 Ramped interface, however participant free responses suggested autonomy was more important.
- Hypothesis H3 (simpler task = greater data coverage) is supported by the finding that more crater clusters were marked using the Ramped (Position) interface.

- Hypothesis H4 (task type and judgement influences participant agreement) is supported by the finding that the Ramped (Position) resulted in greater participant agreement in terms of identification (markings per cluster), and that position markings created using the Stepped and Ramped (Position) interfaces (where it is separated out as an individual task) were significantly more in agreement than those made using the Full and Ramped (Mark) interfaces.

In summary, the hypotheses related to task type and judgements 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 both the self-reported and behavioural findings of the study.

8.7 DISCUSSION

In this section the self-reported quantitative and qualitative findings of the questionnaire are combined with the quantitative measures of participant agreement and performance to paint a broader picture of how task workflow design factors can effect a virtual citizen science platform's output and the motivation of its volunteers. Participants' quotes are used to highlight their issues and concerns and are related back to their behaviour and performance.

8.7.1 The Effect of TWD on Participant Engagement and Motivation

Participant responses to the NASA-TLX type statements regarding the workload of the crater-marking task showed no 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 inherent task of identifying and marking a crater and the imagery presented were constant across the study. This stance 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 regarding accuracy and the marking tools were overwhelmingly about the difficulty of marking

small craters (a point that will be further discussed throughout this section), an issue raised across all study conditions.

Differences in participant responses were present however when considering the number of craters in an image. Participants using the Ramped interface showed a significantly greater agreement with the statement that there were too many craters to mark, suggesting an issue with the repetitive nature that only having one task to complete brings. This is supported by a number of participant quotes, for instance (R11):

> *"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*."

As with the findings regarding the original Planet Four site in chapter 7, this demonstrates the importance of variety when considering the volunteers' intrinsic motivation to take part (Dubinsky & Skinner, 1984).

The greatest concern of participants however, according to their questionnaire responses, was regarding autonomy. Participant Likert-type responses showed that they found the Stepped interface significantly less easy to use and access, and the reasons for this were explained by their 'free' responses (S12):

> *"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."*

This suggested that participants were not in favour of being prescribed an order of completing the tasks, as with the Stepped interface, especially when they were aware of other tasks that needed to be done. Furthermore, the number of comments related to this issue suggests that the importance of autonomy on volunteer satisfaction as found in previous research (Chung-Yan, 2010) could have a larger effect on volunteers' intrinsic motivation than task variety when considering the VCS case.

8.7.2 Greater Coverage: But at What Cost?

Analysis of the crater markings made by participants showed that the Ramped (Position) interface resulted in the greatest coverage; with significantly more participants marking at least one crater per image, a

greater number of markings per image, more crater clusters identified and more participant markings per cluster. This supports hypothesis H3 (Eveleigh et al., 2014), with the simpler position task (1 identification, 1 matching – I mouse click) resulting in greater coverage compared to the more involved tasks of the other interfaces (at least 1 identification, 2 matching – 1 mouse click and drag to size). However, while this greater coverage has resulted in 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, clusters made using the Ramped (Position) interface also required more participant markings (at least 8) before the fraction of true positives reached 100%. Additionally of course the results garnered from the Ramped (Position) interface are less detailed, returning only the position of identified craters and not the size. Although, as explained in the methodology section, the interface advances to include the full marking tools after a number of images - the advantage of greater coverage is then lost, with the Ramped (Mark) interface returning fewer clusters and true positives than the Stepped and Full interfaces.

The practical outcome of these findings is that the science team involved would have to decide if the advantage of greater coverage, and therefore fewer unidentified features, outweighs the disadvantages of having to somehow deal with the increase in false positives and needing a greater number of volunteers to classify each image to ensure the markings made 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.

8.7.3 Volunteers' Time is Precious

Participants' 'free' responses to the questionnaire regarding autonomy indicated a frustration at having to revisit the same crater several times when using the Stepped interface, alluding to a feeling that their time was being wasted. This opinion is supported by findings regarding crater-marking behaviour, as whilst participants spent significantly longer on each image using the Stepped interface, this did not translate into better

performance in terms of inter-participant or expert agreement compared to the other interfaces. In fact when considering crater diameter, markings made using the Stepped interface were considerably less in agreement, both between participants and compared to the expert equivalent. This suggests that not only are participants frustrated with having to revisit craters they have already marked, but that when undertaking the final task in the process (that of marking the size) this has translated into a loss of performance.

Furthermore, while 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. With the time that volunteers can donate to a VCS platform being finite, the extra time taken per image could very well translate into less images being classified overall.

8.7.4 Be Direct with your Volunteers

Analyses of participant markings supports hypothesis H4, showing that the type of task presented and judgement required of the participant can affect marking agreement (Hutt et al., 2013). It has also been shown that this relationship also exists when comparing markings to the expert equivalent. Expanding on this, the results suggest that being direct with volunteers in terms of the specific data or measure that is required improves both inter-participant agreement and performance compared to the expert. For instance, markings of crater position showed a significantly greater inter-participant agreement when made using the Stepped and Ramped (Position) interfaces and a significantly greater agreement with the expert when using the Ramped (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 Ramped (Mark) interfaces where crater position is calculated from the size markings rather than directly measured. Likewise, measures of crater diameter showed significantly more agreement both between participants and with the expert when using the Full and Ramped (Mark) interfaces compared to the Stepped – where marking the size is one continuous task rather than

broken down into marking the position first then adjusting the size in a second step.

Although it may be tempting to both VCS developers and the science teams involved to gather as many different measures as possible from a single volunteer contribution, there are caveats. The results of this study show that if both crater position and size are an important measure to capture 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 that is to be achieved.

8.7.5 Unforeseen Advantages of a Laboratory Trial

Beyond the focus of addressing the research questions regarding task workflow design, this study has also raised a number of other issues independent of the interface manipulation that could have a practical application when designing a VCS project. By testing the platform in a laboratory setting, with invited non-expert participants, and comparing their data with a subset of the expert equivalent, these issues have been discovered and can be subsequently addressed before the platform is released and promoted 'en-masse' – when changes to interface design become an interruption.

For instance, specifically considering the crater-marking theme of the platform in this study, participants consistently had difficulty marking the smaller craters (>0 – 10 pixels in diameter), with over half of those identified by the expert not being marked at all. This could indicate a limitation of the system; whether due to the resolution of the imagery displayed, or the mechanics of the marking tools, its an important fact to know especially when the size of the crater is such a key component of the science case. Connected to this small crater issue, results also show that despite their self-reported problems regarding the difficulty of identifying and marking small craters, participants have still attempted to do so, marking the majority of them (~90%) at a larger size. This highlights that not only could there be an issue with craters being missed, but also a number of the crater markings could be biasing the data, by indicating a greater number of larger craters than actually exist.

Another issue has been raised when analysing the marking clusters, and specifically the number of participant markings that have contributed. With all of the interfaces there is a certain percentage of 'false positives' – markings that are not craters – that exist even when the number of participants contributing rises to higher levels (3-7 participants depending on interface). This suggests the existence of certain surface features or artefacts that are consistently misconstrued as craters, perhaps due to their similar appearance (figure 8.16 shows an example of an incorrectly identified feature). By being able to identify the types of feature that might be misinterpreted as the target, this could be used to inform improved training procedures that help volunteers avoid such mistakes.

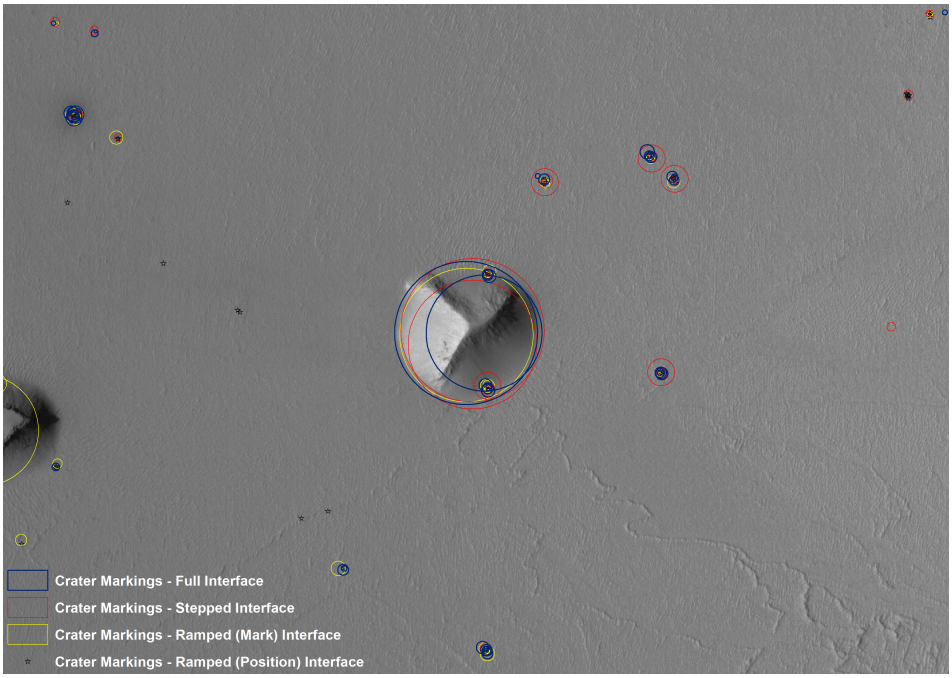


Figure 8.16: Example of false positive participant markings (centre)

Perhaps in an effort to address these issues, a number of participants have suggested their own solution in the form of an additional 'zoom' tool. Although in this case the use of such a tool may well create additional development work up front and effect aggregation when comparing markings made at different zoom levels, by running a laboratory version of the project participants have not only identified issues that exist but through feedback have also suggested solutions.

Finally, while many citizen science projects may have access to a small sample of expert data, they do not have full expert coverage that the

volunteer markings can be compared to for their accuracy and validity (hence the need for a VCS project in the first place). Existing projects have devised a number of different techniques to validate the data (see section 2.5.2) and a laboratory trial can evaluate their effectiveness before the main body of volunteer data is collected. For example the results of this study show that participant agreement is a good measure of accuracy, with the interfaces where volunteer marking variability is smallest (in terms of position and diameter) corresponding to greater agreement with the expert.

At first glance these issues could be regarded as specific to the crater counting case, but similar issues could also exist when considering any VCS project involving the detection of features through the presentation of on-screen imagery. Through implementing a laboratory-based version of the platform, both these issues and others more specific to the task and science case involved can be identified and accounted for at an early stage before full release to the public.

8.8 CONCLUSIONS

Using a laboratory study to test the effect of manipulating task workflow design factors with regards to a specific VCS platform, it was found that factors including autonomy, variety, task type and judgement type had an effect on both the volunteers' preference and their data output.

Which task workflow design configurations do volunteers prefer?

Through self-reported scores and associated free-text responses adding context, participants indicated a preference for interfaces that involve greater variety and autonomy, supporting postulate P7 (see 4.4.3). 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 (postulate P8, Lasecki et al., 2014) – with participants preferring to control when it occurs.

Thesis RQ_3: Does task workflow design influence volunteer results?

The analysis of crater marking behaviour has also indicated that manipulating certain task workflow design factors affects performance, dependent on the type of performance measure considered. It was found

that the interface involving a simpler task, less variety and less autonomy (Ramped Position) resulted in greater coverage (Eveleigh et al., 2014) and greater efficiency (supporting postulate P10). There is a caveat however, as while 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 (Stepped and Ramped (Position) for crater position, Full and Ramped (Mark) for crater diameter) supporting postulate P11.

What is the relationship between preference and performance effects?

When considering if volunteer preference improves performance, crater marking data does not support postulate P5 stating that greater autonomy (the volunteers preferred interface) 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. This reveals a tension when considering task workflow design, where a possible trade-off might exist between choosing an interface that increases volunteer participation, or one that improves performance (either data accuracy or quantity).

Overall, the results of this study support the findings of previous related human factors 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. Directly relating this result to the overall thesis research question RQ_2 (section 4.5.4), this again suggests that TWD factors can effect volunteer motivation and therefore their behaviour in terms of visiting the site. However, when considering research question RQ_3 regarding volunteer performance the picture is more complex, with different task workflow design factors affecting measures such as coverage, participant agreement and expert agreement in different ways. VCS developers and science teams may well have to consider the type of data they need, 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.

8.9 SUMMARY AND OUTLOOK

Building on the previous studies in this thesis, this study supports the findings of chapter 7 that volunteers are aware of task workflow design factors such as autonomy and variety, and have preferences regarding their design that could influence their motivation to take part. Additionally, TWD factors have been shown to affect the data produced by volunteers, in both a positive and negative way depending on the performance measures considered. These relationships do not necessarily correlate, with task workflow design manipulations preferred by the participant not always resulting in a benefit to performance.

Methodologically however, while the study combines the collation of participant opinions with measures of performance in evaluating the manipulation of task workflow design, it does not measure the effect of TWD on volunteer behaviour in terms of website engagement. The study demonstrates that volunteer preference does not necessarily translate to better performance, and therefore similarly it cannot be presumed that positive behavioural outcomes in terms of the website analytic measures described in chapter 6 will result. A second limitation of the study is that the participants involved do not necessarily represent a typical citizen science community, both in terms of previous VCS experience and demography.

The final study of this thesis in chapter 9 addresses these issues, by combining the methods of the previous studies within the scenario of a 'live' experiment.

9

TASK WORKFLOW DESIGN 'IN THE WILD'

This chapter presents the final empirical study of the thesis, involving a 'live' release of the Planet Four: Craters platform to a real world volunteer community. It investigates the effects of manipulating task workflow design upon volunteer motivation and engagement through self-reported indicators and behavioural measures in terms of website analytics. It also considers the effect on performance through comparison with expert analysis. Although there was no significant difference in participants' questionnaire scores, free-text responses suggested a preference for interfaces that allowed greater autonomy and involved more task variety. This position is supported when considering website behaviour with participants using the Full interface (most autonomy and variety) returning more frequently, although the amount of time spent per visit was unaffected by the interface used. However, analysis of participant performance suggested that how this time was used did differ. The interface involving the least complex task (Ramped Position) 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 judgement, 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.

9.1 INTRODUCTION

Although there exists a growing body of research that has studied existing VCS projects beyond their original scientific goals, exploring the motivations and contribution of volunteers, to our knowledge none has

directly investigated the manipulation of design factors and its effect when considering a live citizen science platform. Existing research has instead compared the design (Reed et al., 2012; Prestopnik & Crowston, 2012) and volunteer motivations (Reed et al., 2013) of different VCS projects concerning different scientific disciplines and goals in a retrospective manner.

The first study of this thesis (chapter 6) took a similar approach, comparing the task workflow design and volunteer behaviour of several existing Zooniverse projects. The study in chapter 7 built on this by focussing in on one platform (Planet Four), and considered volunteers' opinions regarding task workflow design. While these two studies have shown that TWD factors can have an effect on volunteer behaviour and opinion, the experiment in chapter 8 furthered this to investigate their effect on volunteer data and performance. Through manipulating the interface design of a purpose-built platform as part of a hypothesis-driven experiment, it was found that task workflow design also affects the data produced by volunteers.

The final empirical study presented in this chapter combines the methodologies used in the previous three studies. It investigates the effect that the different interface designs explained in chapter 8 have on participant behaviour, experience and performance when incorporated into a 'live' platform with an existing volunteer community. In doing this, the experiment explores whether task workflow design factors can directly influence such measures of scientific output in a 'real world' setting. Specifically, it investigates not only if the relationship between task workflow design and volunteer experience and performance exists as found in the studies of chapters 7 and 8, but extends this to consider if any perceived volunteer preference in interface design translates to an effect on volunteer website engagement.

9.2 IMAGE PROCESSING AND PREPARATION

As with the previous study explained in chapter 8, the CTX image G05_020119_1895_XN_09N198W (and the 78 slices it has been converted to) has been used as a stimulus for the experiment in this chapter. While the predominant theme of this thesis is to investigate the effect of task

workflow design factors, there is still a science case to address as derived in section 8.2 and as such the same region of the Martian surface was the focus for the 'live' experiment. Also by using the same image comparisons can again be made between the volunteer and expert data, as well as between participants of the controlled laboratory study and a more representative existing citizen science community.

However, while the laboratory study in the previous chapter involved 30 participants, a 'live' study involving the community of an existing platform has the potential to attract hundreds or even thousands of volunteers to take part, raising some ethical issues that need to be addressed. For instance, as suggested in 8.6, this thesis' crater marking task requires a maximum of 8 volunteers to view each image for the crater markings to be seen as a definite true positive. It could be argued therefore that as the number of volunteers that classify an image increases beyond this point, this constitutes an inefficient use of volunteer time and effort with at least some of their markings contributing to an already certain classification. This is an especially prevalent point considering the self-reported views that participants shared regarding the importance of not having their time wasted, as reported in the previous chapter (see 8.7.3).

In order to address this issue, a second CTX image was used in this study again located near the heart of the Cerberus Fossae region of Mars, known as G23_027332_1907_XN_10N202W (see figure 9.1).



Figure 9.1: CTX image G23_027332_1907_XN_10N202W (left - top section, right - bottom section)

After completing the processing chain as detailed in section 8.3, this second image was 'sliced' as before into 122 smaller chunks. Thus, including the 78 image slices from the previous study, a total of 200 image slices were displayed to the 'live' community for the purpose of this study.

9.3 RESEARCH QUESTIONS AND HYPOTHESIS

Task workflow design factors are studied in a 'real world' setting in the context of a 'live' VCS platform with an existing volunteer community. As with the previous chapter the design-influenced factors of autonomy, variety, task type, and user judgement (independent variables 1-4) were operationalized through the implementation of three separate platform interfaces. Again, as with chapter 8, this experiment attempts to answer the question:

RQ: Which task workflow design configurations do volunteers prefer?

As well as in relation to the research questions introduced in 4.5.4:

RQ_3: Does task workflow design influence volunteer results?

And

RQ_2: Does task workflow design influence volunteer behaviour?

In doing this, the experiment again underpins the results found in the studies of chapters 6 and 7 regarding the opinions of volunteers, whilst also investigating if the effects of task workflow design on the scientific data collected found in chapter 8 are consistent when considering a 'real world' scenario. Additionally, by setting the experiment in a 'real world' context using an established volunteer community, it builds on earlier findings by considering the effect of task workflow design on volunteer behaviour in terms of platform engagement. In addressing these questions regarding the effect of task workflow design, the study again answers the sub-question:

RQ_a: What is the relationship between these effects?

This is not only the two-way relationship between volunteer experience and performance (i.e. does a more satisfying volunteer experience correspond to great data accuracy compared to the expert?) as with chapter 8, but includes a third factor, volunteer behaviour. For example, does a preference for a certain interface correspond to volunteers spending longer on the site per visit or returning more often? Furthermore, how does volunteer behaviour in terms of platform engagement influence the data they produce?

Based on the design-influenced postulates described in section 4.4.3 (P5-11), testable hypotheses were again formulated for a mix of qualitative and quantitative dependant measures regarding volunteer experience and performance:

H1: Volunteers using an interface involving more task variety have more motivation.

H2: Volunteers performing a less 'involved' task and judgement produce greater data coverage.

H3: The type of task performed and judgement made by the volunteer influences volunteer agreement (in terms of the data produced).

Further hypotheses are also formed based on the findings of chapter 8, and the findings regarding volunteer behaviour in chapter 6:

H4: Volunteers using an interface involving more autonomy have more motivation.

H5: Volunteers using an interface involving more task variety and complexity are more likely to return to the platform.

H6: Volunteers using an interface involving more task variety will spend less time classifying per platform visit.

9.4 EXPERIMENT DESIGN

Unlike the laboratory-situated study described in the previous chapter, this 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 judgement* were manipulated. The three classification interface designs described in chapter 8 were again employed, adapted in collaboration with the Zooniverse development team to run online through the Planet Four landing page. The questionnaire including NASA-TLX type statements used in chapter 8 was converted into an online version (using the Qualtrics platform as in chapter 7) in order to assess volunteer opinion and perceived workload. Additionally, volunteers were able to leave 'free-response' messages and create threads on Planet Four's 'Talk' discussion board; an application common to all promoted Zooniverse VCS platforms allowing volunteers to communicate with both the scientific team and other volunteers. Figure 9.2 gives an overview of the experimental design in relation to the model of task workflow design factors derived in section 4.3 and the measures of scientific outcome described in section 4.5.3. The dependant variables shown in red indicate those that build on the analysis of chapter 8 through using a 'live' platform.

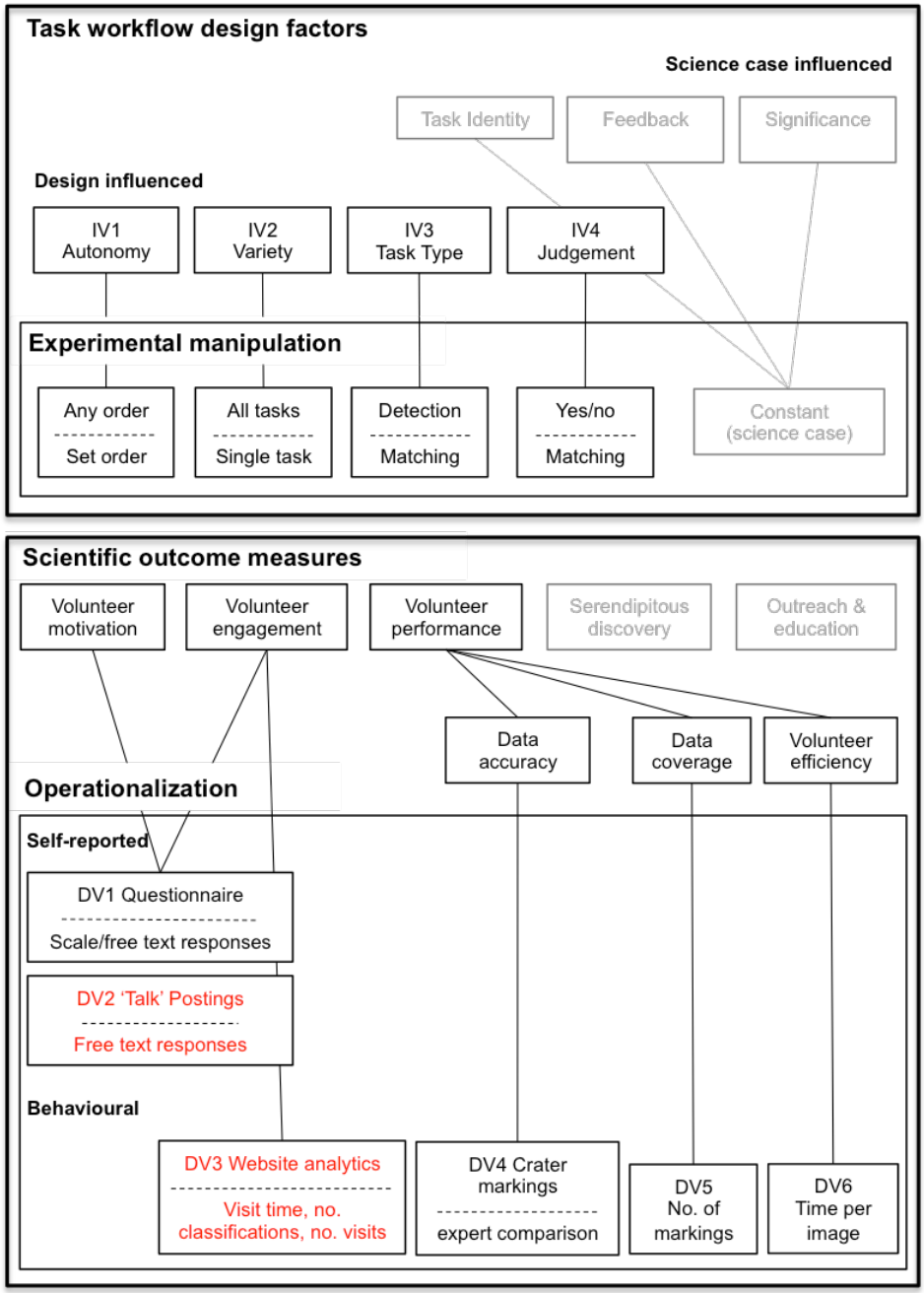


Figure 9.2: Overview of the experimental design in relation to the task workflow design model

The four design-influenced TWD factors (IV1-4) have been experimentally manipulated through the design of the three VCS classification interfaces (see 8.5.2), and the impact of this manipulation on the operationalisation of scientific outcome is then measured. This is done by participants' self-reports giving their opinions on motivation, engagement and experience (DV1 & 2), behavioural measures of engagement through website analytics (DV3), and by measuring performance through participant-expert marking comparison (DV4), the number of markings made (DV5) and the time spent

classifying each image (DV6). As with the laboratory study in chapter 8, other science case influenced TWD factors are held constant by experimental design, i.e. by addressing the same science case across each condition/interface (see 8.2).

