



Grant agreement no. 607379

SPA.2013.2.1-01 - Analysis of Mars Multi-Resolution Images using
Auto-Coregistration, Data Mining and Crowd Source Techniques

- Collaborative project -

D7.3

Report on Data Validation Tests by Citizen Scientific Users

WP 7 – Crowd-sourced features for change discovery and validation of data mining

Due date of deliverable: month 31 – July 2016

Actual submission date: 16 / 11 / 16* *(*) EC approval pending*

Start date of project: January 1st 2014 Duration: 36 months

Lead beneficiary for this deliverable: UCL

Last editor: Robert Houghton

Contributors: RJH, JW, JS (UNOTT)

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

History table

Version	Date	Released by	Comments
1.0	18/07/16	JS	Version 1.0 - appendices created
1.1	19/07/16	JS/JW/RH	Version 1.1 – converted and formatted
2.0	21/10/16	RJH	Version 2.0
2.1	26/10/16	JCS	Version 2.1
2.2	7/11/16	RJH	Version 2.2 - reformatting

Executive Summary

This document reports on a series of five activities carried out within Work Package 7 “*Crowd-sourced features for change discovery and validation of data mining*” formative to the final public form of the citizen science component of the European FP7 project “*iMars: Analysis of Mars Multi-Resolution Images using Auto-Coregistration, Data Mining and Crowd Source Techniques*”. The intention of these experiments was to resolve and explore a range of design issues that are specific to building a citizen science project consonant with, and expressive of, the unique data provided by the wider iMars project. This is important both because we wish to involve citizens with interesting and novel scientific imagery and because optimising citizen performance and engagement with the task improves the volume and quality of potential results.

Because iMars produced co-registered imagery spanning 40 years, this provides an opportunity for a change detection task that allows citizen users the opportunity for *discovery* (as opposed to more common tasks of cataloguing or labelling found in the majority of citizen science projects). We considered the range of potential changes in terms of features, and considered them in terms of their suitability for a citizen science project. Concluding that range of features appeared potentially viable, we examined the question of whether specialisation in asking users to detect single features (greater accuracy at a cost of a lower hit rate and consequent ‘vigilance decrement’) was preferable to asking users to detect multiple features (at a cost to accuracy but with a higher, more motivating hit rate). This empirical issue was also unresolved in the wider visual inspection literature. Overall we observed a pattern where specialist participants suffered from worsening performance over time while generalist participants improved in performance demonstrating skill acquisition. This suggests asking participants to detect multiple features will be a more engaging and productive long-term design option. We then investigated the best ways to provide a task and interface through which users could undertake change detection itself as an activity, settling on manual and automatic flicker as our preferred methods. Further, because iMars also includes development of data-mining techniques leading to considerations of various pipelines for checking and aggregating results between humans and computers, we investigated the interplay between users and simulated computer and crowd opinion. The most striking finding of this investigation was how rapidly naive users gained a feel for and judgement of the performance parameters of an algorithm or crowd and that people were generally more forgiving of algorithms than crowds to whom they tended to impute nefarious intentions for shortcomings, in contrast to general opinion in the area. Finally, a workshop was undertaken with post-graduate planetary scientists. This generated a significant amount of fully completed image classifications (multiple aggregated responses) to inform refinement of iMars data mining techniques in Work Package 6 “Change detection from Data mining & validation” and also led to the identification of additional issues to consider in the final implementation (control over flickering, issues of quality, provision of training and additional website functionality).

Table of contents

History table	2
Executive Summary	3
Table of contents	4
Key word list	6
Definitions and acronyms	6
1. Introduction	6
1.1. Scope and objectives	7
1.2. Relationship to activities in WP7 and relationship to other Work Packages	8
2. Assessment of Mars geomorphological features from a citizen science perspective	8
2.1. Survey of geomorphological features	9
2.1.1. Active Gullies	9
2.1.2 Avalanches	10
2.1.3. Dunes	12
2.1.4. Dust Devils	15
2.1.5. Recurring Slope Lineae (RSL)	17
2.1.6. New Impacts	19
2.1.7. Polar Pits	21
2.2. Suitability for a Citizen Science Project	22
3. Experiments to optimise human performance in the iMars crowdsourcing platform	25
3.1. Detection of single vs. multiple feature types	26
3.1.1. Experimental Methodology	27
3.1.1.1. Participants	27
3.1.1.2 Apparatus & Materials	27
3.1.1.3 Experimental Procedure	29
3.1.2. Results	30
3.1.3. Discussion	34
3.1.4. Conclusions	34
3.2. Optimisation of change detection method	35
3.2.1. Experimental Design	35
3.2.1.1. Participants	35
3.2.1.2. Apparatus & Materials	35
3.2.1.3. Experimental Procedure	38
3.2.2. Results	38

3.2.2.1. Participant Feedback	38
3.2.2.2. Results: Performance	40
3.2.3. Conclusions	42
3.3. Trust in the Crowd vs. Trust in the Machine	43
3.3.1. Introduction	43
3.3.2. Experiment C1: Trust in algorithms	44
3.3.2.1. Experimental Design	44
3.3.2.2. Participants	44
3.3.2.3. Apparatus & Materials	45
3.3.2.4. Experimental Procedure	46
3.3.3. Experiment C2: Trust in the crowd	49
3.3.3.1. Experimental Design	49
3.3.3.2. Participants	50
3.3.3.3. Apparatus & Materials	50
3.3.3.4. Experimental Procedure	50
3.4. Experiment C1 Results: The Computer	51
3.5. Experiment C2 Results: The Crowd	53
3.6. Discussion	56
4. MSSL Workshop	57
4.1. Introduction	57
4.2. Method	58
4.3. Results	60
5. Conclusions and implications for future work	63
6. Outputs & Publications	66
6.1. Publications	66
6.2. Workshops & Demonstrations	66
7. Relevant links	67
8. References	67

Key word list

Citizen Science, Crowdsourcing, Change Detection, Task Design, ICT, Internet Applications, Visual Inspection

Definitions and acronyms

Acronyms	Definitions
CRISM	Compact Reconnaissance Imaging Spectrometer
CSA	Citizen Science Alliance
CTX	Context Camera (a camera carried by MRO)
DEM	Digital Elevation Model
MOC	Mars Orbiter Camera
MOLA	Mars Orbiter Laser Altimeter
MRO	Mars Reconnaissance Orbiter
HiRISE	High Resolution Imaging Science Experiment (a camera carried by MRO)
HRSC	High Resolution Stereo Camera
MSSL	Mullard Space Science Laboratory
PanCam	Panoramic Camera
NASA-TLX	National Aeronautics and Space Administration Task Load Index
RSL	Recurring Slope Lineae
THEMIS	Time History of Events and Macroscale Interactions during Substorms
VCS	Virtual Citizen Science

1. Introduction

This is the deliverable of the iMars project - *D7.3. Report on data validation tests with citizen/scientific users* reports experiments and testing of the citizen science component of the iMars project, “Mars in Motion”. The aim of this work to examine experimentally prominent issues identified in *D7.1. Design guidelines for crowd-sourcing software* that are particularly relevant to iMars citizen science. It also uses, validates and refines software produced for *D7.2. Improved toolkits*.

Subsequent to the publication of the original Description of Work, the title of this deliverable was amended from “scientific users” to “citizen scientific users” to better to reflect the two demographics of user population that it needs to support - citizen scientists in carrying out citizen science tasks and

validation use by professional scientists. Work involving both types of user groups is reported in this deliverable.

1.1. Scope and objectives

Given that the iMars project concerns the co-registration and data-mining of Mars multi-resolution images captured over the last 40 years, this gives us the opportunity to produce a distinctive and exciting citizen science project that is truly expressive of this unique dataset and allows members of the public to interact with it. This leads to the following questions we must ask in order to make this a reality:

1. The iMars dataset (considered at large) encompasses a considerable volume of the Martian planetary surface. In this regard it is not, as in the case of some other citizen science projects, the product of specialised sensors seeking only one kind of event or feature. Rather it represents the planetary surface upon which a range of geological and weather events and the myriad features they produce might be visible. Consequently we have to make decisions about what data are collected by citizen scientists. Positioning this from a user perspective we therefore ask: **how many features (and which features) should we ask people to identify?**
2. iMars data lends itself to the detection of change; such change detection is a relatively unusual and interesting activity for citizen scientists to undertake (planetary science citizen science more typically involves counting or tagging, as for example in *Planet Four: Craters*). A change detection task has appeal in that it allows participants the chance of *discovery* in a planetary science context. Therefore we ask: **what is the best task-interfaced design for change detection?**
3. Another defining part of the iMars project is its use of advanced data-mining techniques. This prompts us to ask: **what is the best way to integrate crowd and machine activity?**

These questions need to be answered as part of the final implementation of an effective citizen science project and, interestingly to us, mark a point of interface between hard science and human capability. Thus, the scientific aims of the wider project find expression here in terms of what citizen scientists will be asked to do and how they will be asked to do it. A set of experiments reported here address these questions that were with one exception, conducted using the Zooniverse Panoptes platform (Bowyer et al., 2015). They should therefore be understood as doing double duty as both controlled experiments but also part of on-going iterative development of the final public citizen science project itself.

The deliverable begins by reporting a survey of Mars planetary features that may be prominent in iMars imagery. These are evaluated not only in terms of their physical and scientific nature, but also in terms of how appropriate we believe they might be for citizen science activities. For example, some features can be productively reported in terms of their mere presence, in the case of others we might need to also provide some sort of measurements (introducing extra task complexity for a user to navigate) to be worthwhile. Alternatively, other features may require intricate yet repetitive annotation unlikely to sustain much volunteer interest. In general we construe the construction of an engaging and productive citizen science experiment appearing out of a negotiation between scientific needs (the “science case”)

and the needs of users to be engaged in their contribution. We then discuss experiments we undertook to examine how the system could be optimised for participant engagement and performance. We conclude by reporting on a workshop activity undertaken with skilled planetary scientists. This provides a beta-test of the system using iMars-generated planetary imagery, generated user-testing data, in terms of both quantitative performance and qualitative feedback, and also provides a raw training dataset to be used in refinement of the data mining techniques developed elsewhere in the project.

1.2. Relationship to activities in WP7 and relationship to other Work Packages

Within Work Package 7, the reported activities follow-on from an initial literature review and scoping exercise (reported as *D7.1 Design guidelines for crowd-sourcing software*) that was performed to understand what human factors issues might be most relevant to an iMars citizen science project. Our analysis of the Martian surface features in terms of their appropriateness for citizen science study and identification and form of the optimisation experiments reported here are both directly informed by this. Further in Work Package 7, we also developed a technical approach to carrying out a citizen science experiment based on the use of iMars imagery (*D7.2 Improved toolkits*) which we seek to deploy at scale before the end of the project (this constituting part of the project's *Milestone 9*). The present work uses the approach we developed leveraging the Zooniverse's Panoptes platform and therefore carries this work forward serving as part of its ongoing iterative development and testing before final deployment.