9.4.1 Interface & Platform Design

As the 'live' study of this chapter addressed the same crater-marking focussed science case of chapter 8, the tools made available to the participants were identical, and are described in section 8.5.1. Likewise, the four independent variables highlighted in figure 9.2 were manipulated through the same three interfaces presenting the tools differently as described in 8.5.2. Participants accessed the site through a URL link that was displayed on the original Planet Four's home page, and disseminated through various email shots and Planet Four blog posts[26]. On visiting the URL (craters.planetfour.org) participants were presented with the platform home page, from where they could access the classification interfaces and also other 'general' pages as detailed in 8.5.3. Figure 9.3 shows the Planet Four: Craters home page with the links to the other pages described.

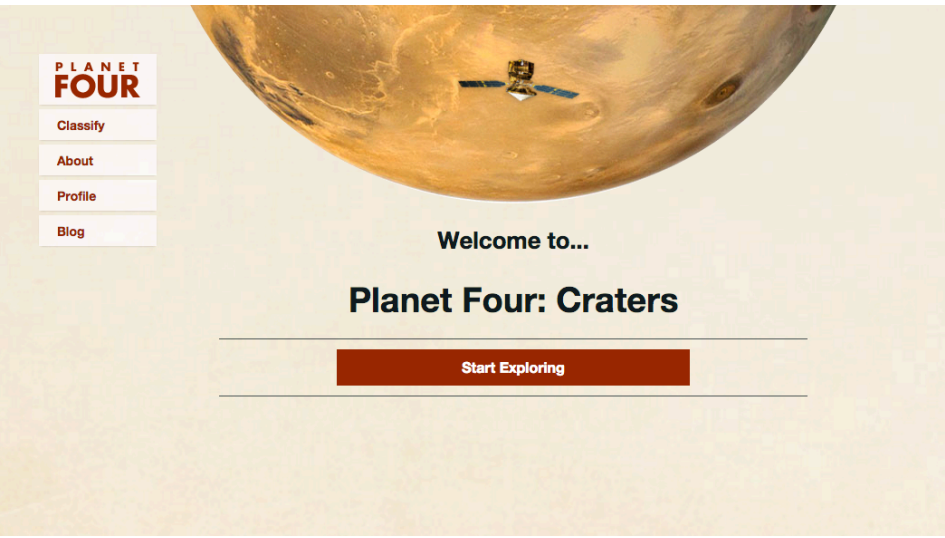


Figure 9.3: Planet Four: Craters home page

[26] An example of a blog post advertising the site can be found in Appendix D of this thesis

9.4.2 Questionnaire & Online 'Talk' Forum

In order to collate participant views and opinions, participants had the opportunity to complete an online questionnaire[27] accessible through a link on the classification interface. The design of the questionnaire followed that of the one used in chapter 7, using the same Qualtrics software. It again contained both Likert-type statements and allowed 'free-text' responses under the categories *design & usability, tasks & tools* and *imagery* as described in section 7.3.1. Prior to completing these sections, participants completed a general demographics section regarding their background and VCS experience.

One of the communication tools included on the study site was the Zooniverse's 'Talk' forum. Launched in 2010, 'Talk' is an online data collection and discussion tool primarily designed to allow volunteers to communicate both amongst themselves and with project researchers (Fortson & Lynn, 2014). This ability to communicate has been identified as a critical engagement factor for VCS systems and as such volunteer interaction with such tools and its correlation with the activity on the primary task has been studied previously (Fortson et al., 2012; Reed et al., 2013). Although this research direction has not been explored for this thesis, the Talk tool has been used to allow participants to communicate when taking part. A special 'Planet Four: Craters' thread[28] was created on the original Planet Four Talk page, and linked to through the Classifying interfaces, that allowed participants to share their views, opinions and concerns regarding anything about the site either in conjunction with or instead of completing the questionnaire.

9.4.3 Participants

The Planet Four: Craters site went live and was accessible through the link described on 26th March 2015. From this time until 26th June 2015, 606

[27] An example of the Planet Four: Craters online questionnaire can be found in Appendix D of this thesis

[28] The Planet Four: Craters Talk thread can be found at: http://talk.planetfour.org/#/boards/BPF000000b

registered Zooniverse volunteers (those who have provided a username and password) visited the site and classified at least one image. A further 13242 classifications (~50.2%) were made by unregistered volunteers. Although it is not possible to know exactly how many volunteers contributed to this figure, previous analysis carried out on other Zooniverse sites (Swanson et al., 2015) suggests that unregistered volunteers can make up as much as two thirds of the total. This being the case, it is estimated that approximately 1500-2000 volunteers took part in the study. Of these, 63 completed the online questionnaire and 61 left posts on the Planet Four: Craters Talk thread.

9.4.4 Procedure

The experiment was carried out under supervision at the Nottingham Geospatial Institute, University of Nottingham. The author had full responsibility in agreement with the supervision team to perform all activities necessary to conduct the experiment. This included experimental design, orchestration, set-up, support and communication. Additionally, the technical infrastructure that supported the experiment was designed and implemented by the author. This was done with the Zooniverse development team, and also in agreement with the original Planet Four platform science team through which the Planet Four: Craters site was promoted and accessed.

When accessing the study URL (craters.planetfour.org), whether through the link on the original Planet Four site or through links provided on email and blog posts, participants were taken to the platform homepage as shown in figure 9.3. From here, they could access the general information pages outlined in section 8.5.3 regarding the background of the scientific case being addressed and examples of the types of features they will be marking. On selecting the 'classify' link, participants were taken to one of the three classifying interfaces (full, stepped or ramped – figure 8.6) on which the image slices were presented and the tools for marking craters made available. The interface displayed was selected at random, in a non-weighted way so that the probability of seeing each was equal. Once presented, the classification interface remained constant for the duration of the visit and for each classification, until the participant closed down the

'classify' page. If the participant subsequently chose to return to the platform to make further classifications, the process 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 (see figure 8.6). As with the laboratory study in chapter 8, the marking and behavioural data collected regarding the example image was separated from the data analysis discussed later in this chapter. The order in which image slices were displayed to the participant was randomised, again in an un-weighted way such that each slice was seen a similar number of times during the study. This was for the same reasons explained in section 8.5.5 regarding image content, learning effects and fatigue which could, if unaccounted for, bias the data collected. At any point when using the classifying interface, the participant could access a link to the online questionnaire, or if they are a registered volunteer the Talk forum, in order to share their views / opinions of the platform. This could also be done at any other time during the life of the study by accessing the relevant URL through the normal channels (search engine, web address bar, bookmark etc.).

9.5 RESULTS AND ANALYSIS: SELF-REPORTED

As outlined in figure 9.2, dependant variable measures have been recorded both through participant self-reporting and more behaviourally through crater marking performance and website analytics. The following section presents the analysis and results regarding self-reported data in terms of their relation to the independent variables regarding task workflow design.

9.5.1 Online Questionnaire: Likert-Type Responses

63 participants that visited the Planet Four: Craters platform subsequently completed the online questionnaire, 46 of which were male and 10 female (7 participants declined to answer). Participants' age ranged from 14 to 79 years old (mean = 49, median = 49). Regarding experience with the Zooniverse, 44 participants (70%) had taken part in sites other than Planet

Four and as with the study in chapter 8 they were predominantly of a similar space theme (Galaxy Zoo – 34, Planet Hunters – 28, Moon Zoo – 24). Most had only taken part in one other project, however this number increases to a dozen in some cases. Figure 9.4 illustrates the participation profile of those who completed the questionnaire.

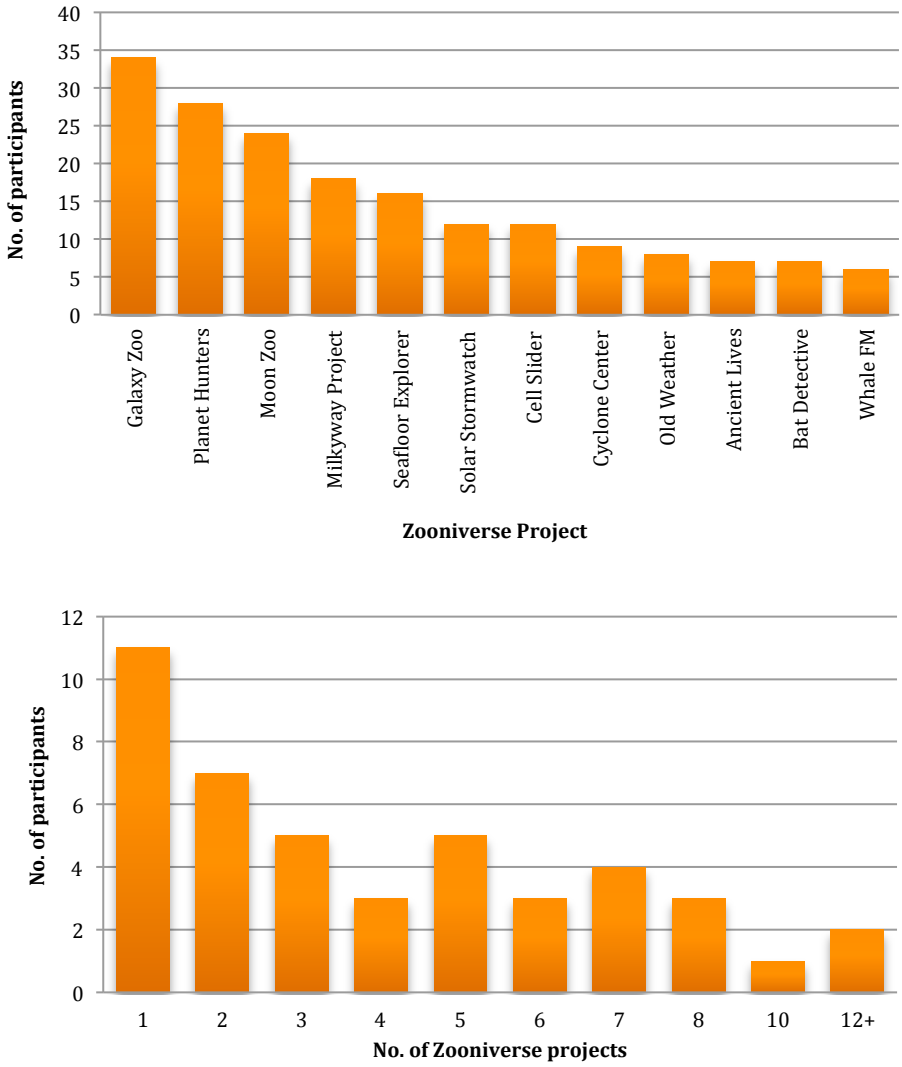


Figure 9.4: Participation in other projects - by site (top) and number (bottom)

Participant Likert responses to a number of statements included in the *design & usability, tasks & tools* and *imagery* sections of the online questionnaire showed no significant difference between each interface. This is perhaps as expected as many of the design features, tasks performed and tools used along with the images presented were constant, as with the laboratory study in chapter 8.

When considering the visual attractiveness of the site, participant responses tended to be positive, with over 80% of all raw scores across each interface lying between 7 and 9 (very attractive). This trend continued when considering the ease of use of the site, with over 75% of all raw scores across each interface lying between 7 and 9 (strongly agree) when indicating agreement with the statements *'I found the website easy to use and access'* and *'I learnt to use the website very quickly'.* Regarding the NASA TLX statements measuring perceived task workload, over 60% of all scores are found between 1 and 4 (low demand), with one-way ANOVA tests showing no significant differences existing between each interface (figure 9.5).

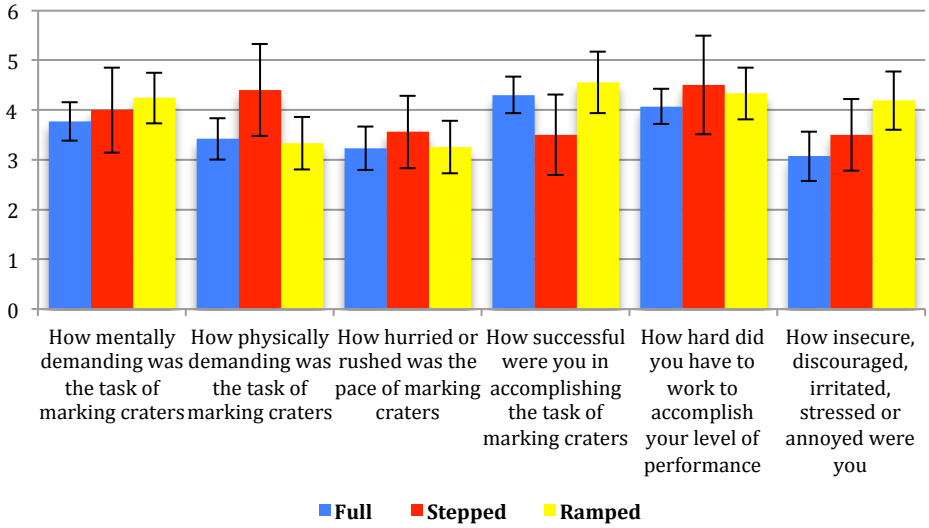


Figure 9.5: Participant responses to NASA TLX statements (with standard error shown)

In terms of responses to the statement *'I found the crater marking tool easy to use'*, again there is no significant difference in scores between each interface. Participant responses showed no clear preference, with the Full, Stepped and Ramped interfaces having mean scores of 5.93 ± 2.72, 6.10 ± 2.81 and 6.35 ± 2.64 respectively (1 being strongly disagree and 9 being strongly agree). This is also true when considering the elliptical crater tool, with mean scores of 5.88 ± 2.25, 6.10 ± 2.28 and 5.63 ± 2.59 respectively. Regarding the statement *'I would like more tools to mark different shapes and features'* again there is no significant difference in participant responses between each interface, with mean scores of 6.23 ± 2.52, 6.40 ± 2.99 and 5.90 ± 2.71 for the Full, Stepped and Ramped interfaces respectively.

Participant responses to statements regarding the imagery presented on the platform again showed no significant difference between interfaces, according to one-way ANOVA statistical testing (figure 9.6). Responses tended to be positive across each interface when indicating agreement with the statements *'I found it hard to pick out craters and features'* and '*I didn't know what to do if an image had no craters or features to mark',* with over 70% of responses lying between 1 and 3 indicating strong disagreement.

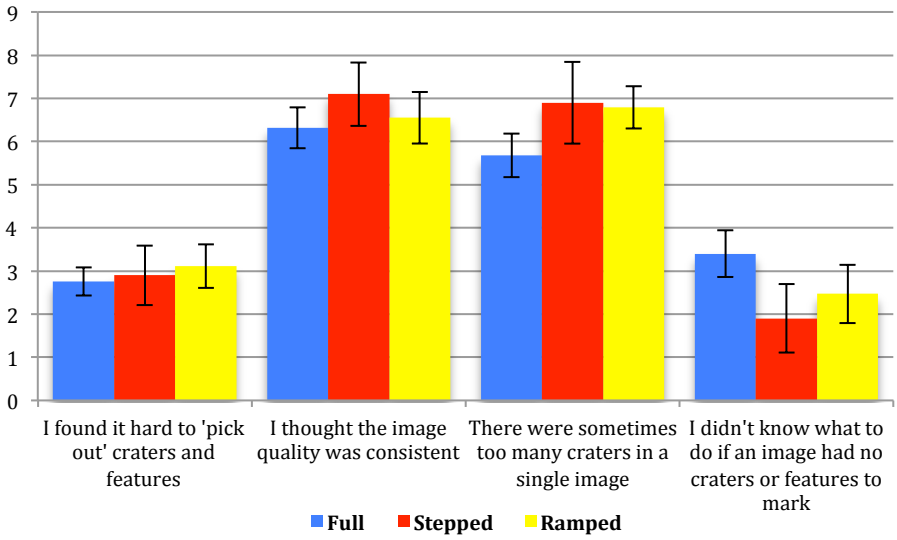


Figure 9.6: Participant responses to statements regarding platform imagery (with standard error shown)

Regarding the statement *'I thought the image quality was consistent'*, mean scores of 6.32 ± 2.38, 7.10 ± 2.33 and 6.55 ± 2.67 were given by participants for the Full, Stepped and Ramped interfaces respectively and showed no significant difference. Finally, mean participant responses to the statement *'there were sometimes too many craters in a single image'* were 5.68 ± 2.55, 6.90 ± 2.99 and 6.79 ± 2.12 for the Full, Stepped and Ramped interfaces respectively, again with no significant difference apparent.

9.5.2 Questionnaire Comments and 'Talk' Forum Posts

As with the laboratory study explained in chapter 8 participants also had the opportunity to provide 'free response' replies, both through the online questionnaire to add context to their scores and through the associated platform 'Talk' forum. These responses again have been collated and subsequently coded under 6 categories, the four defined in chapter 8: *usability, accuracy, tool issues* and *imagery*, and two additional categories:

technical issues and *science enquiries.* Technical issues comments were concerned with problems in accessing the site and general instances of 'crashing' and Science enquiries comments were requests for more information about the science involved regarding types of crater and the Martian surface. The existence of these two new categories reflect the 'live' nature of the study; participants taking part online using their own equipment and internet connection increasing the possibility of technical failure, and the increased interest in the science behind the tasks and imagery involved acting as the extrinsic motivation in visiting the platform in the first instance. Table 9.1 shows a breakdown of the number of comments made by participants regarding each interface, sorted into the 6 different categories.

Table 9.1: Number of responses by category and interface

Topic	Full Interface	Stepped Interface	Ramped Interface
Usability	3	10	11
Accuracy	7	4	5
Tool Issues	12	2	3
Imagery	2	2	2
Technical issues	2	4	5
Science enquiries	1	2	3
Total:	27	24	29

Comments in the *usability* category involved two themes: most concerned the autonomy allowed by each interface, in terms of the tools available and order of use (participant R51 & R42 respectively):

> *"I would have preferred all the tasks available at the same time…"*
>
> *"I do ***not*** like the Ramped interface, It feels like I am working with fragments. It also gives me the (real) feeling that I am being condescended…"*

A few others were positive feedback regarding the ability to choose when to stop (S73):

> *"This actually is what I'm looking for – namely, one in which I can stop after merely indicating that craters are present, or proceed to marking positions, or if time permits, move onto marking size…"*

Comments in the *accuracy* category again highlighted issues when marking smaller craters, and the difficulties that arise (S18):

> *"Some of the images have a lot of tiny craters, which are hard to mark for size and [it] takes a long time to capture all of them."*

The 'smaller craters' issue is again raised by comments in the *tool issues* category, this time relating the problems to tool use and design (participant F2):

> *"There were a lot of very small craters that I found hard to select with the standard tools, due to the handles covering them up…"*

Other comments in this category indicated the benefit of additional tools, either to address the problem of too many craters in an image (R43):

> *"It might be useful to have an option for 'too many to count' for things like craters…"*

Or to allow for the collection of additional, descriptive, scientific data (F73):

> *"Would it be useful to note the relative 'sharpness' (S for Sharp, D for Diffuse) of crater edges to help determine relative amounts of weathering thus age?"*

Comments in the *imagery* category either concerned the content of the imagery, relating to the number of craters found in an image as mentioned previously (F73):

> *"My current image has over 100 craters, all sizes – but too many for me to count, what would you like us to do?"*

Or were requests for more information regarding the image format (F80):

> *"…I am new to this project, how do I know what scale the craters are being displayed in?"*

Finally, comments in the *technical issues* and *science enquiries* categories were independent of the interface they were using, being either about general accessibility / site failure or questions regarding the formation of craters and their typology (R73):

> *"Does the absence of craters in the rift valley indicate that the valley formed after the period of heavy bombardment?"*

Any technical issues raised by participants were passed on directly to the Zooniverse development team in Chicago so that any solutions could be implemented, whilst responses to scientific queries were provided by the author, either through personal knowledge or after consultation with the science team at the University of Bristol.

9.6 RESULTS AND ANALYSIS: BEHAVIOURAL

As outlined in figure 9.2, dependant variable measures have also been recorded behaviourally through crater marking performance and website analytics. The following section presents the analysis and results regarding behavioural data in terms of their relation to the independent variables regarding task workflow design.

9.6.1 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 ANOVA showed no significant difference (F (2,670), p = .577), with the Full, Stepped and Ramped interfaces having mean times of 20m40s ± 3m14s, 20m48s ± 3m50s and 17m11s ± 1m41s respectively (figure 9.7). However, when considering the number of images classified per visit a Kruskal-Wallis test showed that there was a statistically significant difference between each interface (X^2 (2) = 81.75, p = .001). A Bonferroni post-hoc test revealed that the number of images classified when participants used the Ramped interface (25 ± 27) was significantly greater than when using the Full and Stepped interface (18 ± 26, p = .001 and 10 ± 16, p = .001 respectively). The number of images classified using the Full interface was also significantly greater than when using the stepped (p = .001).

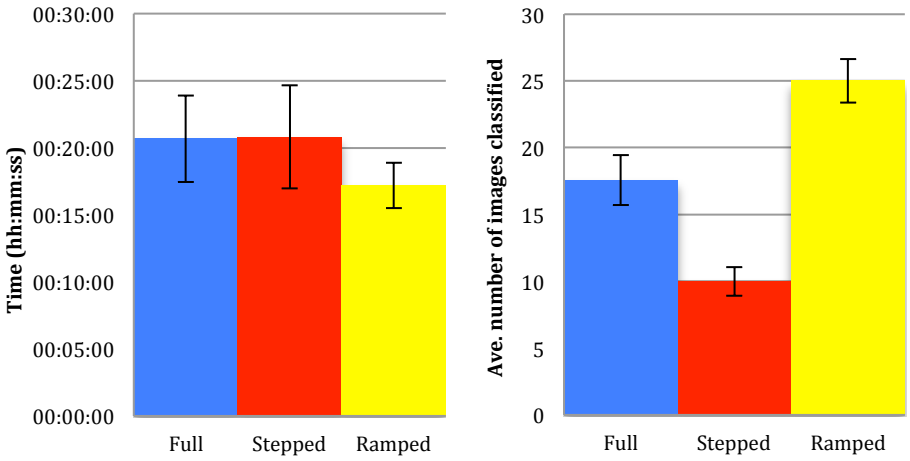


Figure 9.7: Time spent on site and number of images classified (with standard error shown)

Finally when considering the number of occasions participants visited the site and classified at least one image, all three interfaces had a similar rate of 'non-returners' (those who visited only once and did not return) between 83-85%. However there were differences when considering repeat visitors and how often they returned (figure 9.8).

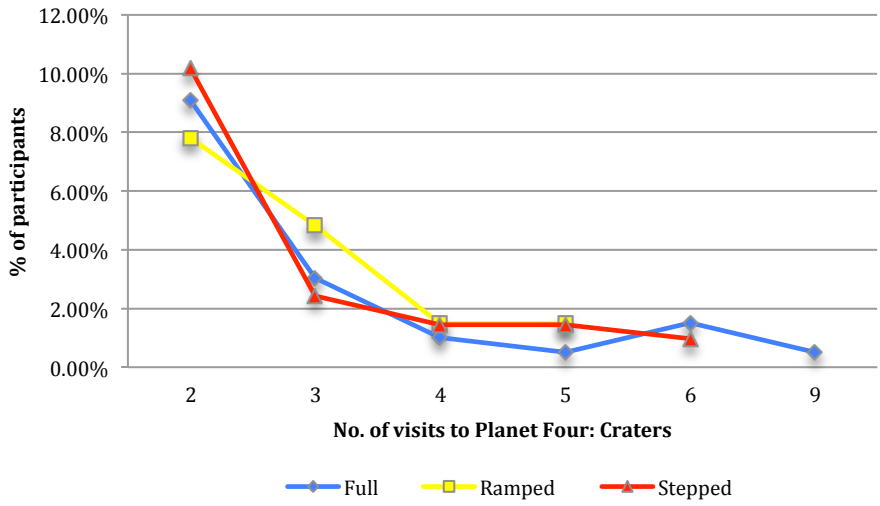


Figure 9.8: Number of return visits per interface

The maximum number of visits any volunteer made to the Ramped interface was 5 (1.5% of participants) over the three month period of the study, whilst 1% of participants using the Stepped interface visited on 6 separate occasions. The most number of visits made by any participant occurred using the Full interface, where 0.5% of participants visited 9 times over the three months.

9.6.2 Crater Marking Behaviour

As with the laboratory study in chapter 8, participant crater marking behaviour 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. Again 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 centre 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, a Friedman test showed a significant difference between each interface (X^2 (3) = 64.56, p = .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 ± 23.21%) compared to the Ramped (mark) and Stepped (36.25 ± 26.59%, Z = 7.068, p = .001 and 39.98 ± 21.09%, Z = 3.679, p = .001 respectively). This is also true when using the Ramped (position) interface (44.63 ± 25.14%, Z = 6.429, p = .001 and Z = 3.04, p = .014 respectively).

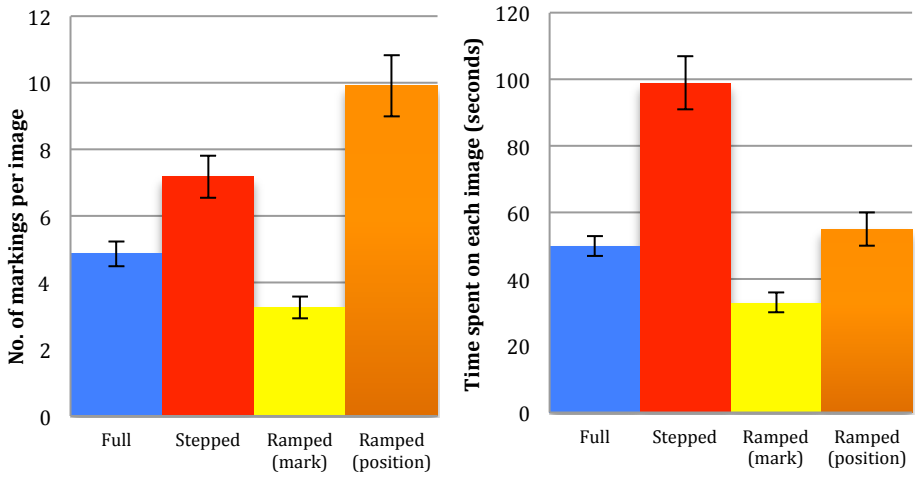


Figure 9.9: Crater marking results (with SE shown)

In terms of the number of crater markings per image, a repeated measures ANOVA showed a statistically significant difference between each interface ($F(1.895, 377.146) = 44.944$, $p = .001$). Post hoc tests using the Bonferroni correction revealed that the Ramped (position) interface resulted in a greater number of markings (9.92 ± 13.02) compared to the Stepped (7.19 ± 8.72, $p = .001$), Full (4.86 ± 5.21, $p = .001$) and Ramped (mark) interfaces (3.26 ± 4.68, $p = .001$). Similarly, the Stepped interface also resulted in a greater number of markings compared to both the Full and Ramped (mark) interfaces ($p = .001$). 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, $p = .001$). Participants spent more time per image using the Stepped interface compared to the Ramped interfaces (99 ± 90 seconds vs. 33 ± 57 seconds and 55 ± 68 seconds, $p = .001$) and more time compared to the Full (50 ± 39 seconds, $p = .001$). Participants significantly spent the least amount of time per image using the Ramped (mark) interface compared to the other three. Figure 9.9 shows the average number of crater markings and the average time spent on each image using each interface, along with standard error.

9.6.3 Participant Agreement

Table 9.2 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 centre position and the average standard deviation of the cluster crater diameter in terms of screen pixels (i.e. the level of agreement between participants). Clusters were identified through the same process described at the beginning of section 8.6.2.2, using ArcMap's *Grouping Analysis* functionality.

Table 9.2: Crater marking agreement for each interface

	Full Interface	**Stepped Interface**	**Ramped (Mark)**	**Ramped (Position)**
No. of crater clusters	379	212	83	496
SD of position	2.51 ± 0.16	2.68 ± 0.46	2.75 ± 0.75	2.10 ± 0.15
SD of diameter	6.62 ± 0.26	3.85 ± 0.41	5.60 ± 0.77	-

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 five-fold 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 less images were classified per participant on average has resulted in less 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. A Friedman test showed a significant difference between each interface (X^2 (3) = 15.042, p = .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 of 2.57 ± 5.75 pixels vs. 3.34 ± 7.22, p = .001). This is also true when compared to markings made using the Ramped (Mark) and Ramped (Position) interfaces, although the difference is not significant (3.08 ± 8.12 and 3.31 ± 9.30 pixels respectively). Regarding participant marking agreement in terms of crater diameter, again a Friedman test revealed a significant difference between each interface (X^2 (2) = 44.947, p = .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 were significantly less in agreement than those made using the Stepped and Ramped (Mark) interfaces (8.52 ± 9.36 pixels vs. 4.96 ± 10.64, p = .001 and 5.52 ± 7.95, p = .001 respectively). Figure 9.10 shows participant agreement in terms of position and size across each interface.

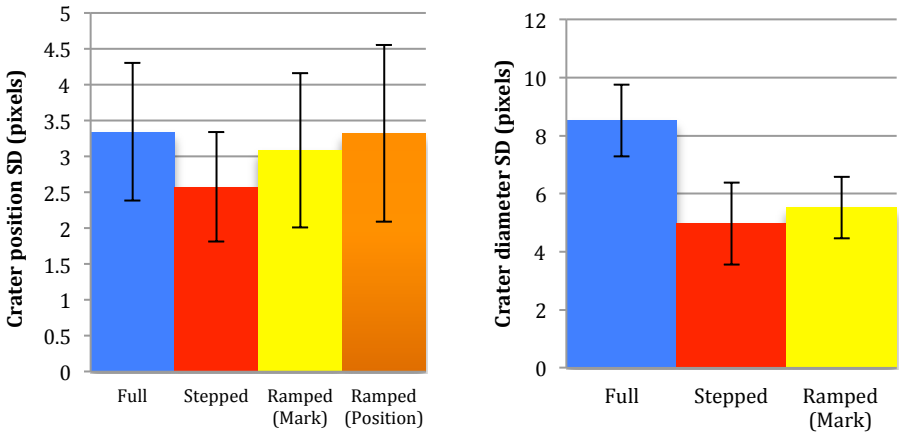


Figure 9.10: Comparison of marking agreement per interface

Breaking down participant agreement in terms of crater identification further, figure 9.11 shows the number of crater clusters marked verses 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.

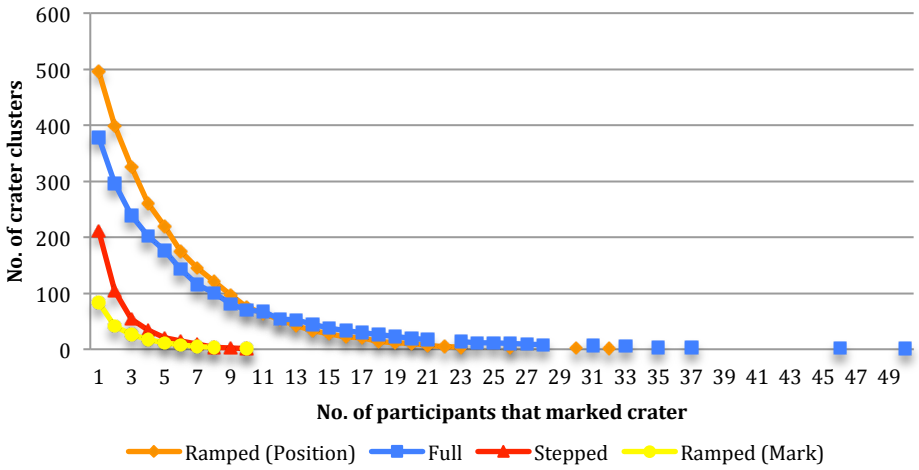


Figure 9.11: Participant crater identification agreement

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.

Finally, table 9.3 breaks down the number of crater clusters identified using each interface in terms of their size. As previous explained in the laboratory study (chapter 8), the crater counting analysis method for aging the surface of Mars is highly dependent on the size distribution of craters. The size 'bins' used have again been chosen to match those previously used by the expert team (J. Taylor et al., 2013).

Table 9.3: Number of crater clusters per size bin

Size Bin (Diameter)	Full Interface	Stepped Interface	Ramped (Mark)
>0-10 Pixels	106	68	32
>10-20 Pixels	219	137	37
>20-40 Pixels	43	6	12
>40 Pixels	11	1	2
Total:	379	212	83

The order in terms of the number of crater clusters identified by participants continues from Full with the most to Ramped (Mark) with the fewest when considering craters < 20 pixels in diameter. When considering craters > 20 pixels however, the Ramped (Mark) interface has identified more (14, ~17%) compared to the Stepped (7, ~3%). Furthermore, proportionally more of the clusters marked using the Ramped (Mark) interface were identified as the smallest size (> 0 – 10 pixels diameter) compared to the Full and Stepped interfaces (39% vs. 28% and 32% respectively).

9.6.4 Participant – Expert Comparison

As with the laboratory study explained in chapter 8, crater clusters identified by participants using each interface have been compared to those identified by experts from the University of Bristol (see section 5.3.2). Table 9.4 compares the expert benchmark with the crater marking clusters made using each interface in terms of crater identification.

Table 9.4: Crater identification compared to expert

Interface	True Positive	False Positive	False Negative
Full	268	111	97
Stepped	166	46	199
Ramped (Mark)	73	10	292
Ramped (Position)	303	193	62

Breaking down the crater clusters identified using the Full interface, 268 (~71%) clusters were confirmed as a crater by the expert, with 111 (~29%) misidentified. Regarding the expert crater markings, 97 (~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, ~78%), with 46 (~22%) misidentified. Due to the reduced number of clusters identified, more of the markings made by the expert have been missed (199, ~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, ~88%), with 10 (~12%) misidentified. Likewise however the reduced number of clusters identified overall has resulted in the most expert crater markings being missed by participants (292, 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, ~17%).

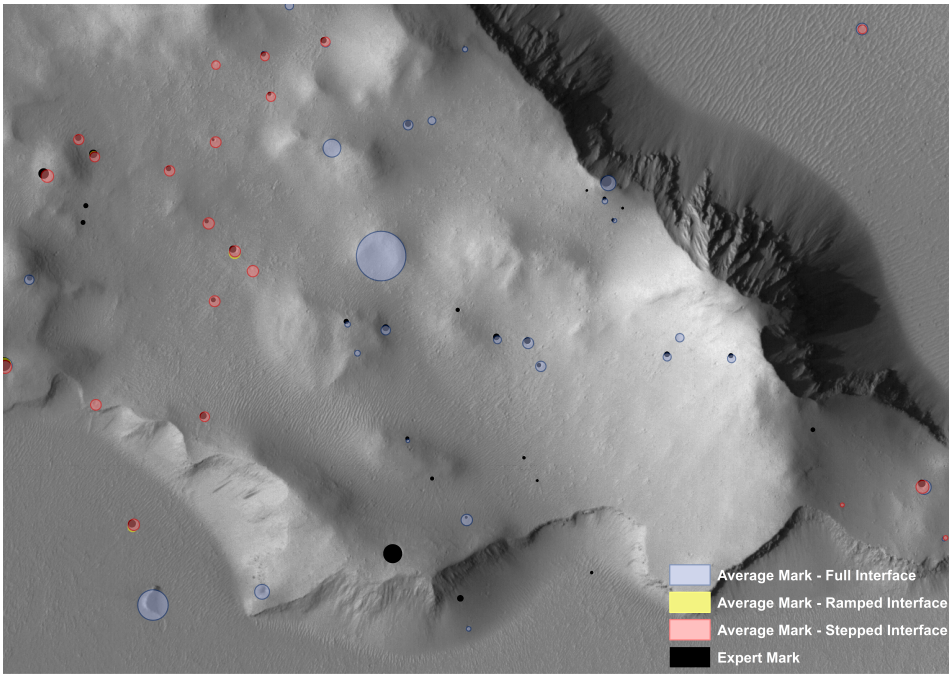


Figure 9.12: Expert and participant average crater markings per interface

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 9.12 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, a Friedman test showed a significant difference between each interface (X^2 (3) = 14.82, p = .002). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that the position of markings made using the Ramped (Position) interface were significantly further away from the expert equivalent than those made using the Full and Stepped (average difference of 4.70 ± 7.79 pixels vs. 4.18 ± 7.11, p = .005 and vs. 3.80 ± 3.93, p = .023 respectively). Markings made using the Ramped (Mark) interface (4.62 ± 8.25) were also further away position-wise from the expert than the Full and Stepped, though the difference was not significant.

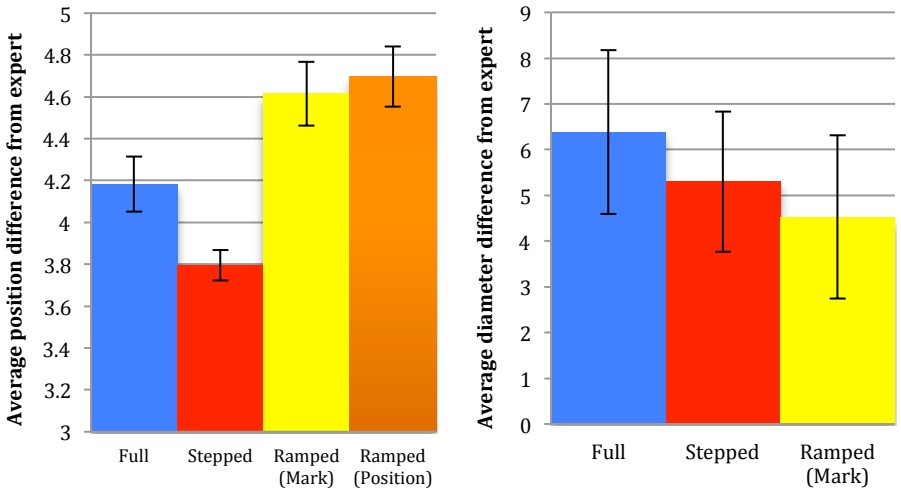


Figure 9.13: Position and diameter difference of participant markings compared to expert (with standard error shown)

Regarding the average difference between participant crater diameter and the expert equivalent, a repeated measure ANOVA with a Greenhouse-Geisser correction applied showed that there was a significant difference between each interface (F (1.858, 800.517) = 6.247, p = .003). Post-hoc tests using the Bonferroni correction 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 interface (difference of 4.53 ± 13.13 pixels vs. 6.39 ± 13.18, p = .001). Figure 9.13 shows the average difference in position and diameter between the expert markings and those made by participants using each interface.

As with the laboratory study described in chapter 8, participant crater identification has been further broken down by comparing the clusters identified using each interface with the expert data in terms of different size bins (see 8.6.2.2). Table 9.5 describes the participant-expert agreement as a function of the size of the crater.

Table 9.5: Participant crater marking sizes compared to expert

Size Bin (Diameter)		Full Interface	Stepped Interface	Ramped (Mark)
>0 – 10 Pixels	True positive:	66	37	29
	False positive:	149	97	11
	False negative:	88	169	263
>10 – 20 Pixels	True positive:	35	25	21
	False positive:	11	4	6
	False negative:	8	25	27
>20 – 40 Pixels	True positive:	4	1	3
	False positive:	0	0	0
	False negative:	1	4	2
>40 Pixels	True positive:	2	1	2
	False positive:	1	1	1
	False negative:	0	1	0

In terms of general crater identification, the greater number of clusters identified by participants using the Full interface again resulted in fewer of the expert markings being missed across each of the different size bins (97 in total). The fewest number of clusters identified using the Ramped (Mark) interface has also again resulted in the highest amount of expert markings being missed when considering smaller craters (< 20 pixels diameter, 290), however when considering larger craters (> 20 pixels) fewer of the expert markings were missed compared to participants using the Stepped interface (2 vs. 5). In terms of marking the size of craters, although participants using the Full and Stepped interfaces identified a greater number of small craters (> 0 – 10 pixels diameter), proportionally more of them were falsely marked as larger craters (149, ~69% and 97, ~72% respectively) compared to the Ramped (Mark) interface (11, ~28%). Participant performance was comparable between the Full and Ramped (Mark) interfaces when considering craters of diameter > 10 – 20 pixels, and an improvement on the smaller size with ~23% misidentified in terms of size. This improvement however is even more evident when considering the Stepped interface, where only ~14% were wrongly identified.

When considering larger craters (> 20 pixels in diameter), performance in terms of crater size was comparable across each of the interfaces, with all craters correctly identified also being marked within the correct size bin

for diameters of > 20 – 40 pixels, and only one crater incorrectly marked as smaller on each interface when considering sizes > 40 pixels.

Finally, figure 9.14 again considers the number of participants that contributed to a cluster verses 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.

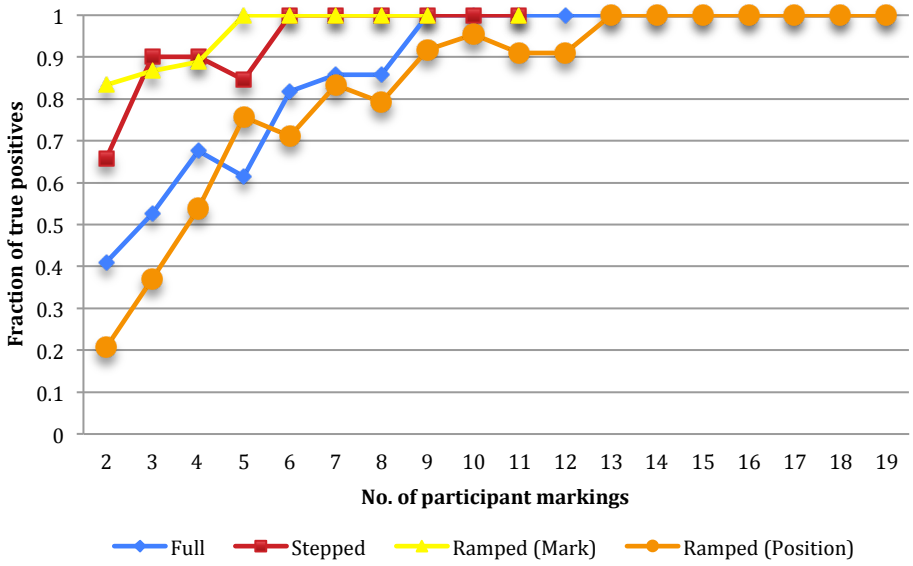


Figure 9.14: Cluster participant contribution verses fraction of true positives compared to expert

Clusters identified by participants using the Ramped (Mark) interface required the least amount of markings, with all clusters made up of 5 or more participant contributions also recognised as a crater by the expert. This was followed closely by those made on the Stepped interface, where 6 or more participant markings were needed. The figure is somewhat higher for clusters made on the Full interface, with those made up of 9 or more participant contributions all recognised 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 2 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 verses 0.41 and 0.21 respectively).

9.6.5 Summary and Hypotheses

- Hypothesis H1 (greater variety = more motivation) is partially supported by the findings of the online questionnaire & Talk forum. Although in terms of the quantitative Likert-type responses there was no significant difference, participant free-text responses raised issues regarding the need for more tools to enable different tasks to be attempted.
- Hypothesis H2 (simpler task = greater data coverage) is supported by the finding that more crater clusters were identified and marked using the Ramped (Position) interface.
- Hypothesis H3 (task type and judgement 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), and that position 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.
- Hypothesis H4 (greater autonomy = more motivation) is partially supported by the findings of the online questionnaire and response on the Talk forum. Participant free-text responses raised issues concerning feeling 'restricted' and 'condescended' when forced to perform tasks in a set order using the Stepped interface.
- Hypothesis H5 (task variety = more return visits) is supported by the finding that participants visited up to a maximum of 9 different occasions using the Full (most variety) interface compared to 5 using the Ramped (least variety) interface.
- Hypothesis H6 (task variety = less time per visit) is not supported by the findings, with no statistically significant difference in time per visit existing between each of the interfaces.

In summary, as with the findings of the laboratory study (chapter 8) the hypotheses related to task type and judgements 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 both the self-reported and behavioural findings of this study.

9.7 DISCUSSION

In this section the self-reported quantitative and qualitative findings of the online questionnaire are combined with the quantitative measures of website behaviour, participant agreement and performance compared to the expert to paint a broader picture of how task workflow design factors can effect a VCS platform's output and the motivation of its community.

9.7.1 The Effect of TWD on Participant Engagement and Motivation

Participant responses to the NASA-TLX statements regarding perceived workload of the crater-marking task showed no statistically significant difference across each interface. This was also the case regarding the 'ease of use' of the tools and the imagery presented. This is again unsurprising as the inherent task of identifying and marking a crater and the imagery presented has remained constant similarly to the laboratory study of chapter 8. There is also no significant difference in participant responses to the remaining Likert-like statements found in the *design & usability, tasks & tools* and *imagery* sections of the questionnaire across each interface.

Participant free-text responses provided either through the online questionnaire or posted on the associated Planet Four: Craters Talk forum however do reveal certain participant issues and concerns, many of which can be associated with the task workflow design factors of each interface. Comments grouped under the *usability* category were predominantly focussed on the Stepped and Ramped interfaces, and alluded to a feeling of being 'restricted' (S59):

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

As with the findings regarding the original Planet Four site in chapter 7 and the feedback from participants of the laboratory study in chapter 8, this further reinforces the importance of autonomy on volunteer satisfaction

(Chung-Yan, 2010) and its potential bearing on volunteers' intrinsic motivation to take part.

Although comments in the *accuracy* category were predominantly about the difficulties of marking small craters and were constant across each interface, a number of comments alluded to the issue of too many craters in an image. This concern was also raised by a number of comments in the *imagery* category, for instance (R73):

> *"My current image must have over 100 craters, all sizes - but too many for me to count…"*

Comments grouped in the *tool issues* category often suggest the development of extra tools that allow the volunteer to do additional, and often more involved tasks (participant F66):

> *"A feature like "Be a Martian" by NASA where one can write a short note of what seems to be interesting in the pic. Just a box to write notes in…"*

Both examples demonstrate an agreement with the findings of the previous two chapters, regarding the importance of variety when considering the volunteers' motivation to take part (Dubinsky & Skinner, 1984). The frequency of comments relating to these issues suggest that variety is at least as important as autonomy when considering VCS volunteer motivation and engagement.