In terms of its relationship to other iMars Work Packages, we have paid close attention to interlinkages, particularly as our eventual goal is to produce an experience for the general public expressive of the unique datasets produced more widely in iMars and the scientific issues they most compellingly speak to. This work therefore takes input in terms of imagery from Work Packages 2 (*Auto-coregistration of NASA S/W & data*) and 6 (*Change detection from Data mining & validation*), collaborates further with Work Package 6 in terms of accepting co-registered image pairs in which change has assessed by data mining and returning a training dataset for further training of the processing algorithms being used. Given the nature of citizen science, it also contributes to the general outreach goals of iMars expressed in Work Package 8, supported specific public outreach events through providing an interactive exhibit (e.g., the Nottingham 'science in the park' event attended by around 7,500 visitors) and contributed to a training workshop for planetary scientists.

2. Assessment of Mars geomorphological features from a citizen science perspective

We began the process of assessing which features it might be interesting and appropriate to draw the attention of citizen scientists with a survey concerning how professional planetary scientists had examined those features, the imagery they had used and how they characterised them. We report the substance of this survey below by feature type.

2.1. Survey of geomorphological features

2.1.1. Active Gullies

Martian gullies are small networks of narrow channels, along with their associated down slope deposits, that occur on steep slopes, especially on crater walls. They have so far been predominantly detected at northern latitudes between 25 and 75 degrees. Current research regarding Martian gullies fall into two categories, comparison and change detection. The first category studies mostly HiRISE and CTX imagery and compares gullies found on Mars to those on Earth, in order to better understand their formation. The second category (more aligned with iMars aims) studies repeat observations of HiRISE imagery in order to derive their movement and growth over time.

Table 2.1. Breakdown of active gully studies

Publication	Detection	Technique	Imagery	Outcomes
Dundas, C.M., Diniega, S., McEwen, A.S., n.d. Long-term monitoring of martian gully formation and evolution with MRO/HiRISE. Icarus. doi:10.1016/j.icarus.2014.05.013	Change/ Movement	Repeat observations of HiRISE	HiRISE (25-75 lat)	Ongoing gully formation rather than degradation
Johnsson, A., Reiss, D., Hauber, E., Hiesinger, H., Zanetti, M., 2014. Evidence for very recent melt-water and debris flow activity in gullies in a young mid-latitude crater on Mars. Icarus 235, 37–54. doi:10.1016/j.icarus.2014.03.005	Comparison	Earth Vs Mars	CTX HiRISE THEMIS HRSC	Gullies formed by melting ice rather than impact heat
Hobbs, S.W., Paull, D.J., Clarke, J.D.A., 2014. A comparison of semiarid and subhumid	Comparison	Earth Vs Mars	DEM derived from HiRISE	Formation not restricted to single process. Slope and

Publication	Detection	Technique	Imagery	Outcomes
terrestrial gullies with gullies on Mars: Implications for Martian gully erosion. Geomorphology 204, 344–365. doi:10.1016/j.geomorph.2013.08.018				Sinuosity need context (environment)
Conway, S., Balme, M., Murray, J., Towner, M., 2014. Comparing the topographic long profiles of gullies on Earth and Mars. Presented at the EGU General Assembly Conference Abstracts, p. 15122.	Comparison	Earth Vs Mars	HiRISE	Study of gully profiles similar to Earth fluvial gullies
Raack, J., Reiss, D., Appéré, T., Vincendon, M., Ruesch, O., Hiesinger, H., 2014. Present-day seasonal gully activity in a south polar pit (Sisyphi Cavi) on Mars. Icarus. doi:10.1016/j.icarus.2014.03.040	Change / Movement	Repeat Observations of HiRISE	HiRISE CTX	Deposits formed by dry flows supported by sub. Of co2 ice

2.1.2 Avalanches

Martian avalanches feature the collapse of ice material down a slope, and therefore occur at the poles. Current research into Martian avalanches use a range of differing techniques, some of which are more suitable for use in a citizen science project than others. As with gullies, research exists looking at comparing Martian avalanches with those on Earth, as well as looking at change detection through repeat HiRISE imagery in order to discover trends in location, season, velocity and inclination. Other research looks at wind tunnel modelling in order to derive avalanche triggering processes, creating an inventory of crater types where avalanches occur and more in depth metric analysis of rock mass properties and slope instability.

Table 2.2. Breakdown of avalanche studies

Publication	Detection	Technique	Imagery	Outcomes
Brunetti, M.T., Guzzetti, F., Cardinali, M., Fiorucci, F., Santangelo, M., Mancinelli, P., Komatsu, G., Borselli, L., 2014. Analysis of a new geomorphological inventory of landslides in Valles Marineris, Mars. Earth and Planetary Science Letters 405, 156–168. doi:10.1016/j.epsl.2014.08.025	Comparison	Inventory of mid to high res images	CTX HRSC	Proportion of large landslides bigger than on Earth, and look to be seismically induced
Russell, P.S., Byrne, S., Dawson, L.C. 2004. Active powder avalanches on the steep north polar scarps of Mars – 4 years of HiRISE observation. 45th Lunar and Planetary Science Conference, 2688.	Detection Change / Movement	Repeat observations of HiRISE	HiRISE	Trends found in location, time of day, season, velocity, inclination
de Vet, S.J., Merrison, J.P., Mittelmeijer-Hazeleger, M.C., van Loon, E.E., Cammeraat, L.H., 2014. Effects of rolling on wind-induced detachment thresholds of volcanic glass on Mars. Planetary and	Samples	Wind-tunnel Experiments	None	Recent sand mobility benefited from rolling as saltation triggering process

Publication	Detection	Technique	Imagery	Outcomes
Space Science 103, 205–218. doi:10.1016/j.pss.2014.07.012				
Weiss, D.K., Head, J.W., 2014. Ejecta mobility of layered ejecta craters on Mars: Assessing the influence of snow and ice deposits. Icarus 233, 131–146. doi:10.1016/j.icarus.2014.01.038	Comparison	Inventory of crater types	CTX THEMIS	Presence of variable thickness icy substrate are consistent with ejecta mobility
Crosta, G.B., Utili, S., De Blasio, F.V., Castellanza, R., 2014. Reassessing rock mass properties and slope instability triggering conditions in Valles Marineris, Mars. Earth and Planetary Science Letters 388, 329–342. doi:10.1016/j.epsl.2013.11.053	Detection	Metric analysis	MOLA THEMIS	Low seismic events induced by impacts, could be a cause of landslides

2.1.3. Dunes

Compared to other features found on the surface of Mars, dunes can be much more complex. They are found everywhere across the planet, either individually or as part of large sand seas, and come in number of different sizes and types. As such, there is a large range of current research concerning dunes looking at a number of different associated metrics including migration, slip face movement, ripple movement, edge position, location of different types etc. This poses an interesting question when considering dunes for a citizen science project. While the range of different metrics that can be measured could offer interesting variability in terms of task design and user engagement, some of the measurements taken by current research could be deemed either too subtle or complex for an untrained community.

Table 2.3. Breakdown of dune studies

Publication	Detection	Technique	Imagery	Outcomes
Fenton, L.K., Michaels, T.I., Chojnacki, M., Beyer, R.A., 2014. Inverse maximum gross bedform-normal transport 2: Application to a dune field in Ganges Chasma, Mars and comparison with HiRISE repeat imagery and MRAMS. Icarus, Third Planetary Dunes Systems 230, 47–63. doi:10.1016/j.icarus.2013.07.009	Change detection	Repeat observations of HiRISE	HiRISE CTX	Southward movement of ~2.6m/E. Year
Chojnacki, M., Johnson, J.R., Moersch, J.E., Fenton, L.K., Michaels, T.I., Bell III, J.F., n.d. Persistent aeolian activity at Endeavour crater, Meridiani Planum, Mars; new observations from orbit and the surface. Icarus. doi:10.1016/j.icarus.2014.04.044	Change detection	Repeat temporal and spatial observations	HiRISE CTX Opportunity PanCam	Dome dunes have highest migration rates (4-12m/ M. Year)
Hayward, R.K., Fenton, L.K., Titus, T.N., 2014. Mars Global Digital Dune Database (MGD3):	Detection	Metric Analysis	CTX HiRISE	NP dunes are part of large seas, EQ & SP dunes individual, in

Publication	Detection	Technique	Imagery	Outcomes
Global dune distribution and wind pattern observations. Icarus, Third Planetary Dunes Systems 230, 38–46. doi:10.1016/j.icarus.2013.04.011				craters
Johnson, M.B., Zimbelman, J.R. 2014. Documentation of sand ripple patterns and recent surface winds on martian Dunes. 45th Lunar and Planetary Science Conference, 1518.	Detection	Study of Ripple patterns	HiRISE	Ripple patterns can be used to infer wind strength and dir.
Bourke, M., McGaley-Towle, Z., 2014. Why do sand furrow distributions vary in the North Polar latitudes on Mars? Presented at the EGU General Assembly Conference Abstracts, p. 13626.	Change detection	Temporal and spatial variation of sand furrows	HiRISE	Furrow formation linked to ice thickness
Chojnacki, M., Burr, D.M., Moersch, J.E., 2014. Valles Marineris dune fields as compared with other martian populations: Diversity of dune compositions,	Change detection	Migration, slip face, ripples, edge and deflation	HiRISE	Dune activity areas lower in elevation (~1km) Aeolian activity largely influenced by environment

Publication	Detection	Technique	Imagery	Outcomes
morphologies, and thermophysical properties. Icarus, Third Planetary Dunes Systems 230, 96–142. doi:10.1016/j.icarus.2013.08.018				
Bridges, N.T., Ayoub, F., Avouac, J-P, Leprince, S., Lucas, A., Mattson, S. 2012. High sand fluxes and abrasion rates on mars determined from HiRISE Images. 43rd Lunar and Planetary Science Conference, 1322.	Change detection	COSI-Corr tools used to measure changes from 1/10 pixel	HiRISE	Migration of ~0.1m/yr for ripples, 5 times larger for lee front

2.1.4. Dust Devils

A dust devil is a strong, well-formed, and relatively long-lived whirlwind, ranging from small (15 metres wide) to large (more than 300 metres wide and more than 1000 metres tall). The primary vertical motion is upward. Martian dust devils can be up to fifty times as wide and ten times as tall as those found on Earth, and there is a wide range of research concerning their behaviour currently ongoing. This research either concentrates on the ‘live’ movement of dust devils across the surface, measuring their speed and direction, or looks at their seasonal recurrence, i.e. how often they appear and in what density across several months or years. Both directions of research could be a good fit for a citizen science project, with the interest/excitement of seeing a ‘live’ dust devil travel across the surface, and the temporal change across months and years facilitating a large image dataset.