9.7.2 Time Gentlemen, Please!

Website analyses regarding participant behaviour indicated that varying the design-influenced task workflow design factors of autonomy, variety, task type and judgement 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, whether they be the more science-influenced factors outlined in 4.3 or other external environmental factors. In terms of the broader aims of the thesis this finding underpins the claim made in 8.7.3 that '*volunteers' time is precious*', and 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 utilise an interface involving a simpler task akin to the Ramped (Position) interface (1 identification, 1 matching – 1 mouse click), resulting in more participants marking at least one crater per image, more crater clusters identified and more images classified per visit (Hypothesis H2, Eveleigh et al., 2014). However although this greater quantity of volunteer analysis results in fewer features being missed, it also increases the number of false positives (crater markings that do not represent an actual crater) that will have to be dealt with. Additionally, as explained in 8.7.2, the results collected from a simpler task are less detailed, returning only position and crater numbers rather than other metrics such as diameter. 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 and with the expert judgement, and although the total number of clusters identified is lower, the number classified per image is comparable to the simpler interface.

The issue with both these approaches is that they either restrict the volunteers' autonomy, or the variety of tasks available to undertake. Both of these factors influence volunteer motivation (hypotheses H1 and H4), a finding supported by the participant feedback provided through the online questionnaire and Talk forum. To negate this, a third 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. Despite the fact that the time spent per visit is not significantly longer, the maximum number of return visits by volunteers using such an interface design does increase (hypothesis H5).

9.7.3 The Right Tools for the Right Job

Analysis of crater markings in terms of inter-participant agreement supports hypothesis H3, showing that the type of task presented and judgement 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 findings outlined in 8.7.4 regarding 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, as also found regarding the laboratory study, 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 inter-participant 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 judgement than those made by participants using the Full interface.

9.7.4 Separating the Wheat from the Chaff

Beyond the hypotheses addressed through this study as stipulated in 9.3, analyses of both the crater marking data and website engagement has given rise to other notable patterns in participant behaviour. In terms of participant-expert comparison, participants using the Stepped and Ramped (Mark) interfaces showed greater agreement with the expert judgement than the other interfaces in terms of crater position and crater diameter respectively. This greater expert agreement also stretches to the marking of smaller craters, with participants using the Ramped (Mark) and Stepped

interfaces correctly marking the diameter of '>0 – 10' and '>10 – 20' pixel diameter craters within the same size 'bin' as the expert more frequently. Furthermore, 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).

These findings regarding apparent differences in participant-expert agreement suggest the existence of a 'filtering effect' across the different interface designs. As previously discussed, participants using the more prescribed 'task by task' Stepped interface spend longer analysing each image, whilst those using the Ramped (Mark) interface have had to analyse at least 20 images beforehand using the other Ramped interfaces (as each tool is used for a set number of images, then changes). Therefore although participants that have used these interfaces on average are fewer in number, return less often and produce less data (both in terms of images analysed and crater clusters identified), their apparent greater familiarity with the crater-marking task (via either spending more time studying each image or viewing more imagery) has resulted in better performance when compared to the expert judgement. This trend is in agreement with previous VCS research, both in terms of the existence of a committed 'super user' group (Luczak-Roesch et al., 2014) and that volunteers with a greater experience of the platform perform better (Prather et al., 2013). Not only that, but it also suggests that through interface design and the manipulation of different task workflow design factors the rate in which these super user groups form can be altered. Although such designs restrict both variety and autonomy, and therefore volunteer engagement and motivation, those that do return to the platform provide more accurate data when compared to the expert judgement.

The practical outcomes of these findings reinforce those found by the laboratory study, that when considering utilising a VCS platform the science team involved would have to balance the advantage of greater data quantity with the disadvantage of having to 'clean' out a greater number of false positives. 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 analysed and the potential size of the volunteer community taking part will all need to be considered.

9.7.5 Other Lessons Learnt from a 'Real World' Study

Beyond this thesis' focus of studying the manipulation of task workflow design factors and its affect, the 'live' situation of this study has also revealed a number of other issues that could have implications in ensuring the success of a VCS platform. They also support the recommendations of the previous study about the importance of a laboratory test (see 8.7.5), whilst additionally adding credence to the process of having a 'dry run' or beta period of the platform in order to discover issues that only arise when the system is 'live'.

As with the laboratory study, participants consistently had difficulty marking smaller craters (>0 – 10 pixels in diameter) – an issue they were aware of and reported both through the online questionnaire and the Talk forum. Although the manner of dealing with this issue varied (by marking them incorrectly as larger craters on the Full, or not attempting to mark them at all on the Stepped and Ramped), more broadly it could indicate a limitation of the imagery presented, a conclusion supported by participant feedback requesting a 'zoom' feature. This highlights that, depending on the interface design used, there could either be an issue regarding smaller craters being missed, or wrongly marked as larger, causing bias in the data.

Either issue is an important one to consider especially when crater size distribution is such a critical component of the science case.

As with the findings in the previous laboratory study (8.7.5), again there is an issue with a certain percentage of false positives existing even when the number of participants that have contributed to the crater marking rises to higher levels (5-12 participants depending on interface). Although such false positives have been derived through comparison with an expert judgement that in itself is subjective (Robbins et al., 2014), on closer inspection it is clear there are certain features and surface artefacts that are consistently misinterpreted as craters (figure 9.15). By being able to identify the types of feature that might be misclassified as a crater, even by an audience that is perhaps considered more familiar with the Martian surface, training procedures could be tailored and improved as well as potential systematic errors in the data understood pre-emptively.

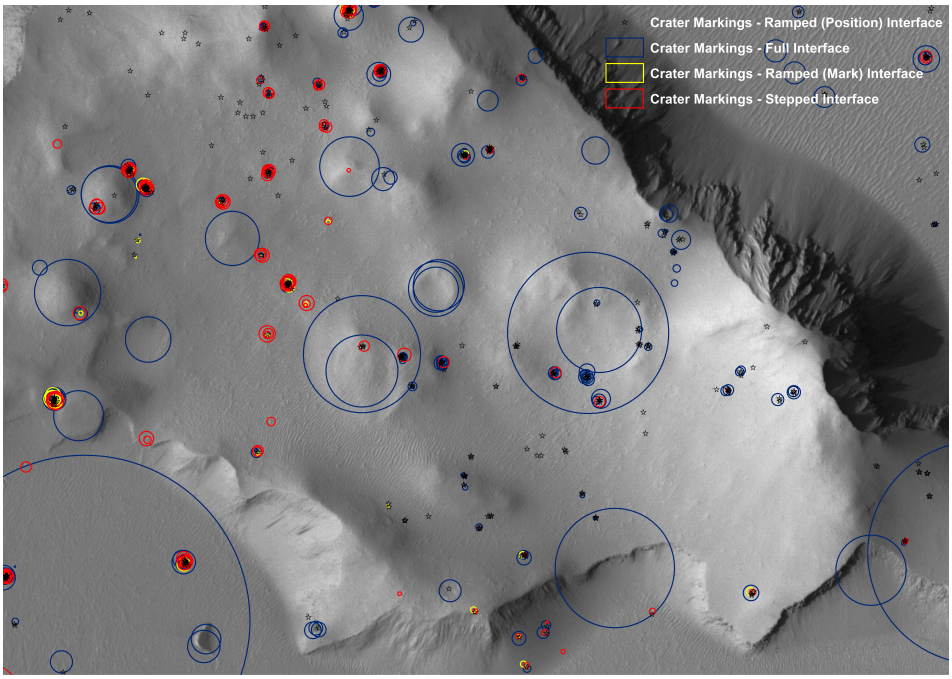


Figure 9.15: Examples of false positive participant markings (larger blue circles – marking peaks rather than craters)

Finally, the findings of the previous laboratory study regarding the relationship between participant agreements both with themselves and with the expert equivalent are reinforced. Participant markings made using the Stepped and Ramped (Mark) interfaces showed better agreement both between participants and with the expert judgement in terms of crater position and crater diameter respectively. This is an important result, as

although existing citizen science projects may have access to a sample of expert data for validation purposes, the general approach has been to use participant agreement as a measure of certainty (more participant markings increasing the likelihood of a true positive), hence ensuring a minimum number of volunteers analyse each image (Smith, 2013). Not only do the findings of this study vindicate this approach (see figure 9.14, a higher number of participant contributions resulting in a higher true positive fraction compared to the expert), but they also suggest that this relationship extends to more involved measurements of the feature themselves, such as their position within the image and their size.

9.8 CONCLUSIONS

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 judgement required had an effect on volunteers' experience, behaviour in terms of site usage and the data they produced.

Which task workflow design configurations do volunteers prefer?

Although participants' self-reported scores in response to the online questionnaire's Likert-type statements showed no significant difference between each interface, free-text responses provided through the questionnaire and the online forum indicated a preference for greater variety and autonomy (supporting postulate P7, see 4.4.3). The frequency of comments on each subject suggests they are of equal importance to the experience of the participant.

Thesis RQ_2: Does task workflow design influence volunteer behaviour?

The results regarding volunteer preference has been reinforced by participant behaviour in terms of interaction with the platform, as although there is no significant difference in terms of visit duration between each interface (~20 minutes) participants that 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 analysing less images in total.

Thesis RQ_3: Does task workflow design influence volunteer results?

When considering crater-marking behaviour, 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. As with the laboratory study in chapter 8, 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 (supporting postulate P10, Jäkel & 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 postulate P5 (greater autonomy = increased performance) especially if data quantity is important, 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 rate of false positives. 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 judgement, 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 (supporting postulate P11, 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 analyse 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 judgement.

Overall, the results of this study support the findings of previous related human factors research and those of the laboratory study when considering the opinions of the volunteer, with preference given to greater autonomy and variety suggesting that interfaces that incorporate these factors can provide an intrinsic motivation to take part. This is supported in terms of volunteer behaviour regarding how often they visit the platform. 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 motivation, engagement and performance – reinforcing the importance that VCS developers and science teams consider the analysis required, the amount needed and the potential size of their volunteer community when considering the design of a project.

9.9 SUMMARY

This chapter finishes up Part III of the thesis – the empirical work conducted in order to achieve a greater understanding of the affect of task workflow design factors on the outcomes of a virtual citizen science platform, and the potential benefits and pitfalls of differing interface designs that manipulate such factors.

While the purpose of experiment 1, presented in chapter 6, was to consider the current state of play regarding VCS platforms and doing so confirmed the diversity of Zooniverse projects in terms of the tasks required and how they are presented to the user, it also implied an affect of these designs on volunteer engagement with the site. The aim of experiment 2 was to add substance to this position, by investigating if task workflow design has a conscious influence on volunteer motivation through gathering volunteer opinions via an online questionnaire aimed at an existing VCS community. Not only did the experiment reveal a number of volunteer concerns regarding the tasks and tools, imagery presented and communication methods provided on the platform; but it also revealed an awareness of task workflow design factors and a preference in their design.

The laboratory study outlined in chapter 8 was conducted in order to observe the effect of directly manipulating task workflow design on a purposely-designed platform, unlike previous research that has used a

retrospective approach where other science-influenced factors are not controlled. It has broadened the thesis to also consider the effect of task workflow design on the data produced by the volunteer, and its accuracy through expert comparison. Not only did the findings support the previous experiments regarding volunteer motivation and experience, but they also showed that TWD factors have an effect on the data produced, in differing ways depending on the performance measures considered. Finally the 'live' study described in this chapter combined the approaches of the previous three. The findings not only showed that TWD has an effect on volunteer motivation, performance and their engagement with the platform, but also that these relationships interact with each other with differing consequences.

This discovery has implications for the design of VCS platforms, and therefore for the following Part IV of this thesis, where the contributions are discussed fully and are related to each other, with overall conclusions drawn.

PART IV
SYNOPSIS

10

DISCUSSION

The final part of the thesis incorporates the separate discussions of the four different empirical studies into an overarching review of the findings brought forward by this work, and their implications for citizen science and the design of virtual citizen science platforms specifically.

10.1 DESIGNING VCS PLATFORMS FOR THE BEST SCIENTIFIC OUTCOMES

The principle question that this thesis has attempted to answer is the seemingly simple: *Which task workflow designs ensure the best scientific outcomes?*

The empirical work presented in chapters 6 – 9 suggests that the design of virtual citizen science platform workflows in order to obtain the best scientific result is far from a simple equation to solve. The two exploratory studies and resultant hypothesis-led experiments have demonstrated the complex manner in which the design-influenced task workflow factors outlined by the model (see 4.5.1, figure 4.2) interact to influence the scientific outcomes of a VCS platform (as defined in 4.5.3).

For example, although the laboratory experiment (chapter 8) and the live experiment (chapter 9) showed that participants preferred the interface affording the greatest autonomy and variety in terms of tasks and their order of completion, this did not translate into an improvement in terms of data coverage or accuracy when compared to expert classifications. Likewise, whilst the interface with the simplest task resulted in the best performance regarding the amount of data collected, it had a detrimental effect on accuracy in terms of crater identification (an increase in false-positives) and also the constrictive, repetitive nature of the design was reported as a frustration by a number of the participants.

Volunteer engagement with VCS platforms has previously been related to motivation, both initial and over time (see 2.6.3) in existing literature, whilst behaviour and performance has been considered in terms of time spent on the site, or compared between different projects with differing scientific goals (2.6.4). These approaches have relied on a retrospective approach, incorporating data collected over extended periods of time from established projects with an established volunteer community.

In contrast, a central methodological choice for this thesis was to forego the use of existing large datasets and instead create a bespoke platform from conception to launch addressing an original scientific case. This approach, with its focus on studying the effects of design-influenced workflow factors, was taken for three key reasons: 1. Specific task workflow factors can be manipulated whilst other, more science-influenced factors shown to influence volunteer motivation can be controlled, 2. A science case can be used that has an existing 'expert' sample dataset for comparison, and 3. The main goal was to study systematic ways in which task workflow factors can be influenced at the design stage, and therefore by going through the entire design process more can be understood regarding the extent and limitations to which such factors can practically be manipulated.

The first 2 studies of the empirical part of this thesis provide an overview of the current virtual citizen science landscape. Study 1 (chapter 6) compared 12 different existing VCS projects, all members of the leading Zooniverse platform, in terms of the task workflow designs they employed and presented to their volunteers. It revealed a varied landscape, with no apparent 'best practice' and no relationship between the science involved and the task workflow designs employed. Probing deeper, an analysis of the website metrics for each site suggested that task workflow design factors can influence volunteer engagement in various ways. Both findings add credence to the thesis' approach, demonstrating not only that design-influenced factors can affect the success of a VCS platform but also that certain designs can be utilised over many different science cases – meaning that the findings can inform the citizen science community beyond the planetary science focus. Study 2 (chapter 7) sought the opinions of an existing volunteer community, specifically that of the Planet Four project. Unlike previous research that considered volunteers' extrinsic motivations

for taking part (Raddick et al., 2010; 2013) it focussed on intrinsic motivation, regarding platform functionality, the design of the tasks and tools and how they were presented. Participants' responses covered a range of topics, and whilst feelings regarding the appearance and general usability of the platform tended to be positive, concerns were raised regarding the completion of the task itself, many of which could be directly related to the task workflow design of the site and the factors that influence it.

The final two experiments of the empirical part of this thesis (chapters 8 & 9) were hypothesis-led in approach, informed by both the findings of the aforementioned studies and previous literature regarding citizen science and crowdwork in general (see postulates, 4.4.3). Through a specifically developed VCS platform, task workflow design factors were directly manipulated and their effect measured. The first (chapter 8) was conducted in a laboratory-based environment, allowing for external factors to be controlled and the conditions directly compared through a within-subjects design. The final experiment of the thesis (chapter 9) was perhaps the 'acid test' of the thesis approach, conducted through launching the Planet Four: Craters platform live online, advertised to the existing Zooniverse community but also accessible to anyone with an internet connection and the associated URL. In testing the effect of task workflow design in this context, where the participant rather than the researcher controls other external factors, it in essence tested the practicality of considering task workflow design factors at the design stage of a VCS project to ensure its success – a key aim of the thesis overall.

Both experiments demonstrated the affect that TWD factors (specifically autonomy, variety, task type and judgement required) can have on VCS volunteer experience and their data output. Self-reported responses revealed participants had a preference for interfaces allowing greater variety in the tasks undertaken, and greater autonomy in choosing how to approach them. When presented with a simpler task, both experiments showed that participants produced a significantly greater number of markings – albeit with the previous caveat regarding accuracy. The laboratory-based experiment (chapter 8), where external conditions were controlled and a within-subjects design used, revealed that task workflow

design factors influenced the performance of individual participants, their agreement with the expert equivalent data being significantly improved when the task is broken down into its constituent parts and the specific metric required clearly communicated. Finally, despite their self-reported preferences, the 'live' experiment showed that TWD factors such as autonomy and variety did not have any significant affect on the amount of time participants spent on the platform, with the average visit duration of ~20 minutes the same independent of the interface used. Observations of participant performance compared to the expert showed that the more restrictive interfaces (least autonomy), which forced participants to take longer on an image or view more images before attempting more complicated tasks – in effect filtering participants' experience, resulted in significantly better agreement with the expert.

In summary, this thesis has made several contributions towards answering the question of the best design of a virtual citizen science platform.

Analysis of participants' behavioural response, in terms of both the scientific output produced and their engagement with the platform, suggests that task workflow design factors considered at the design stage of a platform can have an influence on its outcomes. Self-reported participant responses add context to these findings, and qualify the ways in which TWD factors and their effects interplay with each other on a broader scale. Moreover, these findings add credence to the treatment of virtual citizen science as a form of work and the analogy with 1960s industrial work mechanisms, meaning lessons can be learnt from research in this field. The factors originally identified by Hackman & Oldham's (1975) Job Characterisation Theory and the subsequent research on their effect regarding motivation and performance (Hutt et al., 2013; Eveleigh et al., 2014; Jäkel & Wichmann, 2006) has been shown to be applicable to the citizen science case and therefore suitable to inform future VCS project design in order to improve their scientific outcomes.

However, there is a caveat, as not all of the postulates (4.4.3) derived from previous research have been supported through the empirical work of this thesis. In particular, the assumption that greater volunteer autonomy leads to an increase in performance (Oldham & Hackman, 2010; Dubinsky & Skinner, 1984) has not been confirmed by the experiments of this work,

with the interface affording the most autonomy (Full) not resulting in a significant improvement of performance in terms of accuracy or volume. This finding may well highlight a limitation of considering citizen science as a type of work and also the difference between expert and amateur users. Trained employees, who are experts in their role and understand their position in an organisation, might well benefit from more autonomy through understanding where their skills can and cannot contribute. Alternatively, the untrained volunteers of citizen science projects might feel more separated from the process and less sure of their potential contribution – at least in the short term until their experience increases and motivations change (Rotman et al., 2012).

Whether the effect of manipulating certain task workflow design factors is perceived as beneficial will be determined by the specific aims of each VCS platform, its science team, and its developers. The difficulty faced by any endeavour to develop a VCS platform will be to perform a balancing act, weighing up the importance of volunteer engagement, the data needs of the science case and the resources that can be committed to both reducing the data and interacting with the platform community.

10.2 CONCEPTUAL MODEL OF TASK WORKFLOW DESIGN FACTORS

In considering the importance of task workflow design to the success of a VCS project, how justifiable is it to break down TWD into components? Components such as design and science-case influenced factors that make up the model described in section 4.3 and form the conceptual core of this thesis. Perhaps the biggest challenge of contributing to the field of citizen science is in reconciling the differences between the disciplines it unites in its derivation. The core developers of citizen science platforms often approach problems in terms of their system requirements, dealing with constraints such as limitations in technologies and resources. The scientists who have developed the scientific need for the platform approach the situation in terms of the data they require, its accuracy and validity in answer the research question posed. Promoters of the platform, and associated researchers that consider citizen science as its own field of study, view the development of VCS systems as an opportunity to study and observe the interactions between volunteers, scientists and the platform –

but perhaps are reluctant to give recommendations or draw implications to inform design (Dourish, 2006). Although recent investigations have resulted in the publication of citizen science interface guidelines and best practice recommendations (Simpson et al., 2014; Tinati et al., 2014), some of which directly relate to task workflow (Eveleigh et al., 2014), none yet have attempted to quantify any benefit through direct empirical study.

Therefore, the justification for the conceptual model of task workflow design factors arises from a pragmatic attitude towards the disciplines. The breakdown of TWD factors into those that can be influenced either by the science case or at the design stage makes it possible to systematically study different TWD 'configurations', whilst holding constant other factors that might have varying influence on the outcomes of a project. In doing this, the thesis facilitates the engineering of VCS platforms through the adaptation of TWD factors, in order to improve the scientific outcomes. As such, the findings have a practical application not only to the developers, but also the science teams and researchers of existing and future citizen science projects.

In deriving its constituent factors, the model has also undertaken a novel application of cognitive psychology by applying psychophysical methods theory within a real world environment. In order to categorise the different types of task undertaken by VCS volunteers and the associated judgement required, Pelli & Farell's (2010) more formal derivation regarding observer perception (see 4.2) was applied. Although originally developed to test display characteristics using more abstract visual stimuli such as dots or gratings, the results of the hierarchical task analysis carried out in chapter 6 suggest that they are suitable to be applied to the observations made by the volunteers of virtual citizen science platforms. Moreover, the typology can be used to group VCS projects in terms of their demands of the volunteer, beyond the science discipline they address.

Overall, the model has been synthesised by applying methodologies from several related disciplines involving crowdwork and visual inspection – through an extensive review of the literature involved (see chapter 3), testing their findings applicability to the citizen science domain through the empirical studies described in this thesis.

10.2.1 Empirical Contributions to Validating the Model

The empirical part of this thesis described through the studies of chapters 6 & 7 and the experiments in chapters 8 & 9 supports the pertinence of almost every factor in the model of task workflow design factors – but not always in the way expected as described by the postulates synthesised from previous research. Table 10.1 shows the task workflow design factors of the model, whether they are design or science case influenced, and the disciplines from which the factor was synthesised from the literature – i.e. in relation to job characterisation theory and psychophysical methods work. Additionally, it shows the locus of TWD factors in the design of studies 1 & 2 (chapters 6 & 7) and their manipulations in experiments 1 & 2 (chapters 8 & 9). Where the investigatory hierarchical task analysis and questionnaire studies considered both science case and design influenced factors in a retrospective approach, the laboratory and 'live' experiments controlled only the 4 design-influenced factors (autonomy, variety, task type, and judgement) – holding the other science-influenced factors constant by design. Finally, the table shows the types of data collected in order to understand the effect of task workflow design factors – with the most significant findings also stated.

Overall, the columns of table 10.1 describe the task workflow factors and their empirical manifestations within this thesis, whilst the rows explain the coverage of the factors in relation to the two studies and two experiments of the thesis' empirical work.