Table 2.4. Breakdown of dust devil studies

Publication	Detection	Technique	Imagery	Outcomes
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, Image Processing and Analysis in Space Science 53, 1822–1833. doi:10.1016/j.asr.2013.05.012	Automated	Tested automated results against visual estimations	MOC HiRISE	Best automation by considering directional openings
Lorenz, R.D., Reiss, D., 2015. Solar panel clearing events, dust devil tracks, and in-situ vortex detections on Mars. Icarus 248, 162–164. doi:10.1016/j.icarus.2014.10.034	Comparison	Comparison of dust clearing events with dust devil seasons	Spirit Rover solar array data	Recurrence of 100-700 sols similar to that of dust devil track generation
Reiss, D., Hoekzema, N.M., Stenzel, O.J., 2014. Dust deflation by dust devils on Mars derived from optical depth measurements using the shadow method in HiRISE images. Planetary and Space Science 93–94, 54–64. doi:10.1016/j.pss.2014.01.016	Change detection	Used surface image offset between colour channels, and comparison between HiRISE and CTX, CRISM	HiRISE CTX CRISM	Horizontal speed of dust devil found to be around 4.8 ms ⁻¹
Reiss, D., Spiga, A., Erkeling, G., 2014. The horizontal motion of	Change detection	Time-delayed	HiRISE CTX	Speeds of 4-25ms ⁻¹ ,

Publication	Detection	Technique	Imagery	Outcomes
dust devils on Mars derived from CRISM and CTX/HiRISE observations. Icarus 227, 8–20. doi:10.1016/j.icarus.2013.08.028		image sets between CRISM, CTX & HiRISE	CRISM	diameter 15–280m, majority in northern hemisphere
Lorenz, R.D., 2013. Dust devil populations : Comparing in-situ measurements with imaging and tracks. Mars Atmosphere: Modelling and observation, 5th international workshop, 1, 1406.	Comparison	In situ data with imaging	HiRISE	Most frequent observed track diameter larger than most frequently observed

2.1.5. Recurring Slope Lineae (RSL)

Recurring slope lineae are narrow, dark markings found on steep slopes that incrementally lengthen during warmer periods, then fade over cooler seasons and can recur over multiple Martian years. Current research seems to concentrate on RSL found in the southern mid-latitudes. While there is research into simulating their formation (water volumes required etc.), most concentrates on observing their evolution, from their appearance, lengthening over a season and then fading. This, coupled with the opportunity to also attempt to measure their recurrence over several years, could make RSL a good fit for a citizen science project.

Table 2.5. Breakdown of recurring slope lineae studies

Publication	Detection	Technique	Imagery	Outcomes
Chojnacki, M., McEwen, A., Dundas, C., Mattson, S., Ojha, L., Byrne, S., Wray, J., 2014. Geologic Context of Recurring Slope Lineae in Coprates Chasma. Presented at the Lunar and Planetary Science Conference, p.	Change detection	Repeat HiRISE imagery, DTM's co-registered with MOLA	HiRISE MOLA	Broad scale of spatial and vertical dist. Of RSL. Flows and fading over multiple mars yrs.

Publication	Detection	Technique	Imagery	Outcomes
2701.				
Ojha, L., McEwen, A., Dundas, C., Byrne, S., Mattson, S., Wray, J., Masse, M., Schaefer, E., 2014. HiRISE observations of Recurring Slope Lineae (RSL) during southern summer on Mars. Icarus 231, 365–376. doi:10.1016/j.icarus.2013.12.021	Change detection	RSL sites frequently monitored by HiRISE in MY30 & MY31	HiRISE	13 sites of RSL confirmed. Unique phenomenon on Mars, consistent with wet flow
McEwen, A., Byrne, S., Chevrier, V., Chojnacki, M., Dundas, C., Masse, M., Mattson, S., Ojha, L., Pommerol, A., Toigo, A., Wray, J., 2014. Recurring Slope Lineae and Future Exploration of Mars. Presented at the EGU General Assembly Conference Abstracts, p. 8851.	Change Detection	Monitoring of active RSL in equatorial region (0-15deg S)	HiRISE	Seasonal melting of shallow ice best explains RSL obs.
Stillman, D.E., Michaels, T.I., Grimm, R.E., Harrison, K.P., 2014. New observations of martian southern mid-latitude recurring slope lineae (RSL) imply formation by freshwater subsurface flows. Icarus 233, 328–341. doi:10.1016/j.icarus.2014.01.017	Change Detection	Analysis to see if RSL visible, lengthened or faded	HiRISE CTX	RSL lengthen for 104 sols, are intermittent and only have just started in the southern mid-latitudes
Grimm, R.E., Harrison, K.P., Stillman, D.E., 2014.	Simulation	Modelling of RSL as	None	Required water

Publication	Detection	Technique	Imagery	Outcomes
Water budgets of martian recurring slope lineae. Icarus 233, 316–327. doi:10.1016/j.icarus.2013.11.013		isothermal water flows		volumes are above which can be supplied by melting of near surface ice

2.1.6. New Impacts

New craters form all over the surface of Mars, and though they range in size the smaller are more frequent. Current research either looks at their occurrence in order to calculate cratering rates (a simple detection task, perhaps could become ‘boring’ for a citizen science project) or more specifically targets craters that reveal ice deposits, in order to work out ice shelf depths and glaciation. While more interesting for a non-expert the subtle changes in brightness etc. could be too hard to register for a citizen science community.

Table 2.6. Breakdown of ‘new impact’ studies

Publication	Detection	Technique	Imagery	Outcomes
Balme, M. & the ISSI team, n.d. Northern plains of Mars: Origins, evolution and response to climate change. http://www.issibern.ch/teams/plainsofmars/ accessed 17.10.16	New craters to expose ice	Planetary geomorphologic mapping	HiRISE, CTX	Formation of team to study ice-related geomorphology
Fassett, C.I., Levy, J.S., Dickson, J.L., Head, J.W., 2014. An extended period of	New features	Detection of new craters that superimpose glacial deposits	CTX THEMIS	Northern mid-latitude glaciation was a long-lived recurring process of

Publication	Detection	Technique	Imagery	Outcomes
<p>episodic northern mid-latitude glaciation on Mars during the Middle to Late Amazonian: Implications for long-term obliquity history. Geology G35798.1. doi:10.1130/G35798.1</p>				~600m.y.
<p>Williams, J.-P., Pathare, A.V., Aharonson, O., 2014. The production of small primary craters on Mars and the Moon. Icarus 235, 23–36. doi:10.1016/j.icarus.2014.03.011</p>	Feature counting	Comparison of modelled distribution (flux of terrestrial fireballs) verses observed crater counts	HiRISE	Average cratering rate has been constant between 52.3Ma and 23.9Ma
<p>Daubar, I.J., Atwood-Stone, C., Byrne, S., McEwen, A.S., Russell, P.S., 2014. The morphology of small fresh craters on Mars and the Moon. Journal of Geophysical Research: Planets.</p>	New features	Calculation of depth/diameter ratio for new craters	HiRISE	d/D ratio for craters in last 20 yrs = 0.23. Variations in d/D suggest differences in target material, impact velocity, angle and state of the bolide(s)

Publication	Detection	Technique	Imagery	Outcomes
JE004671. doi:10.1002/2014JE004671				
Dundas, C.M., Byrne, S., McEwen, A.S., Mellon, M.T., Kennedy, M.R., Daubar, I.J., Saper, L., 2014. HiRISE observations of new impact craters exposing Martian ground ice. Journal of. Geophysical Research: Planets 119, 2013JE004482. doi:10.1002/2013JE004482	New features, change detection	Observation of ice excavation by new impacts	HiRISE	Modelling suggests ice requires atmos. Water vapour content of 24 micrometers, double the present value
Daubar, I.J., McEwen, A.S., Byrne, S., Kennedy, M.R., Ivanov, B., 2013. The current martian cratering rate. Icarus 225, 506–516.	New features	44 new craters identified, cratering rate compared to models	CTX, HiRISE	Generally good agreement with models, future multi-decade obs. Needed

2.1.7. Polar Pits

Polar pits are small. Negative relief features that as of yet have only been found in polar troughs. They are around 1-5 metres in diameter, and can appear, change in size and disappear seasonally. Although at first glance the size of such features would be too small for a citizen science project using traditional remotely sensed data (HiRISE etc.), it could be possible using processed ortho-rectified imagery. Other research has looked at phenomena that occur around polar pits – transient bright ‘halos’. This could be

an easier fit for a citizen science project, as in addition to changing in size seasonally they are also much larger – ranging from 10 – 55 metres in diameter.

Table 2.7. Breakdown of polar pit studies

Publication	Detection	Technique	Imagery	Outcomes
Mattson, S., Kilgallon, A., Byrne, S., McEwen, A.S., Herkenhoff, K., Okubo, C., Putzig, N.E., Russel, P. 2014. Meter-scale pits in Mars' north polar layered deposits. 45th Lunar and Planetary Society Conference, 2431.	New features, change detection	Detection, morphology and possible formation mechanisms	HiRISE DTM, Ortho-rectified	1-5 m in diameter, disappear, appear, and change in size over a season
Becerra, P., 2014. Transient bright “halos” on the south polar residual cap of mars: Implications for mass balance. LPI Contributions 1791, 1013.	Change detection	Analysis of halos that appear around pits	CTX HiRISE MOC	Only present in MY28, between 279 & 331. Width changes between 12-55m.

2.2. Suitability for a Citizen Science Project

The table below rates each feature found on the surface of Mars in terms of its suitability to be used in a citizen science project. They are scored using five different criteria, as described below, between 1 and 3, with 3 being the most suitable. The criteria are:

- **Coverage:** How widely the features are present across the Martian surface, spatially.
- **Temporal Variability:** The time scales they change/are created over, be it across several decades, years, seasonal or even smaller.
- **Size:** The size range of the feature, i.e. can they be easily identified in remotely sensed imagery.
- **Task Complexity:** The complexity inherent in making observations about the feature. For example, some features may just require simple judgements about their presence or absence whereas others may require a multi-stage process involving annotation and shape tracing.
- **Task Variability:** The range of different types of measurements or observations that can be made about each feature, to allow for task variability within the same citizen science project.

Table 2.8: Feature scores in terms of suitability for a citizen science platform

Feature Type	Coverage	Temporal Variability	Size	Task Complexity	Task Variability	TOTAL
Active Gullies	2	3	2	3	3	13
Avalanches	1	2	3	2	2	10
Dunes	3	2	3	1	3	12
Dust Devils	3	3	3	3	3	15
Recurring Slope Lineae (RSL)	2	3	3	3	3	14
New Impacts	3	2	2	2	1	10
Polar Pits	1	3	1	2	2	9

Dust Devils – 15 points: In terms of coverage, they have been found all over the surface of Mars, in both the north and south hemisphere. They is great scope in variability, where analysis could look both at their change in numbers over many decades, to seasonal, and even in some images their ‘live’ movement across the surface. The tracks they leave on the surface can be several hundred metres long so easy to spot in most remotely sensed imagery, with a width of around 5m. Simple ‘new feature’ detection, up to marking the direction and length of tracks should be well within the capabilities of citizen scientists, while the scope of tracking both new and existing features and ‘live’ movement allows for good variability of task to keep volunteers interested.