Table 10.1: Model of Task Workflow Design Factors, its grounding in related work, and this thesis' contribution

Task Workflow Design Factors of Virtual Citizen Science											
Type	*Design-influenced factors*				*Science-influenced factors*			*Scientific Outcomes*			
Factor	Autonomy	Variety	Task type	Judgement	Significance	Task identity	Feedback	Motivation	Engagement	Performance	Overview
Related work presented in	3.6.2 & 4.2.1	3.6.2 & 4.2.1	3.2 & 4.2.2	3.2 & 4.2.2	3.6 & 4.2.1	3.6 & 4.2.1	3.6.1 & 4.2.1	2.6.3 & 4.5.3	2.6.4 & 4.5.3	2.5.1 & 4.5.3	Chapters 2, 3 & 4
Coverage of this thesis	*Manipulations (i.e. controlled by design)*							*Data (i.e. observed/measured)*			*Main findings*
Study 1 Chapter 6	Retrospective overview of 12 Zooniverse projects with varying task workflow designs							Website analytics			No current 'best design' re. TWD
Study 2 Chapter 7	Volunteer views and feedback on interface and task design through questionnaire				Kept constant through considering one platform (planetfour.org)			Self-reported questionnaire (Likert & statements)		N/A	Volunteers have issues with TWD
Experiment 1 Chapter 8	Any order vs. set order	One task vs. many	Detection vs. ID vs. Matching 'Yes/No' vs. threshold		Constant – through science case (Martian crater counting)			Self reported – questionnaire responses		Behavioural – crater markings	Sig. diff. in measures affected by TWD
Experiment 2 Chapter 9	Any order vs. set order	One task vs. many	Detection vs. ID vs. Matching 'Yes/No' vs. threshold		Constant – through science case (Martian crater counting)			Self reported - questionnaire Behavioural – website analytics		Behavioural – crater markings	Sig. diff. in measures affected by TWD

10.2.2 Design-Influenced Task Workflow Design Factors

The key aim of this thesis was for the findings to have a practical impact on the design of future virtual citizen science platforms, so that lessons can be learnt and the scientific output improved – this measure depending on the specific scientific aims of the science and development team involved. In order to achieve this aim, the main argument of this thesis was that the design of tasks carried out by citizen science volunteers can be described by component factors (the model of task workflow design) – that are either controlled by the science case involved, or can be influenced at the design stage independent of the discipline it addresses.

The hierarchical task analysis study described in chapter 6 has shown that it is possible to define virtual citizen science projects in terms of the design-influenced TWD factors that make up their classification interfaces. It has also shown that this is a dynamic and variable affair with no clear pattern – and that VCS design-influenced factors are not clearly controlled by the scientific discipline involved and therefore the science case-influenced factors described in the conceptual model. The online questionnaire outlined in chapter 7 has shown that participants are aware of and have concerns regarding issues that can be associated with TWD design-influenced factors – adding credence to the belief that such factors can be manipulated to affect the scientific outcomes of a VCS project.

Through observing the direct manipulation of design-influenced TWD factors, experiments 1 & 2 described in chapters 8 & 9 respectively have shown that such factors can influence the scientific output of VCS projects, including data accuracy, data volume, platform engagement and volunteer motivation. Furthermore, they have demonstrated that these effects exist independent of science-influenced TWD factors, as they were held constant through experimental design. This is an important finding, as it shows that different development strategies regarding the design of the volunteer task can be tailored to the requirements of the project - and that this is possible when in practical terms factors regarding the science case are 'set in stone', decided by the academic need of the discipline to ensure the contribution and credence of the work. Figure 10.1 revisits the model of design-influenced TWD factors described in chapter 4, additionally showing how

the empirical findings of this thesis test the influence of these factors on VCS outcomes and performance.

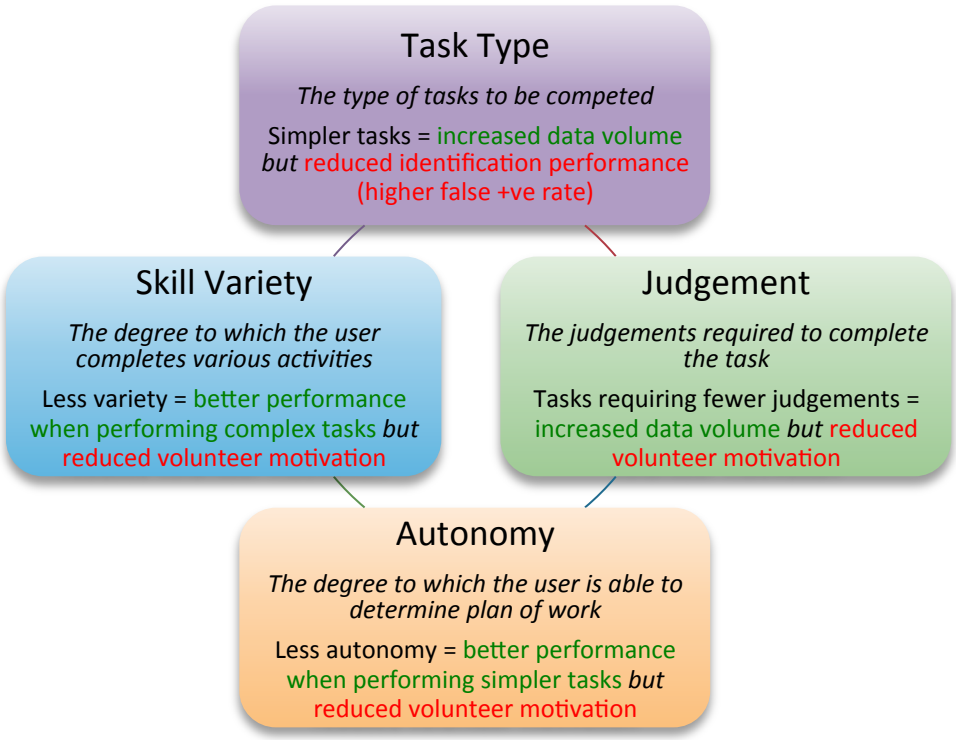


Figure 10.1: Effect of Design Influenced TWD Factors on VCS Outcomes

10.2.3 Science Case Influenced Task Workflow Design Factors

The model described in table 10.1 also outlines science-influenced TWD factors. Although not the focus of this thesis, and therefore not manipulated within the empirical work, their existence and influence have been synthesised from related literature and included so that they can be considered and held constant through experimental design. The need for the outcomes of the thesis to be practically applicable in future VCS design has led to the assumption that such factors cannot be realistically controlled at the design stage, due to their reliance on overarching scientific goals – but this does not tell the whole story.

As explained in section 4.4 (and visualised in the model – figure 4.1), the degree to which a task workflow design factor is either science case or design-influenced is not a binary condition, but more realistically a position on a sliding scale. For instance, although the 'significance' factor – the tasks'

meaningfulness to a specific scientific question – is almost completely controlled by the scientific aims of the project, the 'feedback' factor – the volunteers' awareness of their performance – is influenced both by the science case (ground truth, data visualisation/context) and also design (timing of feedback, positive or negative). In order to constrain the thesis to a realistic scope, a 'line in the sand' has been drawn between design and science-influenced factors, however the influence that design can have on other factors should not be ignored. Feedback especially could become a pivotal consideration during VCS development, as future platforms increasingly combine the tasks and data compiled by volunteers with those performed by automated processes (Marshall et al., 2015). The relationship between these agents, as well as the feedback provided to volunteers regarding automated analysis, could influence the scientific outcomes of a project and therefore its perceived success.

More broadly even if, as hypothesised by this thesis, science case influenced factors are uncontrollable at the design stage of a platform they might well be useful in determining if a VCS platform approach is suitable in addressing a particular scientific problem. Website analytics studied in study 1 (chapter 6) have shown how certain disciplines can resonate with perceived public interest that in turn creates a 'hook', in some cases resulting in exposure on primetime television (BBC's Stargazing Live). This can drive traffic to the platform at launch, an important outcome especially when considering the 'long-tail' behaviour of VCS volunteer engagement. If the scientific question involved is deemed difficult for the public to relate to and its importance hard to communicate (factors 'significance' and 'identity' in figure 10.1), this could become a barrier to the success of a VCS project, especially if the number of classifications required for the scientific output to be impactful is large.

10.2.4 Interplay of TWD Factors

In addition to the influence of the science case and development design choices on task workflow design factors, behavioural and self-reported analysis has revealed the existence of interplay between the factors themselves – collectively contributing to both the volunteers' perceived motivation and the data produced. In experiment 1 (chapter 8), although

participants generally reported favouring both autonomy and variety, a small number referenced an interaction between the two. For instance, one less confident participant indicated a preference for variety only when autonomy was restricted i.e. they were led through each different task step by step, and found having the freedom to choose from a variety of tasks led to a fear of performing poorly or 'missing something out'. Experiment 2 demonstrated the effect of factorial interplay on participant behaviour, both in terms of crater marking and platform engagement. The interface involving only the simplest task resulted in the largest data volume produced by volunteers, with the caveat of restricting variety and autonomy. This restriction had a knock-on effect regarding performance, with perhaps the fatigue factor caused by repeatedly doing the same, straightforward task resulting in a higher rate of false-positive identifications. Furthermore, the self-reported volunteer frustrations regarding limited variety and autonomy translated into a reduction in platform engagement, with volunteers returning to the platform less often.

It could therefore be concluded that it is not the consideration of individual TWD factors alone, but more importantly their interplay and interactions, which determines the best citizen science platform design to ensure it achieves its scientific goals. Conversely, there cannot be a 'one size fits all' task workflow design that is the optimal framework for 100% of virtual citizen science projects. Developers will have to carefully consider the aims of the project, what is needed to achieve them (data volume, accuracy, number of classifications etc.) and how TWD factors both individually and collectively can be designed to best support the process. In the next section, the practical implications for VCS design are outlined further, discussing the potential opportunities and limitations for VCS projects that consider task workflow design in their derivation.

10.3 IMPLICATIONS FOR VIRTUAL CITIZEN SCIENCE DESIGN

The practical connotations of the thesis' findings, specifically regarding task workflow design, are now expanded upon in regards to the virtual citizen science domain. As described in section 4.5.3, three important factors contribute to a successful VCS scientific outcome: volunteer engagement/motivation, data volume and data accuracy. Through

synthesising the behavioural results of chapter 9 and the website analytics of existing platforms in chapter 6, the affect of task workflow design on scientific outcome can be inferred for three different scenarios:

1. *The Heavy Weight:* A virtual citizen science project that involves a highly topical subject, with lots of media attention, and as such benefits from a 'live' launch on primetime television (i.e. planethunters.org)
2. *Mainstream:* A project that involves a subject matter with a reasonably large scientific community to promote it, and might well also benefit through cross-promotion from other established VCS platforms (i.e. ancientlives.org)
3. *Small but Committed:* A smaller project involving a more niche/narrow subject, that therefore doesn't command the same level of promotion and attention. However, the volunteers that do take part tend to be both committed and enthusiastic (i.e. oldweather.org)

Figure 10.2 shows the predicted number of classifications over time for each of these scenarios, depending on the type of task workflow design used as per the empirical work of this thesis.

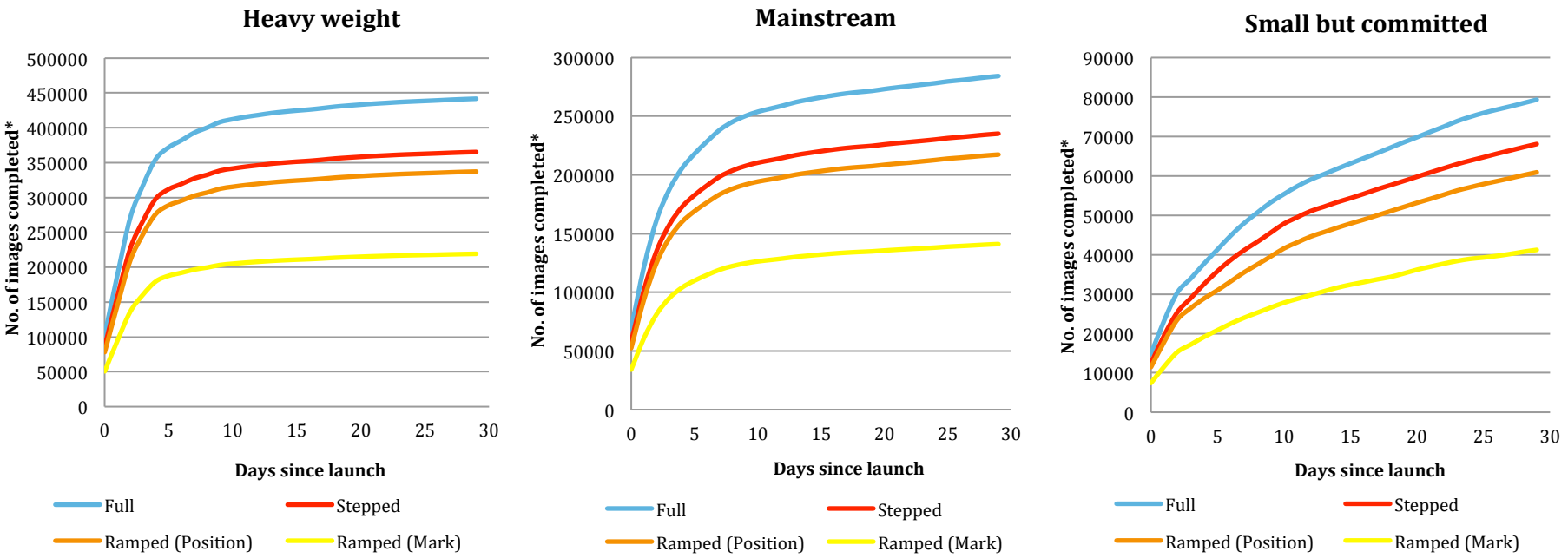


Figure 10.2: Number of classifications (images) completed over time using each TWD interface for each scenario

*NB For an image to be considered 'completed', the number of participant classifications required to assume a 100% true-positive identification rate has been taken into account – as compared with the expert equivalent (see figure 9.14)

In order to calculate the total number of images completed on a given day since launch for each of the scenarios described, the following formula was used incorporating data both from the analytics study in chapter 6 and participant behaviour data from chapter 9:

$$Images\ Completed = \frac{(NV + \sum_{n=1}^{n} V_n R_{n+1})\ I}{C_{tp}}$$

Where:

NV = number of first-time visits (calculated from representative analytics data in chapter six for each scenario)

V_n = Number of nth time visits (calculated from chapter 6 analytics data)

R_n = Percentage of total visitors who will visit for an nth time (calculated from chapter 9 participant behaviour for each interface)

I = Number of images classified per visit (calculated from chapter 9 marking behaviour)

C_{tp} = Number of classifications required for 100% true positive rate (calculated from chapter 9 expert-participant agreement data)

10.3.1 Turn Up the 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 number of combinations tested, then task workflow design factors can have a practical influence. For all three scenarios, figure 10.2 demonstrates that by employing a design that maximises task variety and autonomy (the Full interface), the resultant positive effect on intrinsic volunteer motivation (i.e. return visits) could potentially lead to a greater number of classifications being completed (~20-100%).

As discussed in both the laboratory and live experiments, initial findings agree with Eveleigh et al.'s work (2014) suggesting that presenting a simpler task (Ramped Position) results in a greater number of volunteer classifications. However, due to the corresponding increase in false-positive identifications, the knock-on effect of needing more 'eyes' on each image to ensure a true reading more than negates this advantage overall.

Compared to the full interface, it is predicted that ~30% fewer classifications would be completed in 30 days for each scenario. If the science team of a particular project were more concerned with false-negatives (missed annotations) than false-positive identifications, or have in place a suitable structure for dealing with false markings, then the simpler task approach would be most appropriate – with the caveat of less detailed results (crater position only instead of position and diameter).

10.3.2 Quality over Quantity

If a project's success is conditional on the accuracy of volunteer data, measured either through the variability of their agreement or through expert/gold standard data comparison, again task workflow design can be used to influence the outcome. Classification interfaces that restrict the volunteer, either by forcing them to spend more time on an image or analyse more images before more complex tasks are available (Prather et al., 2013) can be used to improve accuracy. For instance the findings of this thesis have shown that by restricting autonomy alone (Stepped interface), volunteer performance significantly increases when performing a simpler task (marking crater position). When both volunteer autonomy and task variety are restricted (Ramped Mark interface), volunteer performance significantly increases when performing more complex tasks (crater diameter – incorporating identification, position and size measurements). Such approaches however do demonstrate the importance of variety and autonomy on volunteer motivation (Dubinsky & Skinner, 1984), with the number of predicted completed classifications reduced by ~20% and ~100% respectively compared to the Full interface.

10.3.3 A Balancing Act

In the previous two examples, the practical findings of this thesis have been applied to polarised VCS cases, either having a preference for data volume or accuracy. In reality, it is likely that to fulfil a meaningful scientific goal both will be important, the data needing to be a certain level of accuracy and the sample size large enough to draw significant conclusions. Although not directly studied as part of this thesis, the findings add credence to a more fluid approach to task workflow design that changes over time.

For instance, beginners to a VCS project might be presented with a more restrictive TWD interface (such as the Stepped or Ramped) to ensure the accuracy of the data they produce. As their experience with both the platform and the stimulus (i.e. the features to be identified and the imagery presented) increases, the interface would change into one involving greater variety and freedom (Full). The timing of this change will need careful consideration – at the right time to prevent a loss of motivation regarding taking part, but not too soon that any positive performance gains due to familiarisation with the current task are lost (Lasecki et al., 2014; Rzeszotarski et al., 2013). Finally, for the most committed volunteers, the choice of interface and therefore task workflow design could be placed into the hands of the individual, allowing them to choose the tasks they complete and how based on more personal and environmental conditions, i.e. how much time they have, concentration levels etc.

This evolving approach to task workflow design would directly support the motivational arc theory proposed by Nov et al. (2011), adapting to take into account the changing motivation of volunteers as their experience increases over time.

10.3.4 Unforeseen Practical Implications

Beyond the direct implications that the manipulation of task workflow design factors can have on the design of virtual citizen science platforms, the predictive model of classifications over time derived in figure 10.2 has revealed other practical uses. Firstly, it allows developers of prospective projects to make a first attempt at predicting its outcome in terms of the number of completed classifications, based on other known factors. For instance, if a project team is aware that the scientific case being addressed is narrow, not well reported in the media, and they have few resources for publicity (*small but committed* scenario) – the model can prevent unrealistic expectations, showing that only classifications in the order of 10,000s a month should be hoped for rather than 100,000s or millions, independent of the type of task workflow design utilised. Contrastingly, if there is confidence that the science being addressed resonates with the public, which could result in a large media interest with the resources to exploit it, such as an appearance on Stargazing Live (*heavy weight* scenario)

– then a larger number of classifications can be expected. If this is in turn coupled with a relatively small data volume to analyse, then the developers would be free to design tasks and interfaces that produce the most suitable results for their aims, irrespective of the cost to the number of classifications over time.

Secondly, whilst the thesis has shown the effect of task workflow design on volunteer engagement in terms of the number times they return to the platform – it has also revealed that the average time per visit remains stable, at ~20mins. In knowing this information at the design stage, developers can consider how to best use the volunteers' time, in terms of the tasks they undertake, how many images they can classify and ultimately how this influences their output. Furthermore, this process could be tested before launch on both expert and non-expert participants, perhaps to gauge whether the overall scientific goals are realistic and in what timeframe.

Finally, the website analytics studied in chapter 6 and the predicted output rates in figure 10.2 have revealed a pattern regarding the type of volunteer, depending on the scenarios defined. Projects with a larger number of new visitors attracted at launch (i.e. the *heavy weight* & *mainstream* scenarios) tend to keep a higher proportion of new users over the first 30 days of the project, when compared to the *small and committed* scenario (~70% versus ~55%). As shown by the relatively shallower gradient of the curves in figure 10.2 (left and middle graph), the majority of these new users visit near to the launch, with the number of completed classifications levelling off considerably at around the 10-day mark. This difference between new and return visitors is also highlighted by the relative differences between each TWD interface. The gap between the total completed classifications remains more constant from day 10 onwards with new volunteers motivated extrinsically by the science involved, but increases continually until day 30 when considering the scenario with a higher ratio of return users (figure 10.2, right graph), perhaps more likely to be motivated by intrinsic factors involving the task itself.

The practical implications of this trend could manifest itself in two distinct design directions. Firstly, if the development team behind a VCS project are confident that it will resonate with the public and gain sufficient media

attention, a task workflow design approach that concentrates on harvesting the best data to answer the scientific problem could be taken. Although such designs might be at the cost of volunteer engagement, it would be hoped that the constant stream of new visitors provided by the extrinsic motivation of the science case would fill the gap. Contrarily, if the developers had less confidence regarding the reach of the project, and therefore had limited expectations regarding the size of audience it might attract – they could use the *small but committed* scenario as guidance. In this case a task workflow design approach that concentrates on volunteer engagement might be used, providing an intrinsic motivation to return visitors who form a greater proportion of the community. The cost of this approach in terms of the data produced would have to be addressed, and ultimately could form a barrier to the use of citizen science as a solution.

10.4 SUMMARY

This discussion has interlaced the main findings of this thesis, specifically findings regarding the task workflow design of virtual citizen science platforms (10.1), the contributions made by the empirical work towards the conceptual model of task workflow design factors (10.2), and finally the implications for the design of virtual citizen science systems in terms of the task workflow designs they employ (10.3).

The behavioural results from the two experiments showed the different ways in which task workflow design factors can influence the scientific outcomes of a VCS project, whilst the discussion and feedback from participants produced qualitative findings adding context, highlighting the dynamic ways TWD factors can affect their motivation (10.1). The task workflow design factors were then discussed separately in terms of design and science-influenced, and individual design-influenced factors revisited in terms of their interplay with each other, with the aim of supporting the model's credibility (10.2).

Finally, the lessons learnt empirically from the manipulation of task workflow design factors led to a deliberation of the practical implications of different approaches, sustained by a model of predicted classification rates dependent on different launch-day scenarios (10.3).

11

CONCLUSIONS

The final chapter in short summarises the approach of this thesis, by stating the main conclusions in relation to three key themes, namely: methodology, understanding task workflow design, and implications for design. Finally, the thesis concludes with a critique of the research and a consideration of future work that might be inspired.

11.1 SUMMARY

The main argument of this thesis is towards the consideration of virtual citizen science participation as a form of work, making the analogy between that of the repetitive, never-ending tasks involved with 1960s mechanisms of industrial work. In doing this it opens up the opportunity to learn from many decades of related human factors research, specifically Hackman & Oldham's (1975) Job Characterisation Theory originally applied to the industrial case and subsequently adapted to crowdwork, and the consideration of psychophysical methods when performing visual inspection tasks. The thesis bridges the findings of this research with the citizen science domain, considering how they can be adapted to improve motivation and performance – ultimately to improve the scientific output of VCS platforms.

The practical approach adopted by this thesis has led to the derivation of a conceptual task workflow design model, comprising of the influencing factors identified by human factors research but applied directly to the design and presentation of virtual citizen science tasks and classification interfaces. The constituent TWD factors have been defined in two ways: 1) design-influenced – those that can be directly manipulated by the developers at the build stage of the project and 2) science case-influenced – those that are more controlled by the science problem being addressed.

Although in order to define a finite scope of the work a line has been drawn in the sand between these two types, it is recognised that in reality TWD factors lie somewhere on a spectrum between the two definitions, a position that can change on a project-by-project basis.

Overall, the studies of this thesis inform a multi-faceted understanding of current VCS task workflow designs, whilst the direct manipulation of TWD factors in the experiments prompt discussions on different design strategies to ensure the best scientific outcomes. Although the focus of the thesis is on the systematic study of design-influenced TWD factors in the context of a purposely-designed platform, the discussion chapter has also highlighted ways in which science case-influenced factors may be interwoven into the process in order to define realistic expectations.