Active Gullies & Recurring Slope Lineae – 13 & 14 points: In terms of coverage, most coverage centres between 25-75 degrees latitude in the northern hemisphere. As with dust devils, there is again great scope in temporal variability, over several decades, yearly and seasonally. Again the sizes of the gullies and lineae can be several hundred metres in length, but very narrow, so their visibility can vary on remotely sensed imagery. Simple new feature detection, up to marking the direction and change in length over a season should be possible by citizen scientists, which again allows good task variability. An extra advantage of gullies could be that when the original idea for Planet Four was floated, the HiRISE team were very interested in this feature as a target.

Dunes – 12 points: Dunes are found all over the surface of Mars, although research suggests individual examples are mostly located in equatorial and southern Polar Regions. In terms of change over time, movement of both slip faces, ridges and the dune itself have been detected over decades and at a yearly rate, but little seasonal change is observable. Sizes range from metres to kilometres, with all but the

smallest well within the detection of remotely sensed imagery. In terms of complexity, current research tends towards small shifts in slip face position, direction, ridge movement etc. which could be too complex/subjective for a citizen science project. Also the small movement of a few metres per M. year could be very hard to detect without some sort of 'alternate image' mechanism. This does possibly allow for a lot of task variability however.

Avalanches & New Impacts – 10 points: While new impact craters occur ubiquitously across the surface, avalanches currently have only been studied in the Polar Regions. While other features are formed and change as part of a process, avalanches and new impacts are much more random, and as such it is almost down to luck if two images are captured either side of the event. While avalanches are similar in size to RSL and gullies, current research into new impacts tends towards smaller craters, which could be difficult to register in any imagery other than HiRISE. However, where new impacts occur in areas with dust cover they may remove dust hundreds of times larger than them and therefore be easier to spot; their detectability is therefore linked to the location in which they occur. In terms of the measurement task, current research involves either the study of ice abundance and movement or depth vs. diameter calculations, both of which could be rather complex/subtle for citizen scientists. Other new impact work studying the rate of cratering (i.e. just marking the position of new craters) could be quite a boring task for the general public.

Polar Pits – 9 points: Polar pits, as the name suggests, only occur on the polar regions of Mars. They change in abundance over decades and years while they also change seasonally in size and appearance. Regarding size, they are most commonly around the order of 1 metre in diameter. In terms of task complexity, most studies have concentrated in measuring the change of brightness associated with ice abundance, which could be too complex for a citizen science project, unless it is a simple judgement of 'which is brightest?' that in turn does not allow for much task variability. An alternate target feature to measure could be the associated 'halos' that have been detected around them, which are much larger in size (12-55m diameter) and again change in size seasonally.

Overall, the total score across the features has a relatively small range, reflecting the fact that conceivably they could all be suitable targets for a citizen science project. It must be pointed out however this holds true with each category equally weighted, if one category was deemed significantly more important it could quickly rule out certain features depending on which. Of more relevance when choosing the feature to target will be both creating the tasks / measurements that will be needed in such a way that engages the public, and most importantly identifying a science team that will be interested in the results, will engage with the community and carry out the reduction/interpretation of the data.

3. Experiments to optimise human performance in the iMars crowdsourcing platform

In this section we report three sets of studies that investigated human factors issues related to the design of the iMars platform (these issues had been previously identified in a formative literature review reported as iMars D7.1, and our thinking about them in terms of technical constraints informed by technical work behind platform development reported in iMars D7.2).

The first experiment (*3.1 Detection of single vs. multiple feature types*) concerns how many features it is reasonable to ask a participant to identify. On the one hand, an obvious simple strategy is to ask participants to detect only a single change feature at a time (e.g., the appearance of a new crater). However, on deeper consideration on the imagery produced in the iMars project, and our reflections on the viability of crowdsourced detection of different features in the section 2 of this report, an alternative strategy of asking participants to report a range of features emerges as a richer and potentially more engaging task that more fully exploits the distinctive data available and therefore engages the public more fully in iMars itself. However, examination of the available data suggested there might be other human factors, particularly those relating to performance, which might come to bear on this decision which we assess experimentally here.

The second experiment (3.2) explores different approaches to change detection identified through review of the visual perception literature in D7.1. Two main options were identified. The first, flicker, is a practice that we observed satellite imagery and space scientists using. This method exploits neural circuitry devoted to motion detection. As a result of pilot work we identified two variants; automatic flicker and manually controlled flicker. The second option is to permit side-by-side inspection of imagery, a form of task that might be expected to draw more on cognitive (particularly memory-based) processes. These options were again evaluated experimentally in terms of performance, task time and perceived workload.

The third experiment (3.3) draws on ideas we identified in D7.2 as regards options for pipelines including humans and algorithms. Our particular area of interest here was how being informed how either an algorithm or a crowd judgement informed the later processing of images in terms of both performance and trust in the system.

The results of these experiments offer guidance as to the final form of the iMars citizen science platform which we intend to fulfil dual goals of involving the general public in cutting-edge planetary science and producing useful and scientifically useful datasets.

3.1. Detection of single vs. multiple feature types

Having come to the view above that a range of feature types might be viably detected in iMars imagery by citizen scientists, we carried out an optimisation experiment to assess whether it might be better to ask individual participants to search for a range of different features (a “General” strategy) or whether it might be more optimal to partition participants so that they looked only for a specific feature (a “Specific” strategy). Our analysis of the literature, discussed below, demonstrated that there were arguments on both sides generated through research interest primarily in industrial inspection. Thus, we had a fundamental human science issue to resolve. In this experiment non-planetary imagery was used (photographs of cork tiles some of which contained ‘features’ in the form of defects – scratches, notches and so on) in order to rapidly generate a high volume of appropriate imagery.

The most relevant research base to the iMars citizen science task of inspecting planetary imagery for changed features is that which concerns industrial visual inspection, particularly in manufacturing quality control, but also in security and medical contexts (i.e., the inspection of X-rays images). Although the industrial (and consequently, research) emphasis on rapid visual inspection has generally declined in the manufacturing arena as computer-controlled manufacturing made visual inspection of finished goods less important as a locus of quality control, there remains a significant literature that examined the best ways of organising this activity.

Two main strategies of improving performance for inspection have been identified in literature. One approach (the “General strategy”) is to ask inspectors to search for a range of defects. This increases the amount of targets that can be found and makes the task more stimulating and avoids the vigilance decrement, wherein performance falls in a repetitive task where valid targets only rarely appear. Alternatively, another approach (the “Specific strategy”) is to ask inspectors to only look for a specific type of defect. Reducing the number of things to “look out for” and allowing the viewer to specialise can improve performance (Harris & Chaney, 1969, Megaw et al., 1979). While both these claims can be seen as intuitively valid, on consideration one might also notice that in any specific situation these would be contradictory approaches. In a real world task, a Specialised approach means that in any run of images, one is only seeking a particular type of change, thus reducing the number of targets overall and raising the risk of vigilance decrement (Sawyer et al., 2014; Harris, 1968; Broadbent & Gregory, 1965; Su & Konz, 1981; Meuter & Lacherez, 2015).. It is important to strike a balance between over and under stimulation, as this can adversely affect performance (Bexton et al., 1954) with an emphasis on discovering the optimal workload to maximise performance (Wiener et al., 1984). Thus the empirical question arises as to whether the performance increase from Specialised detection is, over a reasonably sustained period and compared in like-for-like circumstances, enough to off-set the increased risk of vigilance decrement. Furthermore, we might also wonder whether, if the human is considered as a sensing and decision making system, the performance profile in terms of tendency towards misses or false alarms. We may, in the context of a citizen science experiment, have a reason for favouring a particular response profile depending on how we analyse the resultant data. For example, in the

absence of ground truth, a set of false alarms may be harder to detect and correct than a series of misses depending on the approach to consensus that is taken.

This study is conducted in the domain of visual inspection on a desktop computer platform (as per a citizen science experiment); however its results have implications for the all other varieties of vigilance-based activity. A similar experiment was carried out by Su and Konz (1981) who suggested that the one defect at time Specialised approach was suited to identifying harder to discern defects. However, this research is novel in that it not only generated its own visual stimuli but also uses actual photographic surface defect imagery whereas most previous experiments resorted to assessing these issues with artificial figurative stimuli or pen and paper assignments. Furthermore, because the above literature approached these issues from a manufacturing perspective, participants were typically asked in all cases whether a defect was present or not even if asked to look out for several specific types. In the iMars science case however we do not wish to know simply that change is present but also, if it is, what kind of change. This alters the nature of the task at hand we were therefore also able to build this variation from the above into the design of the experiment.

3.1.1. Experimental Methodology

3.1.1.1. Participants

30 participants were recruited through email lists, social media posts and subsequent 'word of mouth'. All participants have been or are currently being educated at a university-degree level, and none have had any formal training directly related to visual inspection.

3.1.1.2 Apparatus & Materials

As there was no readily available surface change database (this being somewhat of a moot point), a comprehensive image library was generated manually before commencing the study. Manual generation of the images ensured better control of the study. After extensive surveying of materials available, cork coasters were chosen as the medium for the visual stimuli. They were photographed using a standard orientation using a tripod. Cork coasters were chosen due to the high level of visual information present on their surface as well as the ease in re-creating defects in large volumes. While obviously differing from planetary imagery, we considered these items a reasonable proxy that could be produced in bulk on demand yet also closely in control in as far as cork features a lot of surface detail and the indentations, discolorations etc. we could produce broadly approximate in a familial sense things like RSLs, impacts and suchlike.

The number of different defect types was capped at five, in accordance with literature stating that the practical maximum for an inspector (Rao et al., 2006). The surface defects as seen in Figure 3.1 were chosen as they represented ones commonly seen in industry (Doring et al., 2006). Pilot work, together with our impressions from the surrounding literature, suggested that the defects were of varying

difficulty with the dent and scratch defects harder to detect than the cut, dent and glue defects. The defects were placed at varied locations on the samples with varying magnitudes. There was only one category of defect on any given sample.