11.2 KEY CONTRIBUTIONS AND CONCLUSIONS

As aforementioned, the thesis' work contributes to three key areas:

1. The *methodology* of studying virtual citizen science task and interface design in a pragmatic environment
2. *Understanding how task workflow design* factors can effect the scientific outcomes of a VCS platform
3. *Implications for the design* of VCS systems that have differing aims regarding the data produced and the engagement of their volunteers

These contributions are summated in the following:

11.2.1 Methodology

The methodological approach of this thesis is founded on the consideration of related human factors research involving similar mechanisms that have an analogy to the virtual citizen science case. This posed the challenge of describing the influencing factors identified by such research (i.e. autonomy, variety etc.) in terms of their citizen science counterpart.

In order to achieve this, Hackman & Oldham's JCT characteristics were redefined in terms of their implications regarding citizen science, whilst the types of visual inspection tasks and judgements described by

psychophysical methods theory were related to those required of citizen science platform volunteers. These newly defined factors were then amalgamated to describe the task workflow design of a VCS platform's classification interface. The hierarchical task analysis carried out in chapter 6 has demonstrated the suitability of this novel interpretation, successfully defining 12 current Zooniverse VCS platforms in terms of their differing task workflow design configurations.

Another key aim of the thesis that guided the methodology was to ensure the findings had a practical application to VCS development. This led to the derivation of a model of task workflow design, placing the constituent factors on a spectrum between design and science case influenced. The model allowed for the development of a purpose built citizen science project, which through the experiments described in chapters 8 & 9 allowed both the testing of design-influenced TWD factor manipulation whilst keeping constant other science case-influenced factors. Whilst chapter 8 tested the effect of TWD manipulation in a controlled laboratory environment, the 'real world' scenario of chapter 9 ensured that any findings made in terms of the influence of TWD factors on the scientific outcomes of the project were made in a realistic setting – and hence applicable to future VCS design configurations.

The practical contributions made to citizen science methodology are:

- An adaption of JCT to describe the characteristics of VCS task design and presentation (see table 4.1)
- The derivation of a conceptual Task Workflow Design model, describing factors as either design or science case influenced (see figure 4.1)

11.2.2 Understanding Task Workflow Design

The main conclusion of this thesis is that whether the task workflow design of a virtual citizen science project can have a positive influence on its scientific outcomes depends on the varied ways the different TWD factors interact during the classification process. In its ambition to understand VCS task workflow design, the thesis answers the research question *"What Task*

Workflow Designs are currently used by existing citizen science projects?" by making the following contribution:

- A hierarchical task analysis of 12 existing Zooniverse projects, defining them in terms of their volunteer task configuration (see 6.3.2)

The findings of which reveal a varied picture, with many different task workflow designs existing seemingly independent of the scientific discipline involved. Furthermore, a comparison of the website analytics of each project also revealed differences, appearing to suggest that TWD factors at least have some influence over volunteer motivation and engagement.

Overall, the findings from the empirical part of this thesis support the legitimacy of the conceptual Task Workflow Design model, showing that both design-influenced and science case-influenced TWD factors can influence the different scientific outcomes of a VCS platform. Although practically science-influenced factors cannot be affected, a key argument of this thesis is the importance of design-influenced TWD factors, and was largely supported by the empirical studies conducted. Specifically, the manipulation of autonomy, variety, task type and judgement was investigated in order to answer the principal research question:

$RQ_{principal}$: What task workflow designs ensure the best VCS scientific outcomes?

In this regard, the empirical work has demonstrated that:

1. Design-influenced TWD factors influence the intrinsic motivation and therefore engagement of VCS volunteers, with the findings found by previous human factors work regarding their manipulation holding true for the citizen science case.
2. The data collected by virtual citizen science platforms are similarly affected by the interplay of design-influenced TWD factors, and although for the most-part previous human factors findings persevere some limitations of the work analogy are revealed.

In reference to 1, Study 2 (chapter 7) advanced the work of the hierarchical task analysis and revealed a number of different volunteer concerns through the canvassing of an existing VCS community. A number of these

concerns were directly related to the process of analysing the data (in essence, the work required) and therefore are relatable to task workflow design factors. Hence, the conclusion is that design-influenced TWD factors can effect the intrinsic motivation of VCS volunteers. Experiments 1 & 2 (chapters 8 & 9) supported this finding, with participants' self-reporting a preference for interfaces which allowed greater autonomy and had a greater variety of tasks. This increase in motivation also translated to influence participant engagement, as experiment 2 showed that users were more likely to return to the platform and to do so more often when using the interface with greater autonomy and variety (answering RQ_2, *Does task workflow design influence volunteer behaviour?*).

Concerning 2, analysis revealed that the interplay of different design-influenced TWD factors affected participants' crater marking data in a complex way, answering RQ_3 (*Does task workflow design influence VCS volunteer results?*). The subsequent TWD-related phenomena have surfaced as important volunteer data mediators:

- *Task Type & Judgement* – Simpler tasks with fewer volunteer judgements result in a significantly greater data volume being collected, and at a faster rate, however accuracy is reduced in terms of an increase in false-positive markers.
- *Task Type* – Volunteer accuracy (in terms of inter-participant and expert agreement) is significantly improved when the task can be related to its aim – i.e. the volunteer is aware of the measurement that the task is attempting to capture.
- *Autonomy* – Task workflow designs that constrain autonomy, forcing volunteers to complete every task and in a set order (and therefore spend more time on each classification), result in significantly greater volunteer accuracy - both in terms of inter-participant and expert agreement.
- *Task Variety* – VCS Interface designs that constrain task variety, forcing volunteers to do only one task on several classifications before moving on to a different one, result in significantly greater volunteer accuracy – both in terms of inter-participant and expert agreement.

- *Autonomy & Variety* – TWD's that involve both greater autonomy and variety, allowing the volunteer to do a number of different task types in their chosen order, result in significantly reduced volunteer accuracy, opposing the expected outcome assimilated from previous human factors research.

The main conclusion therefore is that different design-influenced task workflow design factors influence VCS scientific outcomes (volunteer engagement, data volume, data accuracy) in different ways – and it is the interplay between these effects that determines the answer to the principal research question. This being the case, it appears that designing a 'one size fits all' task workflow configuration to ensure the best VCS results is a highly ambiguous undertaking that is unlikely to be suitable for 100% of science aims. It would be more rational for the developers of VCS systems to balance the effects of task workflow design factors against the aims of the project, and the needs of their volunteer community on an individual basis.

11.2.3 Implications for Design

At the beginning of this thesis, the principal research question outlined the importance of the thesis' applicability to the design of virtual citizen science platforms. With this aim in mind, the conceptual model of TWD factors was derived and factors that could be influenced at the design point of a project identified. The empirical work demonstrated that different strategies could be implemented in regards to the manipulation of these factors, in order to have a positive influence on differing scientific outcomes.

The discussion in chapter 10 regarding the empirical works' differing classification interfaces highlighted the concessions inherent when deriving strategies for differing scientific outcomes (those that prioritise data volume, accuracy or volunteer engagement, see 10.3) for which a number of design recommendations were suggested. To add extra substance to these recommendations the empirical findings were combined with the website analytics of existing VCS platforms (chapter 6), so that science case-influenced factors identified in the TWD model were also assimilated. The outcome was a predictive model (see figure 10.2) that

took into account the influence of both differing task workflow design configurations and the potential public interest in the science case on the classification rate.

Through the model, opportunities regarding the development of virtual citizen science platforms were presented (see 10.3.4). They not only exploit the potential influence at the development stage of task workflow design, but also highlight the perhaps limiting factors regarding the science case involved. Not only could the model prevent the presence of unrealistic expectations regarding the rate of classifications, but also allows developers to tweak task designs and tools in order to balance the number of expected classifications, the number required and their accuracy – perhaps ultimately providing a decision tool for the suitability of a virtual citizen science approach.

11.3 CRITICAL REFLECTION

A multi-faceted education in Physics, Astronomy, Remote Sensing and Human Factors has formed the disciplinary background of the author of this thesis. This has allowed for multiple views to be taken regarding its core questions, methodology and approach to gain insight into the problem, and evaluations of several techniques from different but related fields to derive solutions. A practical attitude towards virtual citizen science and the disciplines it involves has influenced the tone of this thesis, rather than following the normal route of a single discipline. Despite this being the case however, the science produced is the main focus. VCS platforms are characterised by the data they produce and how it contributes to furthering the knowledge of the discipline. Therefore improving these outcomes and understanding how techniques from other disciplines can inform the process is the main goal of the thesis.

However, bringing together the perspectives and practices of multiple fields is challenging. Despite the techniques and methodologies that this thesis has adopted being from closely related disciplines, they are not bespoke to the VCS case. Therefore their findings and conclusions are from varying perspectives serving the needs of differing interests regarding the outcomes. Ergo, these perspectives may not necessarily be congruous to those of the stakeholders of virtual citizen science. Acknowledging this

position, the thesis attempts to highlight when such tensions exist where the empirical findings and the related disciplines collide. For instance, regarding the autonomy afforded to the volunteer this work has revealed no significant improvement in performance, a relationship that would have been expected based on the findings of related human factors research. Nonetheless, the justification lies in the practical applicability of the thesis and the requirement to identify task workflow design manipulations that can be influenced at the development stage of a VCS project. In doing so, it suggests a framework that simplifies task workflow design into its constituent components – to inform the engineering of virtual citizen science platforms to ensure the best possible scientific outcomes.

Taking such an approach inevitably has led to various limitations, as described in the following.

11.3.1 Limitations of the Work

Limitations are caused by a number of different agents in this thesis, which are considered thusly.

To begin, the conceptual model of VCS task workflow design factors and the resulting predictive model derived in the discussion chapter can only be used to inform a systems' design when the trade-offs that exist are acknowledged. As was argued in the discussion, by manipulating certain design-influenced TWD factors in order to prioritise a specific scientific outcome, this will invariably have an influence on other outcomes that perhaps are not as favourable. Furthermore, although this thesis has argued that VCS developers have little say over science case-influenced TWD factors their influence still needs to be considered, as well as their interplay with other factors.

Secondly, although the methodological choice to create a purpose-built VCS platform with a specific science case allowed the direct study of TWD factors manipulation, it could arguably make the it difficult to generalise the findings to a wider audience. Whilst the platform created involved the visual analysis of remotely sensed data – a common approach of many current VCS platforms, it is acknowledged that each individual project will have its own nuances and subtleties meaning that not all of the strategies described will be appropriate. However, one of the central aspirations of

this research is that not only does the study of VCS task workflow design need to be realistic and therefore in situ, but the findings need to be actionable at the design stage. Therefore, the argument for going through the entire design process mitigates the need for a wider generalisation.

Finally, the author faced a common problem encountered by many studies of HCI & crowd systems: The synthesis of design recommendations from distinctively measured empirical results. This complication is perhaps amplified by the collision of the different disciplines involved. Whilst the scientists of VCS platforms would be informed by a significant quantitative effect on the results produced, the developers are interested in implications for design practices and promoters concerned with the qualitative findings and how volunteer engagement and motivation might be affected.

11.3.2 Reflection on Task Workflow Design for VCS Systems

This section presents design guidelines derived from the author's experience of devising and studying the manipulation of task workflow design in a real-world VCS application. They are framed in order to take into consideration the perspectives of different VCS stakeholders, and formulated though a subjective review of the empirical findings. As such the guidelines epitomise subjective interpretations, and should not be mistaken as aspiring to a maximum objectivity suitable for all citizen science examples.

The design guidelines may be considered as exercises conducted in a loosely sequential order:

1. *Consider the public reaction* to the science case you are presenting, what you are asking them to contribute to and the level of media attention this might receive. In essence, try and place the project as one of the scenarios in the predictive model (figure 10.2) – in order to gauge realistic expectations.
2. *Consider the data needs* in order to achieve the projects' scientific aims – in terms of its volume, accuracy and granularity. This follows the findings of experiments 1 & 2, showing that task workflow design factors can influence the data collected in differing ways.

3. *Consider if the project is realistic* in terms of the compatibility of public reaction and data needs. For instance, does a task workflow design configuration that will produce the accuracy required also produce enough classifications over time based on your expected visitor size? If not, can a compromise be reached?
4. *Conduct a laboratory trial* of the platform, using non-expert participants. This follows the findings of experiment 1 (chapter 8) that revealed limitations of the system in terms of both its usability and the data collected that might not be apparent to an expert view. By doing this, many issues can be addressed before the platform is released en-masse – and before interface design changes can be an interruption.
5. *Monitor volunteer behaviour* and the data produced regularly. As revealed by their own opinions in the empirical part of this thesis, their preferences can change regarding task workflow design depending on their time on the platform. Whilst new users might prefer more guidance, experienced users might want a greater challenge and more freedom. Identifying when this change occurs and reacting to it in terms of the presented task workflow design could prolong their engagement with the platform, and improve the data they produce.
6. *Acknowledge failure is inevitable*, at least to some degree, and that crowd-sourced information will never be 100% accurate. Virtual citizen science is not conducted in a controlled environment, and therefore a degree of 'noise' will always be present. Consider how the platform will deal with erroneous, null or biased data.

The guidelines were derived from the experience gained through the process of conducting the work for and writing of this thesis. As such, they are unlikely to be complete or comprehensive, or suitable for 100% of virtual citizen science endeavours. As explained in the discussion of implications (see 10.3), the consideration of task workflow design and its factors is a complex undertaking, and its design should be as context-specific as reasonably possible in order to achieve its scientific goals.

11.4 FUTURE WORK

Expanding on the thesis' contribution, its impact has the potential to be expanded into future work in a number of different ways.

In terms of the methodology, making the analogy between VCS platforms and other types of work and adapting the associated research, has the potential to be applied to other crowd-based, HCI systems. Specifically, the application of the TWD model in conjunction with a hierarchical task analysis approach could well assist with identifying design opportunities for performance gain – both in terms of output and user motivation. For example, knowledge transfer opportunities exist to develop the approach to the domain of Mechanical Turk type online practices. Mechanical Turk 'Requesters' could use both the TWD model and the derived guidelines to adapt the tasks they present to the potential crowd workforce to improve both motivation and performance.

At the same time as the empirical work of the thesis has validated the task workflow design model and the categorisation of its constituent factors, it has also highlighted the complexity regarding their interplay and influence. In order to better understand these complexities it could be that further work is needed to create a more detailed model, one that not only represents the different TWD factors at play but also identifies interactions and their summative effects. To achieve this not only must the manipulation of single TWD factors be studied, but also more thought should be given to how they are combined in their configuration.

Future directions could also broaden the thesis' contribution within the citizen science field itself. For instance, although this work has virtual citizen science platforms as its focus, other platforms exist requiring the volunteer to act as a data collector, going out into the real world and collecting data points through various connected devices (see 2.2). Therefore, further work would be needed to expand the TWD model by identifying extra constituent factors specifically related to this endeavour (i.e. geographic context – contributing to the welfare of the community etc.) in order to widen its applicability. Similarly, whilst this work concentrates on the model and its factors' effect on specific scientific outcomes (see 4.5.3) future research is required to understand the models' application

regarding other increasingly influential considerations such as outreach, educational benefits and the ability to make serendipitous discoveries.

This thesis' findings, and the challenges it has revealed and discussed foretell the issues that virtual citizen science stakeholders are increasingly going to have to address on a much larger scale in the future. As the application of VCS platforms covers a broader range of disciplines, involving greater amounts of data and becoming more complex, the importance of the role of the volunteer and how it is presented to them will become increasingly prevalent to the ultimate success of a project.

As such, this thesis is only a first step in the research and development of task workflow design and its influence, with much more work required in order to ensure virtual citizen science projects both assist scientists in furthering their understanding and are rewarding to the volunteer whose commitment is required.

12

REFERENCES

Adamczyk, P. D., & Bailey, B. P. (2004). If Not Now, when?: The Effects of Interruption at Different Moments Within Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 271–278). New York, NY, USA: ACM.

Anderson, D. P., Cobb, J., Korpela, E., Lebofsky, M., & Werthimer, D. (2002). SETI@Home: An Experiment in Public-resource Computing. *Commun. ACM*, *45*(11), 56–61.

Anderson, K. (2005). *Predicting the Weather: Victorians and the Science of Meteorology*. Chicago, Illinois: University of Chicago Press.

Andrews, L. C. (2004). *Field Guide to Atmospheric Optics*. Interntional Society for Optics and Photonics.

Angus, R. (2014, January 16). Stargazing Live. Retrieved from http://astrobites.org/2014/01/16/stargazing-live/

Annett, J., & Duncan, K. D. (1967). Task Analysis and Training Design. *Occupational Psychology*, *41*, 211–221.

Annett, J., Duncan, K. D., Stammers, R. B., & Gray, M. J. (1971). Task Analysis. *Department of Employment Training Information Paper*, *6*.

Antonenko, I., Robbins, S. J., Gay, P. L., Lehan, C., & Moore, J. (2013). Effects of Incidence Angle on Crater Detection and the Lunar Ioschron System: Preliminary Results from the CosmoQuest MoonMappers Citizen Science Project (Vol. 44, p. 2705). Presented at the Lunar and Planetary Science Conference.

Arcanjo, J. S., Da Luz, E. F. P., Fazenda, A. L., & Ramos, F. M. (2014). Evaluating Volunteers' Contributions in a Citizen Science Project. In *2014 IEEE 10th International Conference on e-Science (e-Science)* (Vol. 1, pp. 21–28).

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Barlow, K. E., Briggs, P. A., Haysom, K. A., Hutson, A. M., Lechiara, N. L., Racey, P. A., … Langton, S. D. (2015). Citizen science reveals trends in bat populations: The National Bat Monitoring Programme in Great Britain. *Biological Conservation*, *182*, 14–26.

Barnard, L., Scott, C., Owens, M., Lockwood, M., Tucker-Hood, K., Thomas, S., … Harder, B. (2014). The Solar Stormwatch CME catalogue: Results from the first space weather citizen science project. *Space Weather*, 1–18.

Battersby, S. E., Hodgson, M. E., & Wang, J. (2012). Spatial Resolution Imagery Requirements for Identifying Structure Damage in a Hurricane Disaster: A Cognitive Approach. *Photogrammetric Engineering and Remote Sensing*, *78*(6).

Berardi, A., Raja Parasuraman, James V.Haxby. (2001). Overall Vigilance and Sustained Attention Decrements in Healthy Aging. *Experimental Aging Research*, *27*(1), 19–39.

Berman, D. C., & Hartmann, W. K. (2002). Recent Fluvial, Volcanic, and Tectonic Activity on the Cerberus Plains of Mars. *Icarus*, *159*(1), 1–17.

Beutel, J., Kundel, H. L., & Van Metter, R. L. (2000). *Handbook of Medical Imaging* (Vol. 1). Bellingham, WA: SPIE.

Bolus, R., & Bruzewicz, A. (2002). Evaluation of New Sensors for Emergency Management. *Technical Report ERDC/CRREL*.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next Steps for Citizen Science. *Science*, *343*, 1436–1437.

Borden, K. A., Whyte, L. F., Smith, A., Tarnoff, A., & Schmitt, H. (2012). The Zooniverse: Cutting Edge Scientific Research in the Classroom. *AGU Fall Meeting Abstracts*, *14*.

Boring, E. G. (1942). *Sensation and Perception in the History of Experimental Psychology*. New York: Irvington Publishers.

Bowser, A., Hansen, D., & Preece, J. (2013). Gamifying Citizen Science: Lessons and Future Directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France.

Boyajian, T. S., LaCourse, D. M., Rappaport, S. A., Fabrycky, D., Fischer, D. A., Gandolfi, D., … Szewczyk, A. (2016). Planet Hunters IX. KIC 8462852 - where's the flux? *Monthly Notices of the Royal Astronomical Society*, *457*, 3988–4004.

Braddy, P. W., Thompson, L. F., Wuensch, K. L., & Grossnickle, W. F. (2003). Internet Recruiting The Effects of Web Page Design Features. *Social Science Computer Review*, *21*(3), 374–385.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, *62*(2), 161–178.

Britton, S. A., & Tippins, D. J. (2015). Teaching with Citizen Science—It's More than Just Putting Out Fires! In M. P. Mueller & D. J. Tippins (Eds.), *EcoJustice, Citizen Science and Youth Activism* (pp. 207–222). Springer International Publishing.

Brown, S., Gay, P. L., & Daus, C. S. (2011). Motivation of Citizen Scientists Participating in Moon Zoo (Vol. 217, p. 15814). Presented at the Bulletin of the American Astronomical Society.

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

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., … Vandenberg, J. (2009). Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, *399*, 1191–1205.

Carifio, J., & Perla, R. J. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2014). Modeling false positive detections in species occurrence data under different study designs. *Ecology*.

Chargaff, E. (1978). *Heraclitean fire: sketches from a life before nature*. New York: Rockefeller University 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* (pp. 4061–4064). New York, NY, USA: ACM.

Christian, C., Lintott, C., Smith, A., Fortson, L., & Bamford, S. (2012). Citizen Science: Contributions to Astronomy Research. *arXiv:1202.2577 [Astro-Ph]*.

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*(3), 237–251.

Cormick, C., Nielssen, O., Ashworth, P., Salle, J. L., & Saab, C. (2014). What Do Science Communicators Talk About When They Talk About Science Communications? Engaging With the Engagers. *Science Communication*, 1075547014560829.

Cormier, S., Prather, E. E., Brissenden, G., Lintott, C., Gay, P. L., Raddick, J., & Collaboration of Astronomy Teaching Scholars CATS. (2011). Assessment Of The Effect Of Participation In Zooniverse Projects On Content Knowledge (Vol. 217, p. 247.05). Presented at the Bulletin of the American Astronomical Society.

Cowen, D. J., & Jensen, J. R. (1998). Extraction and Modeling of Urban Attributes Using Remote Sensing Technology. In *People and Pixels: Linking Remote Sensing and Social Science*. Washington D.C.: National Academy Press.

Crowston, K., & Fagnot, I. (2008). The Motivational Arc of Massive Virtual Collaboration. *Proceedings of the IFIP WG 9.5 Working Conference on Virtuality and Society: Massive Virtual Communities*.

Curran, T., Gibson, L., Horne, J. H., Young, B., & Bozell, A. P. (2009). Expert image analysts show enhanced visual processing in change detection. *Psychonomic Bulletin & Review*, *16*(2), 390–397.

Curtis, V. (2014). Online citizen science games: Opportunities for the biological sciences. *Applied & Translational Genomics*, *3*(4), 90–94.

Dai, P., Mausam, D., & Weld, D. (2010). Decision-theoretic control of crowd-sourced workflows. *Proceedings of 24th AAAI Conference on Artificial Intelligence*.

Daubar, I. J., McEwen, A. S., Byrne, S., Kennedy, M. R., & Ivanov, B. (2013). The current martian cratering rate. *Icarus*, *225*(1), 506–516.

Davidson, R., & Henderson, R. (2000). Electronic Performance Monitoring: A Laboratory Investigation of the Influence of Monitoring and Difficulty on Task Performance, Mood State, and Self-Reported Stress Levels. *Journal of Applied Social Psychology*, *30*(5), 906–920.

Davis, J., Arderiu, J., Lin, H., Nevins, Z., Schuon, S., Gallo, O., & Yang, M. H. (2010). The HPU. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (pp. 9–16).

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* (pp. 2425–2428). New York, NY, USA: ACM.

Di, K., Liu, Y., Hu, W., Yue, Z., & Liu, Z. (2014). Mars Surface Change Detection from Multi-temporal Orbital Images. *IOP Conference Series: Earth and*

Environmental Science, *17*(1), 12015. https://doi.org/10.1088/1755-1315/17/1/012015

Diaper, D., & Stanton, N. (2003). *The Handbook of Task Analysis for Human-Computer Interaction*. CRC Press.

Dickinson, J. L., & Bonney, R. (2012). *Citizen Science: Public Participation in Environmental Research*. Cornell University Press.