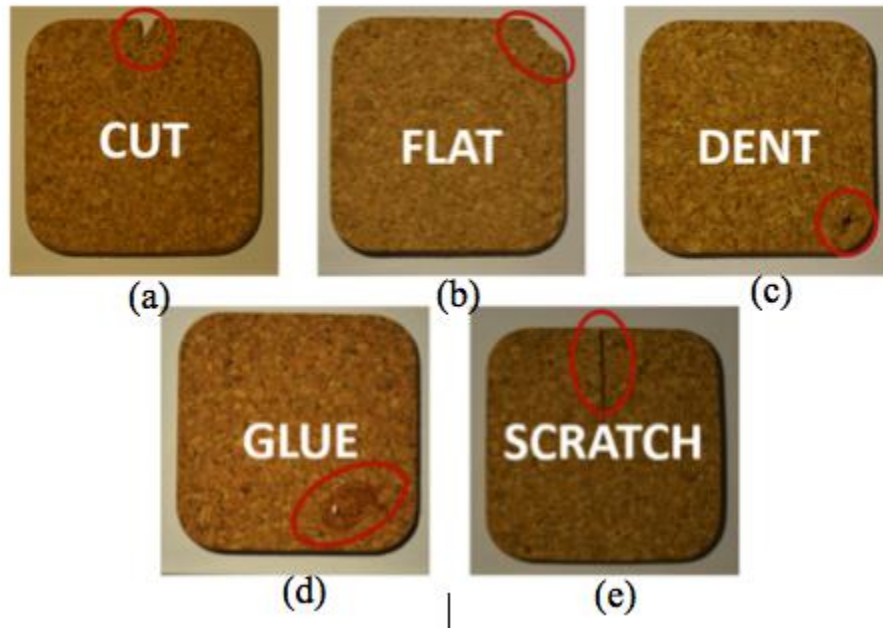


Figure 3.1: Defect categories used (a-e)

An image library of 1200 visual stimuli was generated with 200 containing defects distributed equally between the five defect types previously described. This would result in the General strategy having 200 defects, and the Specific strategy having 40 defects that translates into a 16.67% and a 3.33% defect rate respectively. Stimuli distributions for both strategies are visually represented in Figure 3.2.

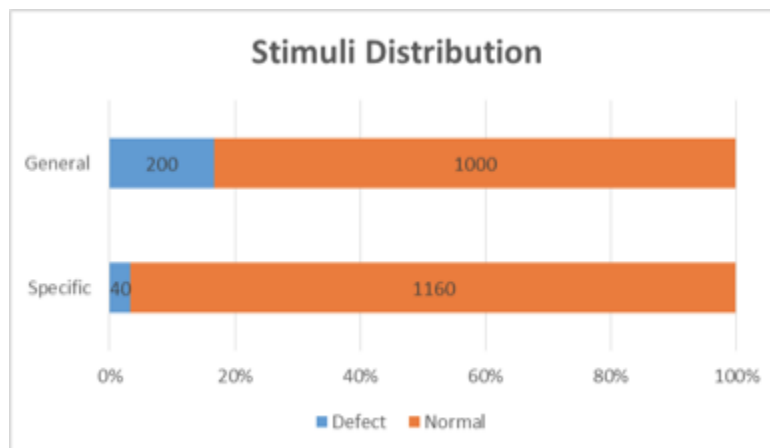


Figure 3.2: Stimuli distribution

3.1.1.3 Experimental Procedure

The study was built and conducted using the PsychoPy2 software (Pierce, 2009; 2007). The software allowed the experiment to call and display images from the library and record participant responses. The experiment was run on a desktop computer with participants responding using the keyboard. The study was designed to have participants in both strategies view stimuli drawn without replacement from the image library in a random order and respond accordingly. The experiment would record which stimuli was displayed, and the participant response and response time for that stimuli. Table 3.1 displays the breakdown of the 30 participants into their respective conditions.

Table 3.1: Study participant blocking

Condition		Participants	
General Strategy		x15	
Specific Strategy	Cut condition	x3	x15
	Flat condition	x3	
	Dent condition	x3	
	Glue condition	x3	
	Scratch condition	x3	

After listening to a briefing and having their questions answered, they would then sign their consent. Both strategies of the study would then have participants assigned to a condition. As the participants were all university students and none of them could have been considered as experts in visual inspection. In order to familiarise them with the task, they would complete a short practice session with instantaneous feedback. It was limited to only 30 trials to avoid any potential bias to be introduced by the presence of rapid feedback (Drury & Addison, 1973). This form of training is also commonly found in citizen science projects and is an approach we also intend to take. The General strategy asked participants to indicate which of the categories the stimuli displayed belonged to or if it was a normal sample. In the Specific condition, participants were assigned a defect category and were asked to only reject samples if the stimuli presented contained a defect from their assigned category.

Upon conclusion of the study the participants were asked to complete the NASA Task Load Index (NASA-TLX), a cognitive workload assessment tool (NASA, 1986). This tool is used throughout the remainder of this report and is worth going into some detail here. The NASA TLX is generally regarded as an industry standard for the self-report of workload on a task and as such has been used in at least 4,000 published studies across myriad situations and industries (Hart, 2006). It is multi-dimensional subjective rating procedure that generates workload scores based upon six subscales:

- **Mental demand** (How much mental and perceptual activity was required?).
- **Physical demand** (How much physical activity was required?)
- **Temporal demand** (How much time pressure did you feel due to the pace at which the task or task elements occurred?),
- **Performance demand** (How successful were you in performing the task?)
- **Effort** (How hard did you have to work, mentally and physically, to accomplish your level of performance?)
- **Frustration** (How irritated, stress and annoyed, versus context, relaxed and complacent, did you feel during the task?)

This would allow analysis of the perceived workload in the task. The NASA TLX includes a weighting procedure where subscale elements thought to most specifically define the task can be emphasised in overall summative score. In practice, this procedure is only relatively rarely used, we were on this occasion most interested in how the strategies were generally characterised in terms of the subscales (for example, did people *feel* that they were performing worse in one condition than the other?) and thus the so-called “Raw TLX” is was reported here (Hart, 2006).

3.1.2. Results

The responses for the two competing strategies were analysed along the individual defect categories after which it was aggregated to derive the overall empirical and theoretical measures of performance.

The average responses while using the Specific strategy are shown in Table 3.2. The Specific strategy experienced miss and false alarm rates of 15.78% and 1.50% respectively.

Table 3.2: Response breakdown using specific strategy

	Accept	Reject
Normal	1142.5	17.5
Defected	6	34
Recall	0.995	
Precision	0.985	
Harmonic mean (F-score)	0.495	

The average responses of a participant using the General strategy are as shown in Table 3.3. The General strategy experienced overall miss and false alarm rates of 14.79% and 0.91% respectively.

Table 3.3: Response breakdown using general strategy

	Accept	Reject
Normal	991	9
Defected	30	170
Recall	0.991	
Precision	0.971	
Harmonic mean (F-score)	0.490	

The miss and fault rates of the individual defect types can be seen in Table 3.4. The General strategy experienced fewer false alarms in every category except that of flat defects. When considering the miss rate exclusively, the Specific strategy performed better for dents and scratches, only slightly outperforming the General for flat defects.

Table 3.4: Comparing miss and false alarm rates for individual defect types

	Miss Rate		False Alarm Rate	
	General	Specific	General	Specific
Cut	9.73%	26.74%	0.20%	0.29%
Flat	11.15%	11.06%	0.52%	0.17%
Dent	30.12%	13.46%	2.11%	2.47%
Glue	10.61%	22.50%	1.45%	4.28%
Scratch	12.34%	5.20%	0.29%	0.32%

General condition participants took an average of 1.15 seconds per response which when compared to the 0.74 seconds per response taken using the Specific strategy was 36% longer. This meant the Specific condition participants took an average of 15 minutes and 16 seconds while General condition participants completed the study in 22 minutes and 56 seconds. This conforms to literature stating it takes longer to complete inspection tasks that have more target types (Kristofferson et al., 1973). Figure 3.3 the response times for both strategies improve over time in agreement with Neisser's findings (1963).

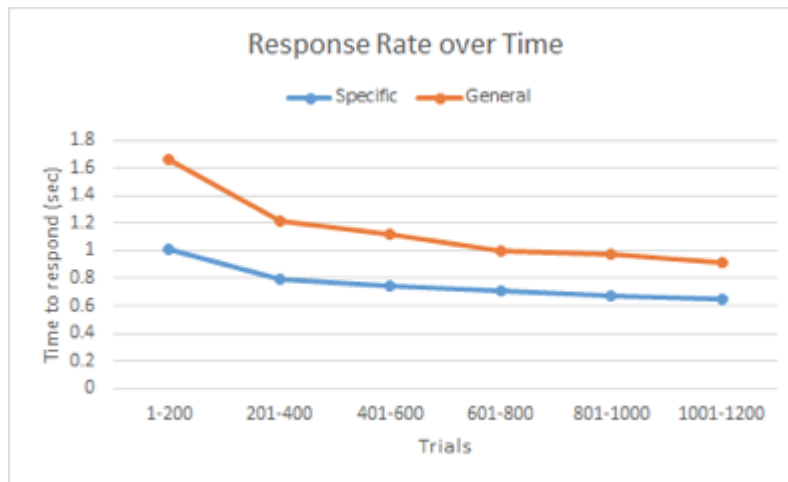


Figure 3.3: Response rate over time

As seen in Figure 3.4, the miss rate for both strategies is almost identical for most of the study until the Specific strategy miss rate spikes towards the end.

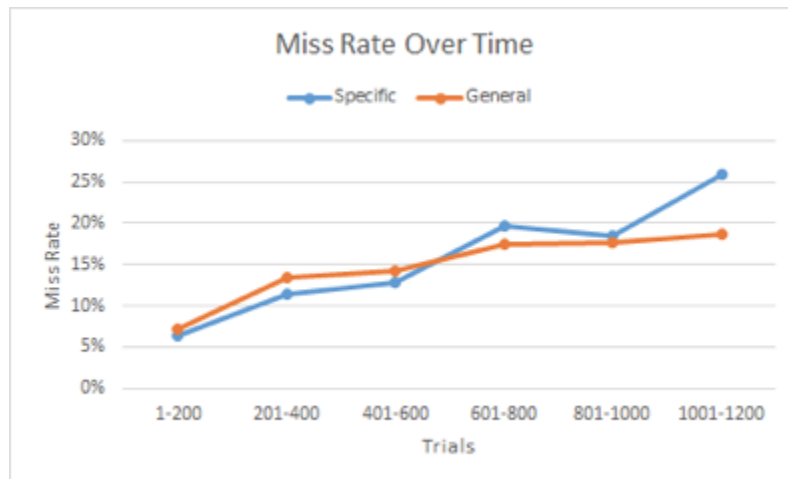


Figure 3.4: Miss rate over time

In Figure 3.5 the General condition starts with a higher false alarm rate. With time however, the participants looking for more defects improved in their ability to identify normal samples. It can be implied that the higher signal rate improved sensitivity to the different defect types over time, overcoming potential losses in vigilance. Participants in the Specific condition became poorer in their ability to identify normal samples.