Ding, W., Stepinski, T. F., Bandeira, L., Vilalta, R., Wu, Y., Lu, Z., & Cao, T. (2010). Automatic Detection of Craters in Planetary Images: An Embedded Framework Using Feature Selection and Boosting. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 749–758). New York, NY, USA: ACM. https://doi.org/10.1145/1871437.1871534

Dodd, N. G., & Ganster, D. C. (1996). The interactive effects of variety, autonomy, and feedback on attitudes and performance. *Journal of Organizational Behavior*, *17*(4), 329–347.

Dourish, P. (2006). Implications for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 541–550). New York, NY, USA: ACM.

Dubinsky, A. J., & Skinner, S. J. (1984). Impact of job characteristics on retail salespeople's reactions to their jobs. *Journal of Retailing*, *60*(2), 35–62.

Durso, F. T. (Ed.). (2007). Handbook of Applied Cognition. In *Handbook of Applied Cognition* (pp. i–xxvi). John Wiley & Sons Ltd.

Dutta, P., Aoki, P. M., Kumar, N., Mainwaring, A., Myers, C., Willett, W., & Woodruff, A. (2009). Common Sense: Participatory Urban Sensing Using a Network of Handheld Air Quality Monitors. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems* (pp. 349–350). New York, NY, USA: ACM.

Ede, A., & Cormack, L. B. (2012). *A History of Science in Society: From Philosophy to Utility* (2nd ed.). North York, Canada: University of Toronto Press.

Elliott, D. B., Whitaker, D., & Bonette, L. (1990). Differences in the legibility of letters at contrast threshold using the Pelli-Robson chart. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, *10*(4), 323–326.

Essock, E. A., Sinai, M. J., DeFord, K., Hansen, B. C., & Srinivasan, N. (2004). Human perceptual performance with nonliteral imagery: region recognition and texture-based segmentation. *Journal of Experimental Psychology. Applied*, *10*(2), 97–110.

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* (pp. 2985–2994). New York, NY, USA: ACM.

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* (pp. 79–82). New York, NY, USA: ACM.

Fassett, C. I., Levy, J. S., Dickson, J. L., & Head, J. W. (2014). An extended period of episodic northern mid-latitude glaciation on Mars during the Middle to Late Amazonian: Implications for long-term obliquity history. *Geology*, G35798.1. https://doi.org/10.1130/G35798.1

Fechner, G. (1966). *Elements of psychophysics. Vol. I*. New York: Holt, Rinehart and Winston.

Feyerabend, P. (1982). *Science in a free society*. London: New Left Books.

Firehock, K., & West, J. (1995). A Brief History of Volunteer Biological Water Monitoring Using Macroinvertebrates. *Journal of the North American Benthological Society*, *14*(1), 197–202.

Fischer, D. A., Schwamb, M. E., Schawinski, K., Lintott, C., Brewer, J., Giguere, M., … Zimmermann, V. (2012). Planet Hunters: the first two planet candidates identified by the public using the Kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, *419*, 2900–2911.

Fitts, P. M. (Ed). (1951). Human Engineering for an Effective Air Navigation and Traffic Control System. *National Research Council*.

Follett, R., & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PLOS ONE*, *10*(11), e0143687. https://doi.org/10.1371/journal.pone.0143687

Forbes, D. A. (2008). So you want to be a professional astronomer! *arXiv:0805.2624 [Astro-Ph]*.

Fortson, L., & Lynn, S. (2014). Talking in the Zooniverse: A collaborative tool for citizen scientists. In *2014 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 1–2).

Fortson, L., Masters, K., Nichol, R., Borne, K. D., Edmondson, E., Lintott, C., … Wallin, J. (2012). Galaxy Zoo: Morphological Classification and Citizen Science. In *Advances in Machine Learning and Data Mining for Astronomy*. Boca Raton, FL.: CRC Press.

Freitag, A. (2016). A typology for strategies to connect citizen science and management. *Environmental Monitoring and Assessment*, *188*(9), 519.

Freitag, A., Meyer, R., & Whiteman, L. (2016). Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. *Citizen Science: Theory and Practice.*, *1*(1).

Garg, R. K. (1996). The influence of positive and negative wording and issue involvement on responses to Likert scales in marketing research. *Journal of the Market Research Society*, *38*(3), 235–246.

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*(3), 366–373.

Gescheider, G. A. (1985). Psychophysics: Methods, Theory, and Application (2nd ed., pp. 174–191). Hillsdale, New Jersey: Erlbaum and Associates.

Gescheider, G. A. (1997). *Psychophysics: The Fundamentals*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ghani, J. A., & Deshpande, S. P. (1994). Task Characteristics and the Experience of Optimal Flow in Human—Computer Interaction. *The Journal of Psychology*, *128*(4), 381–391.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory*. Chicago: Aldine.

Gollan, J. (2013). Citizen science can produce reliable data. Retrieved 18 August 2016, from http://theconversation.com/citizen-science-can-produce-reliable-data-10815

Greenhill, A., Holmes, K., Lintott, C., Simmons, B., Masters, K., Cox, J., & Graham, G. (2014). Playing with science : gamised aspects of gamification found on the Online Citizen Science Project - Zooniverse. In *GAME-ON 2014 :* (pp. 15–24). Lincoln: EUROSIS.

Grier, D. A. (2005). *When Computers Were Human*. Princeton, New Jersey: Princeton University Press.

Hackman, J. R. (1980). Work redesign and motivation. *Professional Psychology*, *11*(3), 445–455.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, *60*(2), 159–170.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, *17*(2–3), 107–145.

Hand, E. (2010). Citizen science: People power. *Nature News*, *466*(7307), 685–687.

Hansen, B. C., & Essock, E. A. (2012). The Relationship Between Human Perceptual Performance and the Physical Attributes of Night Vision Imagery. In *Trends in Experimental Psychology Research*. New York: Nova Science Publishers.

Hardy, J. W. (1998). *Adaptive Optics for Astronomical Telescopes*. Oxford University Press.

Harris, H. E., & Russo, P. (2014). The influence of social movements on space astronomy policy: The cases of 'Hubble Huggers', JWST's 'Science Warriors' and the ISEE-3 'Reboot Team'. *Space Policy*, 1–4.

Harrower, M., & Sheesley, B. (2005). Designing Better Map Interfaces: A Framework for Panning and Zooming. *Transactions in GIS*, *9*(2), 77–89.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

Hartmann, W. K. (1999). Martian cratering VI: Crater count isochrons and evidence for recent volcanism from Mars Global Surveyor. *Meteoritics & Planetary Science*, *34*(2), 167–177.

Hartmann, W. K. (2005). Martian cratering 8: Isochron refinement and the chronology of Mars. *Icarus*, *174*(2), 294–320.

Hawthorne, T. L., Elmore, V., Strong, A., Bennett-Martin, P., Finnie, J., Parkman, J., … Reed, J. (2015). Mapping non-native invasive species and accessibility in an urban forest: A case study of participatory mapping and citizen science in Atlanta, Georgia. *Applied Geography*, *56*, 187–198.

Hay, G. J., Niemann, K. O., & McLean, G. F. (1996). An object-specific image-texture analysis of H-resolution forest imagery. *Remote Sensing of Environment*, *55*(2), 108–122.

Heaton, L., Millerand, F., Liu, X., & Crespel, É. (2014). Participatory Science: Encouraging public engagement in ONEM. *International Journal of Science Education, Part B*, *0*(0), 1–22.

Hennon, C. C., Knapp, K. R., Schreck, C. J., Stevens, S. E., Kossin, J. P., Thorne, P. W., … Carley, I. (2014). Cyclone Center: Can Citizen Scientists Improve Tropical Cyclone Intensity Records? *Bulletin of the American Meteorological Society*.

Hodgson, M. E. (1998). What Size Window for Image Classification? A Cognitive Perspective. *Photogrammetric Engineering & Remote Sensing*, *64*(8), 797–807.

Hoffman, R. R., Markman, A., & Carnahan, W. H. (2001). Angles of Regard: Psychology Meets Technology in the Perception and Interpretation of Nonliteral Imagery. In *Interpreting Remote Sensing Imagery: Human Factors*. New York: Lewis Publishers.

Hollnagel, E. (2003). *Handbook of Cognitive Task Design*. CRC Press.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, *14.06*.

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (1st ed.). New York, NY, USA: Crown Publishing Group.

Huang, C.-Y. (2002). Distributed manufacturing execution systems: A workflow perspective. *Journal of Intelligent Manufacturing*, *13*(6), 485–497.

Hueso, R., Legarreta, J., Pérez-Hoyos, S., Rojas, J. F., Sánchez-Lavega, A., & Morgado, A. (2010). The international outer planets watch atmospheres node database of giant-planet images. *Planetary and Space Science*, *58*(10), 1152–1159.

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)* (pp. 136–143). https://doi.org/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* (pp. 1101–1106). New York, NY, USA: ACM.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371.

Jackson, C. B., Osterlund, C., Mugar, G., Hassman, K. D. V., & Crowston, K. (2015). Motivations for Sustained Participation in Crowdsourcing: Case Studies of Citizen Science on the Role of Talk. *48th Hawaii International Conference on System Sciences*, 1624–1634.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 13.

Jamieson, G. A., Miller, C. A., Ho, W. H., & Vicente, K. J. (2007). Integrating Task- and Work Domain-Based Work Analyses in Ecological Interface Design: A Process Control Case Study. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *37*(6), 887–905.

Jamieson, S. (2004). Likert Scales: how to (ab)use them. *Medical Education*, *38*, 1212–1218.

Joost, S., Baumann, R., Ertz, O., Ingensand, J., Widmer, I., & Rappo, D. (2014). A participatory WebGIS platform to support biodiversity inventory in the Geneva cross-border area. *Proceedings of the 3rd Open Source Geospatial Research & Education Symposium OGRS 2014*, *3*, 143–148.

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

Joy, K., Crawford, I., Grindrod, P., Lintott, C., Bamford, S., Smith, A., … Zoo, M. (2011). Moon Zoo: citizen science in lunar exploration. *Astronomy & Geophysics*, *52*(2), 2.10-2.12.

Kaufman, A. (2014). Let's Talk about Citizen Science: What Doesn't Work. *Animal Behavior and Cognition*, *1*(4), 446–451.

Kim, J. S., & Hamner, W. C. (1976). Effect of performance feedback and goal setting on productivity and satisfaction in an organizational setting. *Journal of Applied Psychology*, *61*(1), 48–57.

Kintisch, E. (2011). How to Grow Your Own Army of Citizen Scientists [Text]. Retrieved 18 May 2015, from http://news.sciencemag.org/space/2011/02/how-grow-your-own-army-citizen-scientists

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013). The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 1301–1318). New York, NY, USA: ACM.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Knapp, T. R. (1990). Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy. *Nursing Research*, *39*, 121–123.

Knaving, K., & Björk, S. (2013). Designing for Fun and Play: Exploring Possibilities in Design for Gamification. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 131–134). New York, NY, USA: ACM.

Koller, S. M., Drury, C. G., & Schwaninger, A. (2009). Change of search time and non-search time in X-ray baggage screening due to training. *Ergonomics*, *52*(6), 644–656.

Lagoze, C. (2014). eBird: Curating Citizen Science Data for Use by Diverse Communities. *International Journal of Digital Curation*, *9*(1), 71–82.

Lahav, O., Naim, A., Buta, R. J., Corwin, H. G., de Vaucouleurs, G., Dressler, A., … Storrie-Lombardi, M. C. (1995). Galaxies, human eyes, and artificial neural networks. *Science (New York, N.Y.)*, *267*(5199), 859–862.

Lasecki, W. S., Marcus, A., Rzeszotarski, J. M., & Bigham, J. P. (2014). Using Microtask Continuity to Improve Crowdsourcing. *Carnegie Mellon University Human-Computer Interaction Institute - Technical Reports*, *CMU-HCII-14-100*.

Lazar, D. J., Feng, D. J. H., & Hochheiser, D. H. (2010). *Research Methods in Human-Computer Interaction*. John Wiley & Sons.

Leckie, D. G., Gougeon, F. A., Walsworth, N., & Paradine, D. (2003). Stand delineation and composition estimation using semi-automated individual tree crown analysis. *Remote Sensing of Environment*, *85*(3), 355–369. https://doi.org/10.1016/S0034-4257(03)00013-0

Lintott, C. J., Schawinski, K., Keel, W., van Arkel, H., Bennert, N., Edmondson, E., … Vandenberg, J. (2009). Galaxy Zoo: `Hanny's Voorwerp', a quasar light echo? *Monthly Notices of the Royal Astronomical Society*, *399*, 129–140.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., … Vandenberg, J. (2011). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, *410*(1), 166–178.

Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2010). TurKit: Human Computation Algorithms on Mechanical Turk. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology* (pp. 57–66). New York, NY, USA: ACM.

Lohmann, A. W., Weigelt, G., & Wirnitzer, B. (1983). Speckle masking in astronomy: triple correlation theory and applications. *Applied Optics*, *22*(24), 4028.

Luczak-Roesch, M., Tinati, R., Simperl, E., van Kleek, M., Shadbolt, N., & Simpson, R. (2014). Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Ma, H., Chandrasekar, R., Quirk, C., & Gupta, A. (2009). Improving Search Engines Using Human Computation Games. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 275–284). New York, NY, USA: ACM.

Mackworth, N. H. (1948). The Breakdown of Vigilance During Prolonged Visual Search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21.

Macmillan, N. A., & Creelman, C. D. (1991). *New Developments in Detection Theory*. Cambridge, UK: Cambridge University Press.

Mankowski, T. A., Slater, S. J., & Slater, T. F. (2011). An Interpretive Study Of Meanings Citizen Scientists Make When Participating In Galaxy Zoo. *Contemporary Issues in Education Research (CIER)*, *4*(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*.

Marks, L. E. (1988). Magnitude estimation and sensory matching. *Perception & Psychophysics*, *43*(6), 511–525.

Marshall, P. J., Lintott, C. J., & Fletcher, L. N. (2015). Ideas for Citizen Science in Astronomy. *Annual Review of Astronomy and Astrophysics*, *53*(1), 247–278. https://doi.org/10.1146/annurev-astro-081913-035959

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual Skills in Airport-Security Screening. *Psychological Science*, *15*(5), 302–306.

McGill, G. E. (1977). Craters as 'fossils': The remote dating of planetary surface materials. *Geological Society of America Bulletin*, *88*(8), 1102–1110.

McGregor, D. (1960). *The Human Side of Enterprise*. New York: McGraw-Hill.

Michael, G. G., & Neukum, G. (2010). Planetary surface dating from crater size–frequency distribution measurements: Partial resurfacing events and statistical age uncertainty. *Earth and Planetary Science Letters*, *294*(3–4), 223–229.

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285–290. https://doi.org/10.1890/110278

Mojsilovic, A., Hu, J., & Soljanin, E. (2002). Extraction of perceptually important colors and similarity measurement for image matching, retrieval and analysis. *IEEE Transactions on Image Processing*, *11*(11), 1238–1248.

Muratet, A., & Fontaine, B. (2015). Contrasting impacts of pesticides on butterflies and bumblebees in private gardens in France. *Biological Conservation*, *182*, 148–154.

Murray, J. B., Muller, J.-P., Neukum, G., Werner, S. C., van Gasselt, S., Hauber, E., … Team, T. H. C.-I. (2005). Evidence from the Mars Express High Resolution Stereo Camera for a frozen sea close to Mars' equator. *Nature*, *434*(7031), 352–356.

Nerbonne, J. F., & Vondracek, B. (2003). Volunteer macroinvertebrate monitoring: Assessing training needs through examining error and bias in untrained volunteers. *Journal of the North American Benthological Society*, *22*(1), 152–163.

Nevin, J. A. (1969). Signal Detection Theory and Operant Behavior: A Review of David M. Green and John A. Swets' Signal Detection Theory and Psychophysics.1. *Journal of the Experimental Analysis of Behavior*, *12*(3), 475–480.

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304.

Nov, O., Arazy, O., & Anderson, D. (2011a). Dusting for Science: Motivation and Participation of Digital Citizen Science Volunteers. In *Proceedings of the 2011 iConference* (pp. 68–74). New York, NY, USA: ACM.

Nov, O., Arazy, O., & Anderson, D. (2011b). Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceedings of the AAAI International Conference on Weblogs and Social 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. *Journal of Organizational Behavior*, *31*(2–3), 463–479.

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance* (pp. 1–39). Oxford, England: John Wiley & Sons.

Parasuraman, R., & Giambra, L. (1991). Skill development in vigilance: Effects of event rate and age. *Psychology and Aging*, *6*(2), 155–169.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America. A, Optics and Image Science*, *2*(9), 1508–1532.

Pelli, D. G., & Farell, B. (2010). Psychophysical Methods. In *Handbook of Optics* (3rd ed., Vol. 3, p. 3.1-3.12). New York: McGraw-Hill.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212. https://doi.org/10.1109/TIT.1954.1057460

Pinkowski, J. (2010). How to Classify a Million Galaxies in Three Weeks. *Time*. Retrieved from http://content.time.com/time/health/article/0,8599,1975296,00.html

Ponciano, L., Brasileiro, F., Simpson, R., & Smith, A. (2014). Volunteers' Engagement in Human Computation for Astronomy Projects. *Computing in Science & Engineering*, *16*(6), 52–59.

Powell, R. L., Matzke, N., de Souza Jr., C., Clark, M., Numata, I., Hess, L. L., & Roberts, D. A. (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sensing of Environment*, *90*(2), 221–234.

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. *Astronomy Education Review*, *12*(1).

Prestopnik, N. R., & Crowston, K. (2011). Gaming for (Citizen) Science: Exploring Motivation and Data Quality in the Context of Crowdsourced Science through the Design and Evaluation of a Social-Computational System. In *2011 IEEE Seventh International Conference on e-Science Workshops (eScienceW)* (pp. 28–33).

Prestopnik, N. R., & Crowston, K. (2012). Citizen Science System Assemblages: Understanding the Technologies That Support Crowdsourced Science. In *Proceedings of the 2012 iConference* (pp. 168–176). New York, NY, USA: ACM.

Price, H. E. (1985). The Allocation of Functions in Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *27*(1), 33–45.

Raddick, J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., … Vandenberg, J. (2013). Galaxy Zoo: Motivations of Citizen Scientists. *Astronomy Education Review*, *12*(1).

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., … Vandenberg, J. (2009). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *arXiv:0909.2925 [Astro-Ph, Physics:physics]*.

Reason, J. (1987). Cognitive aids in process environments: prostheses or tools? *International Journal of Man-Machine Studies*, *27*(5), 463–470.

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)* (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* (pp. 623–625). New York, NY, USA: ACM.

Reed, J. T., Cook, R., Raddick, M. J., Carney, K., & Lintott, C. (2013). Participating in Online Citizen Science: Motivations as the Basis for User Types and Trajectories. In P. Michelucci (Ed.), *Handbook of Human Computation* (pp. 695–702). Springer New York.

Reeves, S., & Sherwood, S. (2010). Five Design Challenges for Human Computation. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 383–392). New York, NY, USA: ACM.

Reips, U.-D. (2000). The Web Experiment Method: Advantages, Disadvantages, and Solutions. In *Psychological Experiments on the Internet* (pp. 89–117). San Diego, California: Academic Press.

Robbins, S. J., Antonenko, I., Kirchoff, M. R., Chapman, C. R., Fassett, C. I., Herrick, R. R., … Gay, P. L. (2014). The variability of crater identification among expert and community crater analysts. *Icarus*, *234*, 109–131.

Rogowitz, B. E., Frese, T., Smith, J. R., Bouman, C. A., & Kalin, E. B. (1998). Perceptual image similarity experiments (Vol. 3299, pp. 576–590). Presented at the Human Vision and Electronic Imaging III.

Rojas, D., Kapralos, B., & Dubrowski, A. (2013). The Missing Piece in the Gamification Puzzle. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 135–138). New York, NY, USA: ACM.

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

Rzeszotarski, J. M., Chi, E., Paritosh, P., & Dai, P. (2013). Inserting Micro-Breaks into Crowdsourcing Workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Sarmento, P., Carrao, H., Caetano, M., & Stehman, S. V. (2009). Incorporating reference classification uncertainty into the analysis of land cover accuracy. *International Journal of Remote Sensing*, *30*(20).

Scepan, J., Menz, G., & Hansen, M. C. (1999). The DISCover validation image interpretation process. *Photogrammetric Engineering and Remote Sensing*, *65*(9), 1075–1081.

Schmidt, L. (2010). Crowdsourcing for Human Subjects Research. *Crowdconf*.

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

Schweitzer, B., & McLeod, B. (1997). Marketing Technology that is Changing at the Speed of Light. *Earth Observation Magazine*, *6*(7), 22–24.

Scollon, C. N., Prieto, C.-K., & Diener, E. (2009). Experience Sampling: Promises and Pitfalls, Strength and Weaknesses. In P. E. Diener (Ed.), *Assessing Well-Being* (pp. 157–180). Springer Netherlands.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., … Obersteiner, M. (2013). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS ONE*, *8*(7), e69958.

Sharples, S., & Houghton, R. J. (2016). The field becomes the laboratory? The impact of the contextual digital footprint on the discipline of E/HF. *Ergonomics*, *0*(0), 1–14.

Shaw, T. H., Matthews, G., Warm, J. S., Finomore, V. S., Silverman, L., & Costa Jr., P. T. (2010). Individual differences in vigilance: Personality, ability and states of stress. *Journal of Research in Personality*, *44*(3), 297–308.

Sheridan, T. B., Cott, H. P. V., Woods, D. D., Pew, R. W., & Hancock, P. A. (1998). Allocating Functions Rationally between Humans and Machines. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *6*(3), 20–25.

Sidiropoulos, P., & Muller, J.-P. (2014). Temporal Analysis of all High-Resolution Mars Imaging Products. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XL-4*.

Simpson, R. J., Povich, M. S., Kendrew, S., Lintott, C. J., Bressert, E., Arvidsson, K., … Wolf-Chase, G. (2012). The Milky Way Project First Data Release: a bubblier Galactic disc. *Monthly Notices of the Royal Astronomical Society*, *424*(4), 2442–2460.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the World's Largest Citizen Science Platform. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion* (pp. 1049–1054). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Smith, A. (2013). How the Zooniverse Works: The Domain Model. Retrieved from http://blog.zooniverse.org/2013/06/20/how-the-zooniverse-works-the-domain-model/

Sprinks, J., Houghton, R., Bamford, S., Morley, J. G., & Wardlaw, J. (2015). Is that a Crater? Designing Citizen Science Platforms for the Volunteer and to Improve Results. *European Planetary Science Congress*, *10*(EPSC2015-694).

Stanton, N. A. (2006). Hierarchical task analysis: Developments, applications, and extensions. *Applied Ergonomics*, *37*(1), 55–79.

Statella, T., Pina, P., & da Silva, E. A. (2014). Automated determination of the orientation of dust devil tracks in mars orbiter images. *Advances in Space Research*, *53*(12), 1822–1833. https://doi.org/10.1016/j.asr.2013.05.012

Swanson, A., Kosmala, M., Lintott, C., & Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, *30*(3), 520–531.

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. *Scientific Data*, *2*, 150026.

Szpir, M. (2002). Clickworkers on Mars » American Scientist. Retrieved 12 September 2016, from http://www.americanscientist.org/issues/pub/clickworkers-on-mars

Tarable, A., Nordio, A., Leonardi, E., & Marsan, M. A. (2014). The Importance of Being Earnest in Crowdsourcing Systems. *arXiv:1411.7960 [Cs]*.

Taylor, F. W. (1911). *Principles of Scientific Management*. New York: Harper and Row.

Taylor, J., Teanby, N. A., & Wookey, J. (2013). Estimates of seismic activity in the Cerberus Fossae region of Mars. *Journal of Geophysical Research: Planets*, *118*(12), 2570–2581.