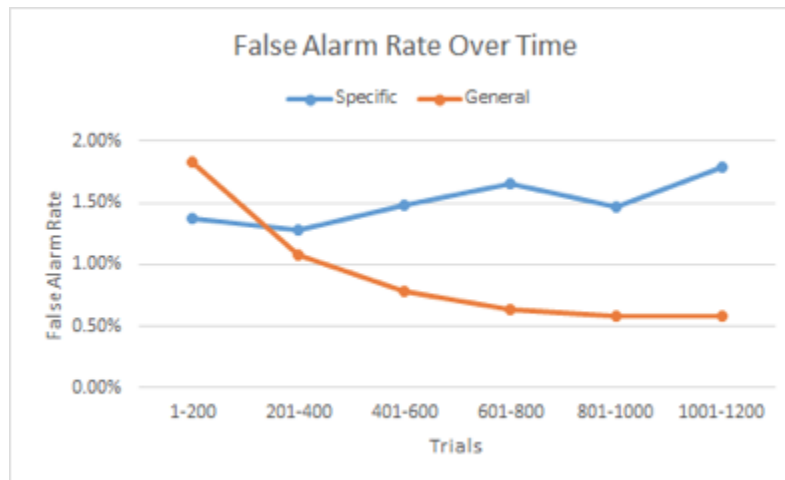


Figure 3.5: False alarm rate over time

A common behaviour observed during the study was the tendency to repeatedly tap the accept key. When a defect did occur, they indicated an accept response and then reacted (belatedly) by indicating the identified defect category for the next stimulus displayed. This was a common trend in the data, but it also showed a decline over time.

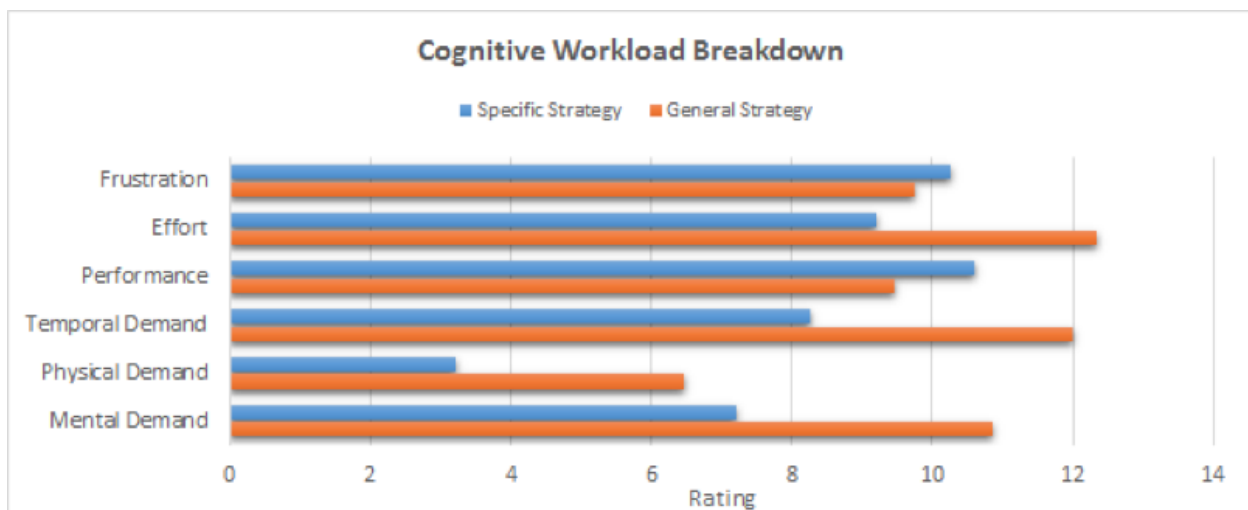


Figure 3.6: Average raw ratings from the NASA Task Load Index

Finally, the levels of workload reported in each condition using NASA TLX (Figure 3.6) are consistent with the general view that in the General strategy, participants had more to do (reflecting greater effort, and perceived temporal, physical and mental demands) whereas in the Specialised case participants experienced slightly more frustration. Interestingly, on average at least participants thought they performed better (or were more satisfied with their performance) when using the Specialised strategy; this is not necessarily supported by the performance data and more likely reflects the perception that the task itself as relatively easier to carry out (and thus do well at). However, as discussed earlier, it is a

defining feature of the vigilance decrement that performance deteriorates precisely because the task fails to offer sufficient stimulation over time.

3.1.3. Discussion

The General and Specific strategies lead to different performance profiles. The improved ability to correctly identify normal samples (correct reject) over time was noted in the General condition and could be attributed to an increase in sensitivity to the different defect categories with time. The increasing miss rate over time (seen in both conditions) indicates there is a vigilance decrement taking place, however this “learning effect” seems to counteract it when considering the false alarm rates of the General strategy. This reduction in false alarm rates supports the assertion that the General strategy actually improves the ability of participants to identify normal samples from defected ones.

This improved ability to identify normal samples is also reflected by a lower false alarm rates for almost all the individual defect types using the General strategy. The Specific strategy has a lower miss rate for the glue and scratch conditions which agrees with Su and Konz (1981) who stated that difficult defects are better detected when looking for one defect at time.

The Specific strategy allowed the participants to process all 1200 images nearly 8 minutes faster than their counterparts. However in order to attain complete coverage of all the defect types, more inspection time would be needed. In the case of this study, full coverage using the specific strategy would take in excess of an hour, and would result in more misses and false alarms. On the whole, the study managed to verify classic findings: there was an improvement in response rate over time (Neisser et al., 1963), an increase in response time for tasks with more target types (Kristofferson et al., 1973) and the better overall performance in the General strategy.

3.1.4. Conclusions

Compared using the inspection parameters prescribed by literature, looking for multiple feature types was slightly more effective overall than looking for a single type. This together with its more economic use of inspector time means that the General strategy could be considered a better overall paradigm when examining imagery in which multiple defect types are likely to present. The study also found that performance improved as a result of learning fairly rapidly while using the General strategy that led to a steady improvement in sensitivity consistent with a relatively brief exposure to the task.

The study also confirms that “harder to detect” features are more suited to a Specialised approach. However a General strategy lent itself to detection of the easier defect types. Overall the results could be taken as suggesting that the task design does influence and impact performance and therefore that a deliberate informed choice can be made depending on the nature of the stimuli (nature of the imagery and feature set of interest) and the kind of response profile that will generate the results desired.

3.2. Optimisation of change detection method

In the preceding experiment the ‘feature’ is always a in the form of a defect from a standard cork tile. However, when considering change detection within iMars imagery, each pair of co-registered images presented to a citizen scientist are likely to depict fairly widely differing terrain. Thus, we require an additional stage of processing, not just feature identification but a prior stage of change detection. Within the experimental platform we have developed using the Zooniverse Panoptes system, two options (and one variant) were generated. In our earlier literature review (D7.1) we identified that flicker was a promising option as a change detection technique as a change creates motion in the visual field orienting the observer to the source. However, it might also be the case that this mode of presentation would be perceived as irritating and citizen scientists might instead do better to compare images simply side-by-side. An additional potential option to this, identified through feedback from a pilot experiment, was that participants might prefer to manually control the flicker rate themselves. More critically, it might simply be the case that the task is impossible, however presented, to complete. Hitherto our analysis of the viability of change detection was based on small pilot and we extended this to compare the three options with actual Mars imagery.

3.2.1. Experimental Design

Using a within-subjects design, three different image comparison presentation styles were manipulated, side-by side, automatic flicker and manual flicker - where the participant used an on-screen button to change between each image. Three separate classification interfaces that varied in relation to these presentation styles were employed, again in conjunction with a questionnaire including NASA Task Load Index (TLX) type statements to assess volunteer opinion and perceived workload. The questionnaire also allowed ‘free-text’ responses so participants could raise issues and add context to their responses.

3.2.1.1. Participants

30 participants were recruited through email lists, social media posts etc. and were asked to attend the Nottingham Geospatial Institute at a set appointment time. All participants have been educated to a university-degree level, however none have had any formal training directly relating to planetary science. As such, this is representative of the education and experience regarding existing citizen science volunteer communities. Additionally, none have had any experience or have used other planetary citizen science platforms, such as Planet Four. Participants were gifted a £5 Amazon voucher for their participation in the study.

3.2.1.2. Apparatus & Materials

For the study, participants analysed a specially selected batch of 84 image pairs, 70 of which did not contain any changes and 14 that did. Of these 14, the features that changed were slope streaks, crater impacts, gullies or dunes (Figure 3.7). This imagery was made up of both HiRISE and CTX examples, taken from the work of Mattson and colleagues (Mattson et al., 2014).

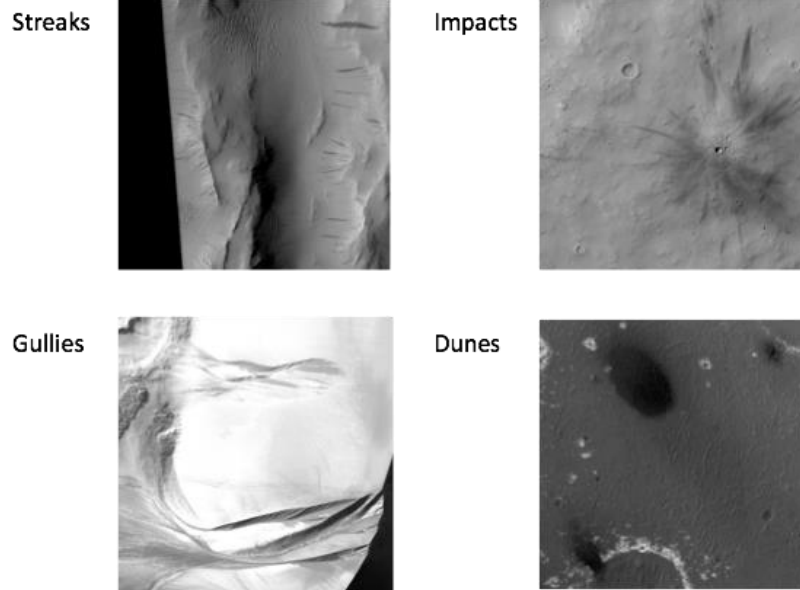


Figure 3.7: Surface features marked by participants

Before being uploaded to the platform, the image was ‘sliced’ into a number of smaller images that can be more easily handled. Original NASA imagery is often gigabytes in size, making it time-consuming to render to a web browser. The image ‘slices’ created were 450 x 450 pixels with an included overlap of 100 pixels to ensure features on the edges were adequately displayed.

Regarding the 3 interfaces, for the ‘side-by-side’ condition the ‘before’ and ‘after’ images were stitched together into one 450 x 950 image (the extra 50 pixels in width being a black ‘gap’ in the middle providing a separation) and displayed as shown in Figure 3.8.

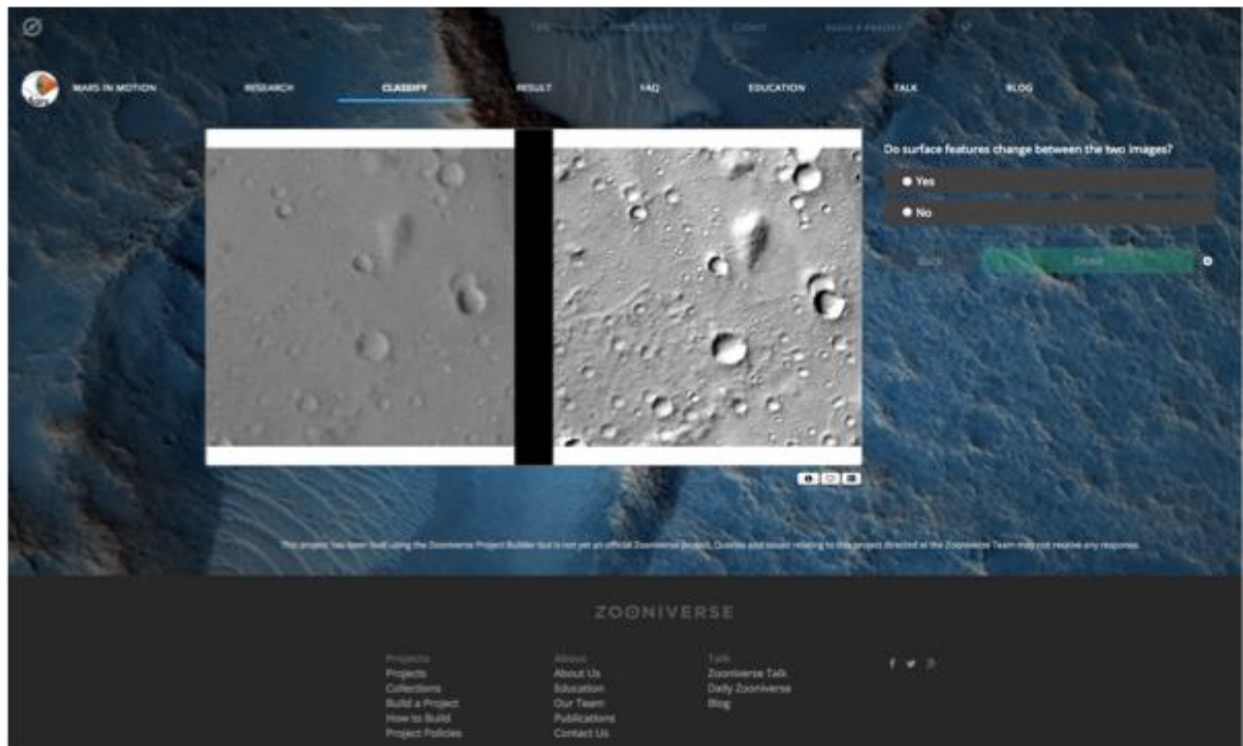


Figure 3.8: Study interface showing 'side-by-side' image presentation

Regarding the manual and automatic flicker conditions the 'before' and 'after' images of each pair were displayed overlapped. In the automatic example, each image was displayed in turn in a flicker type motion with an interval of 0.3 seconds. In the manual condition, the participant could see each image in turn at a rate of their choosing by clicking the circular 'switches' representing each image – as highlighted in the yellow box in in Figure 3.9.

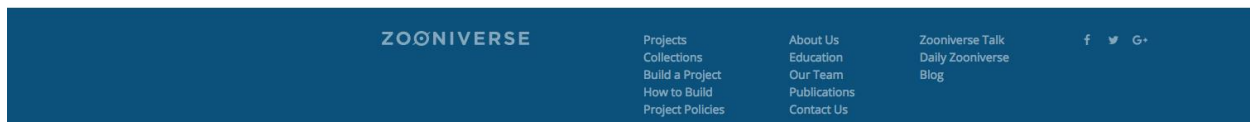
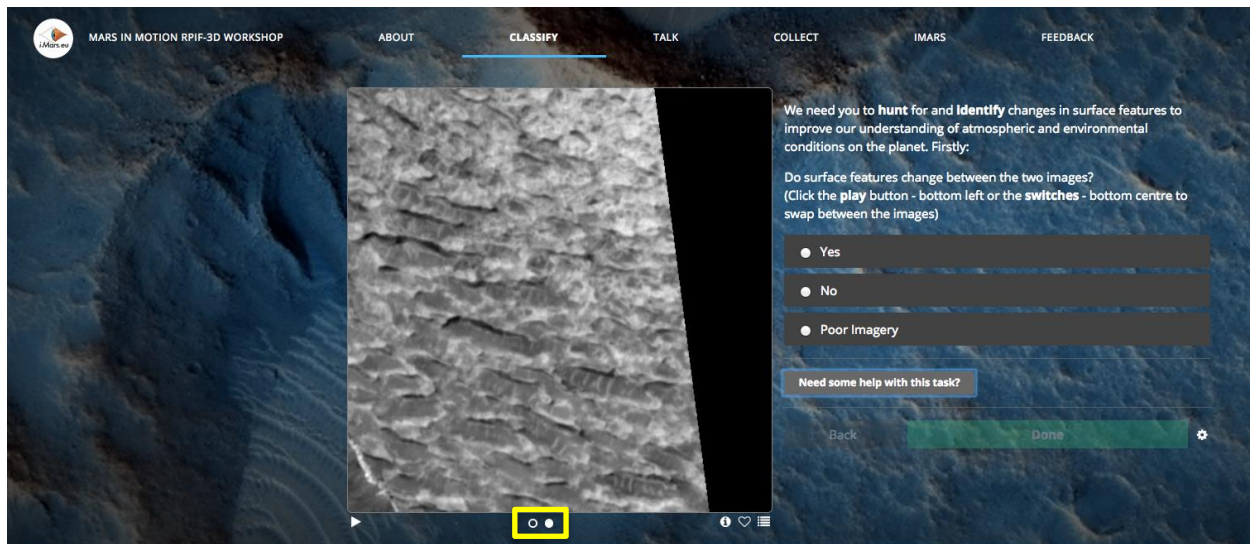


Figure 3.9: Study interface showing ‘flicker’ image presentation

3.2.1.3. Experimental Procedure

All study participants came to the same room (individually) and carried out the experiment on the same laptop, to keep factors such as lighting conditions and screen setup constant and ensure that they did not influence the image analysis task. Before using each interface, each participant completed an online tutorial to learn how to use the tools, looking for change on a separate example image. Participants then used each of the interfaces in turn to for 10 minutes - answering a simple yes/no question for each image pair: *“do surface features change between the two images?”*; to mitigate bias caused by learning of the system, the order in which the interfaces were presented was manipulated so that the same number of participants tested the interfaces in the same order. The order in which image pairs were displayed to each participant was also randomised, to prevent bias being caused by image content (images with or without change appearing in the same interface each time etc.). After using each interface, participants completed the questionnaire to share their views as previously described.

3.2.2. Results

3.2.2.1. Participant Feedback

Participant responses to a number of statements included in the questionnaire showed no significant difference between each interface, perhaps to be expected as many of the design features, tasks performed and images displayed are constant throughout the experiment. However, differences between the interfaces emerged in participants’ scores when considering performance and physical effort (figure 3.10).

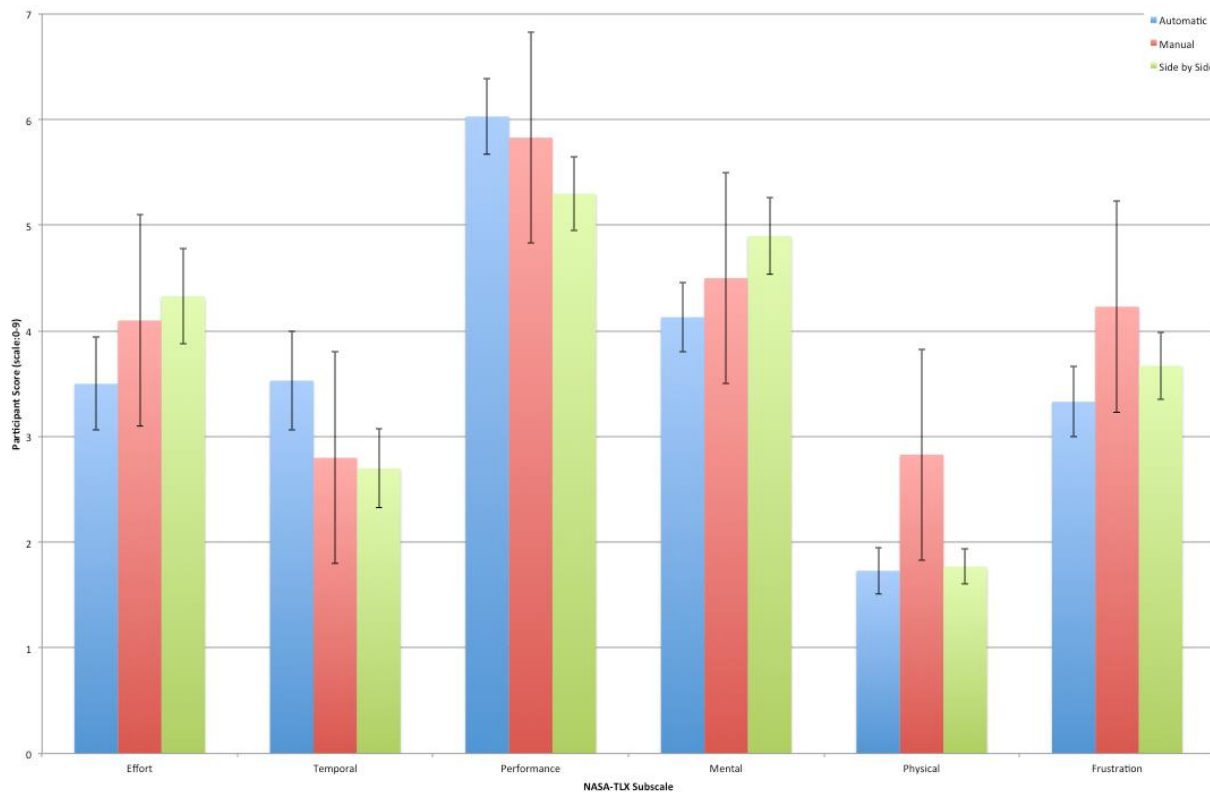


Figure 3.10: NASA-TLX participant responses for each interface (with standard error shown)

When considering their own performance, a repeated measures ANOVA with a Greenhouse-Geisser correction shows a significant difference in scores between interfaces ($F(1.987, 57.622) = 3.288, p = 0.045$). Post hoc tests using the least significant difference correction revealed that participants felt more successful using the automatic interface (6.03 ± 1.96) compared to the manual flicker (5.83 ± 1.82) and side-by-side interfaces ($5.30 \pm 1.91, p = 0.05$). When considering the statement “The amount of physical effort required” again there was a significant difference ($F(1.584, 45.928) = 6.436, p = 0.006$) where participants thought the manual flicker interface took the most physical effort (2.83 ± 1.78) compared to the automatic ($1.73 \pm 1.20, p = 0.024$) or the side-by-side ($1.77 \pm 0.89, p = 0.039$).

In addition to their responses to the Likert-type statements, several participants also provided ‘free response’ replies that add context to their scores, for example:

Regarding automatic flicker:

“Flickering too fast, distracting a bit as well, gives the sense of having to rush, although I probably managed to compare more pictures than in the other 2 tasks - I would make the break between changing the images longer...”