Thomas, J. P. (1985). Detection and identification: how are they related? *Journal of the Optical Society of America. A, Optics and Image Science*, *2*(9), 1457–1467.

Tinati, R., Simperl, E., Luczak-Roesch, M., Van Kleek, M., & Shadbolt, N. (2014). Collective Intelligence in Citizen Science -- A Study of Performers and Talkers. *arXiv:1406.7551 [Physics]*.

Trist, E. L., & Bamforth, K. W. (1951). Some Social and Psychological Consequences of the Longwall Method of Coal-Getting: An Examination of the Psychological Situation and Defences of a Work Group in Relation to the Social Structure and Technological Content of the Work System. *Human Relations*, *4*(3).

Vakharia, D., & Lease, M. (2013). Beyond AMT: An Analysis of Crowd Work Platforms. *arXiv:1310.1672 [Cs]*.

Van Coillie, F. M. B., Gardin, S., Anseel, F., Duyck, W., Verbeke, L. P. C., & De Wulf, R. R. (2014). Variability of operator performance in remote-sensing image interpretation: the importance of human and external factors. *International Journal of Remote Sensing*, *35*(2), 754–778.

Van Den Berg, H., Dann, S. L., & Dirkx, J. M. (2009). Motivations of adults for non-formal conservation education and volun-teerism: implications for programming. *Applied Environmental Education and Communication: An International Journal*, *8*(1), 6–17.

Vetter, J. (2011). Lay Participation in the History of Scientific Observation. In *Science in Context* (2nd ed., Vol. 24, pp. 127–141). Cambridge, UK: Cambridge University Press.

Von Ahn, L. (2005). *Human Computation* (Doctoral Dissertation). Carnegie Mellon University, Pittsburgh, PA, USA.

Ward, D. F. (2014). Understanding sampling and taxonomic biases recorded by citizen scientists. *Journal of Insect Conservation*, *18*(4), 753–756.

Wiggins, A., & Crowston, K. (2010). Developing a conceptual model of virtual organisations for citizen science. *International Journal of Organisational Design and Engineering*, *1*(1), 148–162.

Wiggins, A., & Crowston, K. (2012). Goals and Tasks: Two Typologies of Citizen Science Projects. *45th Hawaii International Conference on System Science*, *HICSS-45*.

Wilk, L. A., & Redmon, W. K. (1997). The Effects of Feedback and Goal Setting on the Productivity and Satisfaction of University Admissions Staff. *Journal of Organizational Behavior Management*, *18*(1), 45–68.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., … Thomas, D. (2013). Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey. *arXiv:1308.3496 [Astro-Ph]*.

Winter, J. C. F. de, & Dodou, D. (2011). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, *16*(1), 1–11.

Zhou, W., Schwarz, K., & Cadenasso, M. L. (2009). Mapping urban landscape heterogeneity: agreement between visual interpretation and digital classification approaches. *Landscape Ecology*, *25*(1), 53–67.

Zhu, Z.-L., Yang, L., Stehman, S. V., & Czaplewski, R. L. (2000). Accuracy Assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region. *Accuracy Assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region*, *66*, 14251438.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors*, *40*(4), 554–568.

A

APPENDIX TO STUDY 1

This appendix includes additional material to Study 1 (see chapter 6):

- Figure A.1 depicts the full hierarchical task analysis for each of the twelve Zooniverse platforms.

In addition to the material shown, a Github repository has been created (github.com/zooniverse/NottMarsZoo) from which extra material and data can be downloaded.

Additional contents available through the thesis Github repository include:

- The source website analytics and HTA data for each of the platforms for the period studied in directory:

  thesis/hierarchical task analysis (chapter 6)/website analytics.csv.

- The Hierarchical Task Analysis diagram (printable on A1): thesis/hierarchical task analysis (chapter 6)/HTA.png.

Hierarchical Task Analysis Diagram to be inserted in final hard-bound thesis. Available on request at: github.com/zooniverse/NottMarsZoo/thesis/hierarchical task analysis (chapter 6)/HTA.png.

B

APPENDIX TO STUDY 2

This appendix includes additional material to study 2 (see chapter 7):

- Figure B.1 shows the online questionnaire.

Additional contents available through the thesis Github repository include:

- Participant response data from the questionnaire in directory: thesis/online questionnaire (chapter 7)/questionnaire results.csv.

PLANET FOUR - USABILITY QUESTIONNAIRE

Q1. Please state your age (years): (Leave blank if you prefer not to answer)

Q2. Please tell us your location: (Leave blank if you prefer not to answer)

Q3. Are you: (Leave blank if you prefer not to answer)

Male

Female

Q4. How did you find out about Planet Four?

- ❑ BBC Stargazing Live programme
- ❑ Zooniverse website / email / newsletter
- ❑ Other website / media article
- ❑ Family / friend/ colleague
- ❑ Another Source (please provide)
- ❑ Decline

Q5. Which other Zooniverse projects have you taken part in?

- ❑ Galaxy Zoo
- ❑ Moon Zoo
- ❑ Solar Stormwatch
- ❑ Planet Hunters
- ❑ Milkyway Project
- ❑ Old Weather
- ❑ Cyclone Center
- ❑ Ancient Lives
- ❑ Whale FM
- ❑ Seafloor Explorer
- ❑ Bat Detective
- ❑ Cell Slider
- ❑ None

Q6. The first set of statements relate to your understanding, motivation and the ease of use of the Planet Four website:

	Strongly Disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I learnt to use the website very quickly	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the website visually attractive to look at	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the website easy to use and access	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q7. Please add any further comments you have regarding the design and usability of the website:

Q8. The next set of statements relate to the task and using of tools for marking craters and features on Mars:

	Very Low 1	2	3	4	5	6	7	8	Very High 9
How mentally demanding was the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How physically demanding was the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How hurried or rushed was the pace of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q9.

	Perfect 1	2	3	4	5	6	7	8	Failure 9
How successful were you in accomplishing the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q10.

	Very Low 1	2	3	4	5	6	7	8	Very High 9
How hard did you have to work to accomplish your level of performance?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How insecure, discouraged, irritated, stressed or annoyed were you?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q11.

	Strongly Disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found the 'Crater' marking tool easy to use	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the 'Elliptical Crater' marking tool easy to use	❑	❑	❑	❑	❑	❑	❑	❑	❑
I would like more tools	❑	❑	❑	❑	❑	❑	❑	❑	❑

to mark different shapes and features									

Q12. What, if anything, would you change about the website tools or images and why?

Q13. The next set of statements relate to the images you are shown on the Planet Four website:

	Strongly disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found it hard to 'pick out' craters and features	❑	❑	❑	❑	❑	❑	❑	❑	❑
I thought the image quality was consistent	❑	❑	❑	❑	❑	❑	❑	❑	❑
There were sometimes too many craters in a single image	❑	❑	❑	❑	❑	❑	❑	❑	❑
I didn't know what to do if an image had no craters or features to mark	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q14. Please add any further comments you have regarding the images shown on the Planet Four website:

Q15. Have you signed up to planetfour.org? (Provided a username and password)

Yes

No

Q16. Did you 'favourite' an object you found interesting or different?

Yes

No

Q17.

	Strongly disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found the 'profile' section useful & interesting	❑	❑	❑	❑	❑	❑	❑	❑	❑
I visited the 'discuss' forum and sometimes contributed	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q18. Did you visit the Planet Four blog?

Yes

No

Q19. Did you find the blog interesting and informative?

	Strongly disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found the Planet Four blog interesting	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the Planet Four blog informative	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q21.

	Strongly disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I feel through the website I have a better understanding of Mars	❑	❑	❑	❑	❑	❑	❑	❑	❑
I feel that through using the website I have contributed to science	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q22 Please add any further comments you have regarding discussion and feedback you would like available.

Figure B.1: Planet Four online questionnaire

C

APPENDIX TO EXPERIMENT 1

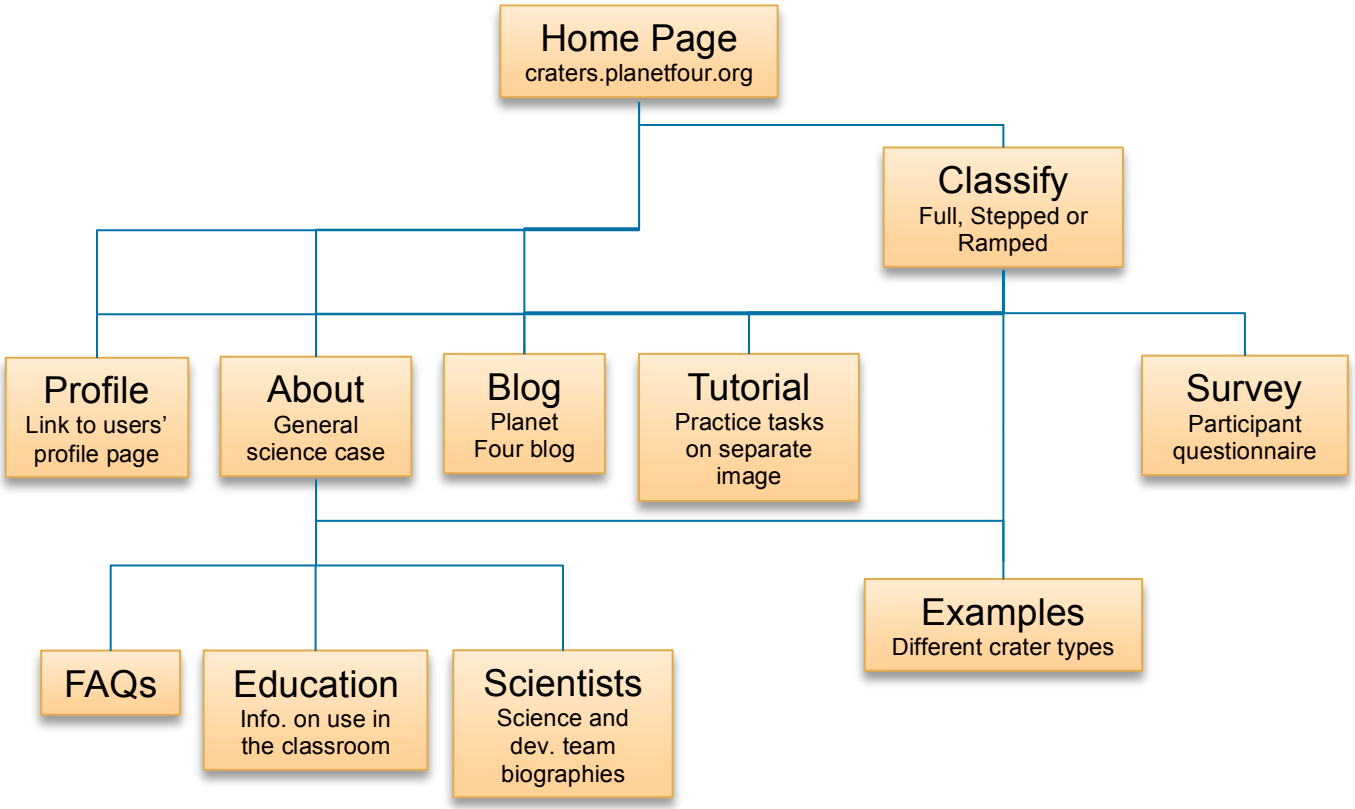
This appendix includes additional material to experiment 1 (see chapter 8):

- Figure C.1 depicts a hierarchical diagram of the Planet Four: Craters page layout.
- Figure C.2 shows the participant questionnaire completed after using each interface.

The Planet Four: Craters platform code is available on request through the Zooniverse Github repository: github.com/zooniverse/NottMarsZoo.

Additional contents available through the thesis Github repository include:

- Participant crater marking data for each interface in directory: thesis/lab experiment (chapter 8)/labcratermarkings.csv.
- Participant position marking data using the Ramped interface: thesis/lab experiment (chapter 8)/labpositionmarkings.csv.
- Expert crater marking data from experts at the University of Bristol: thesis/lab experiment (chapter 8)/expertmarkings.csv.
- Reconstructed CTX image G05_020119_1895_XN_09N198W with participant markings shown (blue – full interface, red – stepped interface and yellow – ramped interface): thesis/lab experiment (chapter 8)/labstudy.jpg.
- Folder of image 'slices' presented to participants through the Planet Four: Craters platform, again showing participant markings colour-coded as previously described: thesis/lab experiment (chapter 8)/image slices.
- Participant questionnaire response data for each interface: thesis/lab experiment (chapter 8)/questionnaire results.csv.



C.1: Planet Four: Craters page hierarchy

Participant Questionnaire - Planet Four

Participant: _____

Age: ☐ **Decline** ☐

Have you used Planet Four before? **Yes** ☐ **No** ☐

Sex: **Male** ☐ **Female** ☐ **Decline** ☐

How did you hear about the Planet Four website? (If you had before this study)

BBC's 'Stargazing Live' Programme ☐ **Zooniverse web/email/newsletter** ☐

Other website / media article ☐ **Family / friend / colleague** ☐

Another source ☐

Which other Zooniverse projects have you taken part in?

Galaxy Zoo ☐ **Moon Zoo** ☐ **Solar Stormwatch** ☐

Planet Hunters ☐ **Milkyway Project** ☐ **Old Weather** ☐

Cyclone Center		**Ancient Lives**		**Whale FM**	
Seafloor Explorer		**Bat Detective**		**Cell Slider**	

Participant Questionnaire - Planet Four

Website Version:

Participant:

The questionnaire should take no more than fifteen minutes to complete.

The first set of statements relate to your understanding, motivation and the ease of use of the Planet Four website:

	Strongly Disagree								**Strongly Agree**
I learnt to use the website very quickly									
	1	2	3	4	5	6	7	8	9
I found the website visually attractive to look at									
	1	2	3	4	5	6	7	8	9

I found the website easy to use and access									
	1	2	3	4	5	6	7	8	9

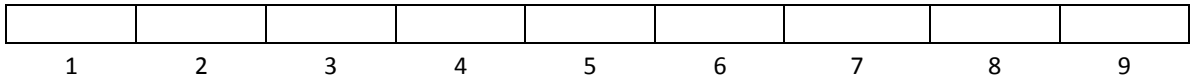
Please add any further comments you have regarding the design and usability of the website:

The next set of statements relate to the task and using the tools for marking fans and features on Mars:

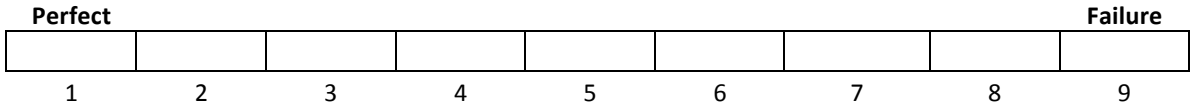
	Very Low								**Very High**
How mentally demanding was the task of marking craters?	1	2	3	4	5	6	7	8	9

How physically demanding was the task of marking craters?	1	2	3	4	5	6	7	8	9

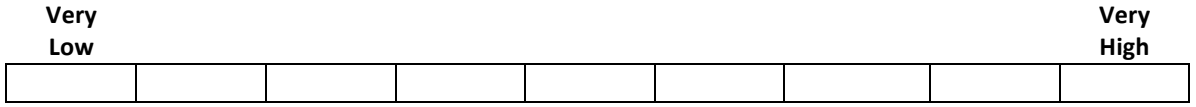
How hurried or rushed was the pace of marking craters?



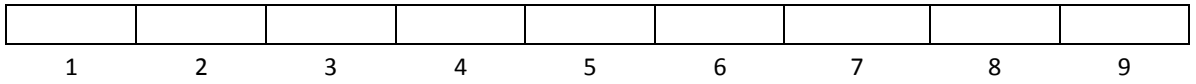
How successful were you in accomplishing the task of marking craters?



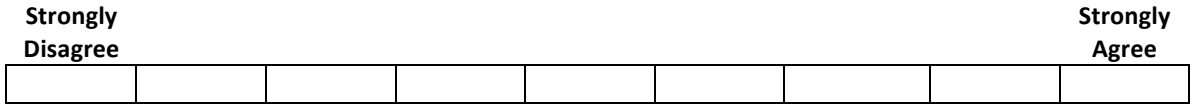
How hard did you have to work to accomplish your level of performance?



How insecure, discouraged, irritated, stressed or annoyed were you?



I found the 'position' marking tool easy to use



I found the 'crater' marking tool easy to use									
	1	2	3	4	5	6	7	8	9
I found the 'elliptical crater' marking tool easy to use									
	1	2	3	4	5	6	7	8	9
I would like more tools to mark different shapes and features									
	1	2	3	4	5	6	7	8	9

What, if anything, would you change about the tools or images and why?

The next set of statements relate to the images you are shown on the Planet Four website:

	Strongly Disagree							**Strongly Agree**

I found it hard to 'pick out' craters and features

1	2	3	4	5	6	7	8	9

I thought the image quality was consistent

1	2	3	4	5	6	7	8	9

There were sometimes too many craters in a single image

1	2	3	4	5	6	7	8	9

I didn't know what to do if an image had no craters or features to mark

1	2	3	4	5	6	7	8	9

Please add any further comments you have regarding the images shown on the Planet Four website:

Figure C.2: Laboratory study participant questionnaire

D

APPENDIX TO EXPERIMENT 2

This appendix includes additional material to experiment 2 (see chapter 9):

- Figure D.1 is an example Planet Four Blog post created to advertise Planet Four: Craters to the existing community.
- Figure D.2 shows the online questionnaire made available to participants through the Planet Four: Craters platform.

The Planet Four: Craters platform code is available on request through the Zooniverse Github repository: github.com/zooniverse/NottMarsZoo.

Additional contents available through the thesis Github repository include:

- Participant crater marking data using each interface for CTX image G05_020119_1895_XN_09N198W in directory: thesis/live experiment (chapter 9)/livecratermarkingsG05.csv.
- Participant position marking data using the Ramped interface for CTX image G05_020119_1895_XN_09N198W: thesis/live experiment (chapter 9)/livepositionmarkingsG05.csv.
- Participant crater marking data using each interface for CTX image G23_027332_1907_XN_10N202W: thesis/live experiment (chapter 9)/livecratermarkingsG23.csv.
- Participant position marking data using the Ramped interface for CTX image G23_027332_1907_XN_10N202W: thesis/live experiment (chapter 9)/livepositionmarkingsG23.csv.
- Reconstructed CTX image G05_020119_1895_XN_09N198W with participant markings shown (blue – full interface, red – stepped interface and yellow – ramped interface): thesis/live experiment (chapter 9)/livestudyG05.jpg.
- Reconstructed CTX image G23_027332_1907_XN_10N202W with participant markings shown (blue – full interface, red – stepped

interface and yellow – ramped interface): thesis/live experiment (chapter 9)/livestudyG23.jpg.

- Folder of image G05_020119_1895_XN_09N198W 'slices' presented to participants through the Planet Four: Craters platform, again showing participant markings colour-coded as previously described: thesis/live experiment (chapter 9)/image slices G05.
- Folder of image G23_027332_1907_XN_10N202W 'slices' presented to participants through the Planet Four: Craters platform, again showing participant markings colour-coded as previously described: thesis/live experiment (chapter 9)/image slices G23.
- Participant questionnaire response data for each interface: thesis/live experiment (chapter 9)/questionnaire results.csv.

Welcome to Planet Four: Craters!

Recently, a new version of the Planet Four project went live, and we are asking you to mark craters on the surface of Mars.

By counting the craters we will be able to figure out how old various geological surfaces are! This will be a big help for missions such as the 2016 NASA InSight experiment, which will use geophysical techniques such as seismology and heat flow to figure out how Mars has evolved. Knowing the age of the surface will help us to put a time-scale on that evolution.

Another aim of the project is to help improve the design of future Zooniverse sites.

When marking craters on Mars using this new version of Planet Four, you will be using one of three different classification interfaces. They each have different tools for you to use (on the left hand side of the page), and will ask you to complete tasks in differing orders for each image. Don't worry! When you use one of these interfaces for the first time a tutorial will guide you through how to use it, in the same way the original Planet Four site did.

The reason for the different interfaces is that we want to know which one works the best, the one that you enjoy using the most and find the easiest to use, and gives the best crater marking results.

To find this out, we would really value your feedback. If you have a spare 15 minutes, there is a questionnaire on the Talk page (accessed through clicking the discuss button on the classify page) where you can answer questions and give your opinions about using the interfaces.

The information you give us will then be used to help design future versions of Planet Four, and other Zooniverse projects – so your opinions really count!

Get involved now at craters.planetfour.org

Happy crater marking!

James

Figure D.1: Example Planet Four Blog post for advertising site

PLANET FOUR: CRATERS USABILITY QUESTIONNAIRE

Q1. Please state your age (years): (Leave blank if you prefer not to answer)

Q2. Please tell us your location: (Leave blank if you prefer not to answer)

Q3. Are you: (Leave blank if you prefer not to answer)

- ❑ Male
- ❑ Female

Q4. Which other Zooniverse projects have you taken part in?

- ❑ Galaxy Zoo
- ❑ Moon Zoo
- ❑ Solar Stormwatch
- ❑ Planet Hunters
- ❑ Milkyway Project
- ❑ Old Weather
- ❑ Cyclone Center
- ❑ Ancient Lives
- ❑ Whale FM
- ❑ Seafloor Explorer
- ❑ Bat Detective
- ❑ Cell Slider
- ❑ None

Which Interface have you been using?

- ❑ Full - All the tools are available (Craters Present, Craters, elliptical, Interesting Feature)
- ❑ Stepped - The tools were available one after the other, when the green 'Next Task' button was pressed
- ❑ Ramped - You used one tool for a number of images (Craters Present), then another (Position) for the next set of images, and then finally the 'Crater' and 'Elliptical' tools

Q5. The first set of statements relate to your understanding, motivation and the ease of use of the Planet Four website:

	Strongly Disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I learnt to use the website very quickly	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the website visually attractive to look at	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the website easy to use and access	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q6. Please add any further comments you have regarding the design and usability of the website:

Q7. The next set of statements relate to the task and using of tools for marking craters and features on Mars:

	Very Low 1	2	3	4	5	6	7	8	Very High 9
How mentally demanding was the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How physically demanding was the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How hurried or rushed was the pace of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q8.

	Perfect 1	2	3	4	5	6	7	8	Failure 9
How successful were you in accomplishing the task of marking craters?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q9.

	Very Low 1	2	3	4	5	6	7	8	Very High 9
How hard did you have to work to accomplish your level of performance?	❑	❑	❑	❑	❑	❑	❑	❑	❑
How insecure, discouraged, irritated, stressed or annoyed were you?	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q10.

	Strongly Disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found the 'Position' marking tool easy to use	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the 'Crater' marking tool easy to use	❑	❑	❑	❑	❑	❑	❑	❑	❑
I found the 'Elliptical Crater' marking tool easy to use	❑	❑	❑	❑	❑	❑	❑	❑	❑
I would like more tools to mark different shapes and features	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q11. What, if anything, would you change about the website tools or images and why?

Q12. The next set of statements relate to the images you are shown on the Planet Four website:

	Strongly disagree 1	2	3	4	5	6	7	8	Strongly Agree 9
I found it hard to 'pick out' craters and features	❑	❑	❑	❑	❑	❑	❑	❑	❑
I thought the image quality was consistent	❑	❑	❑	❑	❑	❑	❑	❑	❑
There were sometimes too many craters in a single image	❑	❑	❑	❑	❑	❑	❑	❑	❑
I didn't know what to do if an image had no craters or features to mark	❑	❑	❑	❑	❑	❑	❑	❑	❑

Q13. Please add any further comments you have regarding the images shown on the Planet Four website:

Figure D.2: Planet Four: Craters online questionnaire