Manual flicker:

"This task was more user-friendly because I wasn't feeling stressed by the quick change of the image, I could give my attention to the details I wanted to and this gave me more self-confidence."

General rate of change:

"...getting more accustomed to the task (and the images) and not finding any differences it became a bit frustrating this time round."

"Because I couldn't find any changes I felt a bit confused about doing the right stuff. It would be nice to show at least one set of images with changes so that someone could make sure (or at least have the chance) is not losing the changes between images."

Participants raised a number of issues and concerns both regarding specific methods of image presentation and also with the general task of change detection itself.

3.2.2.2. Results: Performance

Figure 3.11 shows participant performance using each of the interfaces, in terms of the time spent analysing each image pair and the amount correctly classified (with or without change). A repeated measures ANOVA with a Greenhouse-Geisser correction shows a significant difference in time spent analysing each image between interfaces ($F(1.997, 165.764) = 36.178, p = 0.001$). Post hoc tests using the least significant difference correction revealed that participants analysed imagery significantly quicker using the automatic flicker interface (average of 9.0 ± 5 seconds, $p = 0.001$) compared to the manual flicker (15.0 ± 5 seconds) and side-by-side interfaces (13.0 ± 4 seconds). Additionally, a Friedman test revealed a statistically significant difference in the percentage of correct classifications between interfaces ($X^2(2) = 22.268, p = 0.001$). Post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction applied revealed that participants using the automatic and manual flicker interfaces were significantly more accurate at detecting change ($93\% \pm 0.1$ correct and $92\% \pm 0.13$ correct respectively) compared to when using the side-by-side interface ($87\% \pm 0.17$ correct, $p = 0.001$).

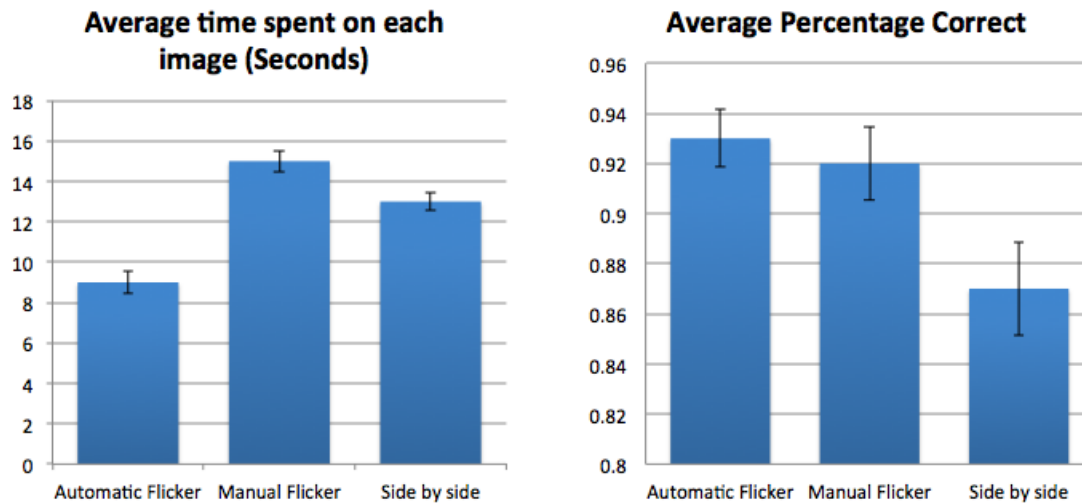


Figure 3.11: Participant performance when detecting change on the Martian surface

Breaking down the results further, Figure 3.12 shows performance in terms of the different types of change that occurred: gullies, impact craters, slope streaks, dunes and no change. In general the same pattern is true independent of the type of change, with the automatic flicker interface being the quickest, and the side-by-side interface the least accurate. However, a difference can be seen in terms of accuracy when considering images with change and those without. When an image pair did feature a change, participants were between 64-77% successful at detect it depending on the feature (i.e., a miss rate between 23 and 36 depending on the feature). When there was no change, participants were 95% successful at correctly responding (i.e., a false alarm rate of 5%). This balance between miss and false alarm is broadly characteristic of similar tasks (Green & Swets, 1968) and also the foregoing experiments.

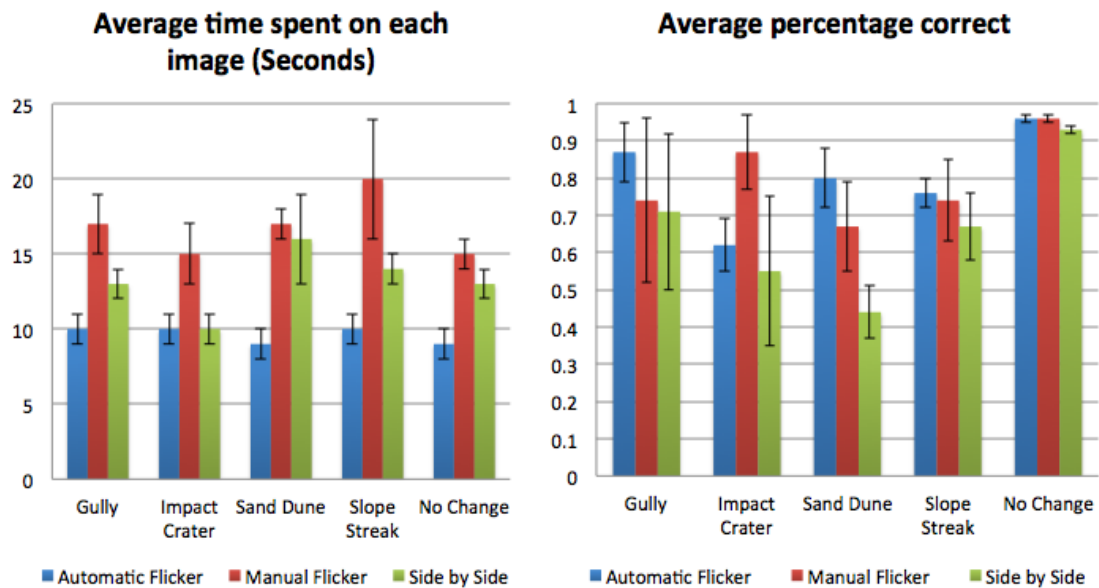


Figure 3.12: Participant performance in terms of the types of change present

3.2.3. Conclusions

Using a laboratory study to test the effect of different image presentation methods on participants' ability to detect change on the surface of Mars, it was found that the way that image sets are shown to the volunteer can influence both their experience and performance. Self-reported NASA-TLX responses suggested that the automatic flicker interface gave the greater perception of success, a feeling supported by the performance results - with volunteers using the automatic interface being both faster and more accurate. On the opposite side of the coin, NASA-TLX responses suggested that the manual flicker interface required the greatest physical workload. This does not tell the whole story however, as participants' free-text responses suggested a frustration of being 'rushed' by the automatic interface, much preferring the autonomy afforded by the manual flicker design. Independent of the interface used, participant accuracy was greatest when detecting no change over change.

The general implication here is that flicker leads to better change detection performance with ecologically valid imagery (this contrasts to a prior pilot experiment we ran where a reversed set of results were found when nameable objects were substituted for more complex surface features). However, it is also confirmed that while automatic flicker increases the rate with which change can be detected without significantly harming performance, the temporal demand inherent might become an irritation over time. As both manual and automatic flicker have essentially the same "form factor" on-screen and in terms of setting up citizen tasks, we conclude that simply giving participants the option to switch between both options will be optimal. Another implication for the iMars platform concerns the analysis of images with no change - is this a 'real' result or do participants default to this position when they are unsure of the correct response. A final observation is that with ecologically valid Mars imagery with its inherent variability, we might have been concerned that this inherent visual complexity might have generated a higher false alarm rate than the 5% we found and thus pollute the dataset with false

positives that might be hard to detect in the absence of the ground truth we have here. While miss rate can be addressed through simply raising the numbers of volunteers “eyes on”, false rates may, depending on the scheme of aggregation, prove harder to remove. This also suggests to us that participants benefitted from the induction offered and despite the complexity of morphological features on Mars, could generally discern between artefactual differences between imagery and the underlying differences caused by features.

3.3. Trust in the Crowd vs. Trust in the Machine

3.3.1. Introduction

An important link between WP7 and other iMars work packages resides with the change detection algorithm under development in WP6. Previous Citizen Science projects, most notably in the Space Warps project (Marshall et al., 2015, Marshall et al., 2016, More et al., 2016) have already demonstrated an impressive potential for a crowd to train computer algorithms and facilitate the ‘deep’ learning required for an algorithm to be able to process images and identify, let alone classify, changes as effectively and efficiently as the human visual system. Such is the volume of Martian images from the last forty years that it would be both inefficient for the iMars project to ask volunteers to process every image pair for changes and ethically unreasonable; especially consider the large swathes of the Red Planet that are likely inactive and unchanged. Even if image pairs for these regions have been co-registered, it is necessary to consider carefully the experience of volunteers.

Alternatively, we might consider other possible relationships between volunteer and algorithmic change detection within a pipeline. Perhaps volunteers and algorithms work independently and their outputs only brought together in aggregate at a late stage. Alternatively one might construe of a situation where human data trains an algorithm or alternatively, where an algorithm effectively screens the dataset for potential change and the product of this is passed on to humans for further examination. One might, furthermore, also consider where the crowd stands in relationship to itself; should we for example consider splitting the task into detection and identification activities? And in all these scenarios, should we give some sort of indication of these prior assessments?

We investigated the effect perceived by participants when a change detection algorithm has pre-filtered the images they see and the images are visually colour-coded to convey to volunteers whether the algorithm found a change or not. As the following sections will describe, we experimented with the accuracy of this algorithm, or the threshold at which it found change, to see how its performance affected participants’. We then extended this to experiment with the provision of a crowd’s judgement, and the strength of their judgement. This feeds into the wider societal interest in trust in automation and would inform not only how the change detection algorithm and crowd-sourced data would interplay within the iMars project.

3.3.2. Experiment C1: Trust in algorithms

3.3.2.1. Experimental Design

Five interfaces were designed to present two images taken of the same place on the Martian surface but at different times and asked participants to ‘Spot the difference’. Participants were told that image pairs had been assessed in advance by an algorithm, which classified image pairs as either 1) an area in which the surface has changed or 2) an area in which the surface has not changed.

The accuracy of the algorithm was varied across five sets of images (Table 3.). One of the sets, labelled A, represented an algorithm that was 100% correct in its classifications i.e. the borders of all images were coloured correctly, according to whether it was an area in which real surface change had occurred or not. Two more sets of image pairs were coloured correctly for 75% of both the correct and incorrect imagery, with a further two with image pairs coloured according to the presence of change with 50% accuracy (Table 3.5). Two sets of each were created to mitigate against the possibility that changes in some feature types are easier to identify than others, anticipated in our review and demonstrated in earlier experiments.

Table 3.5: Variation in the accuracy of the information provided by the algorithm

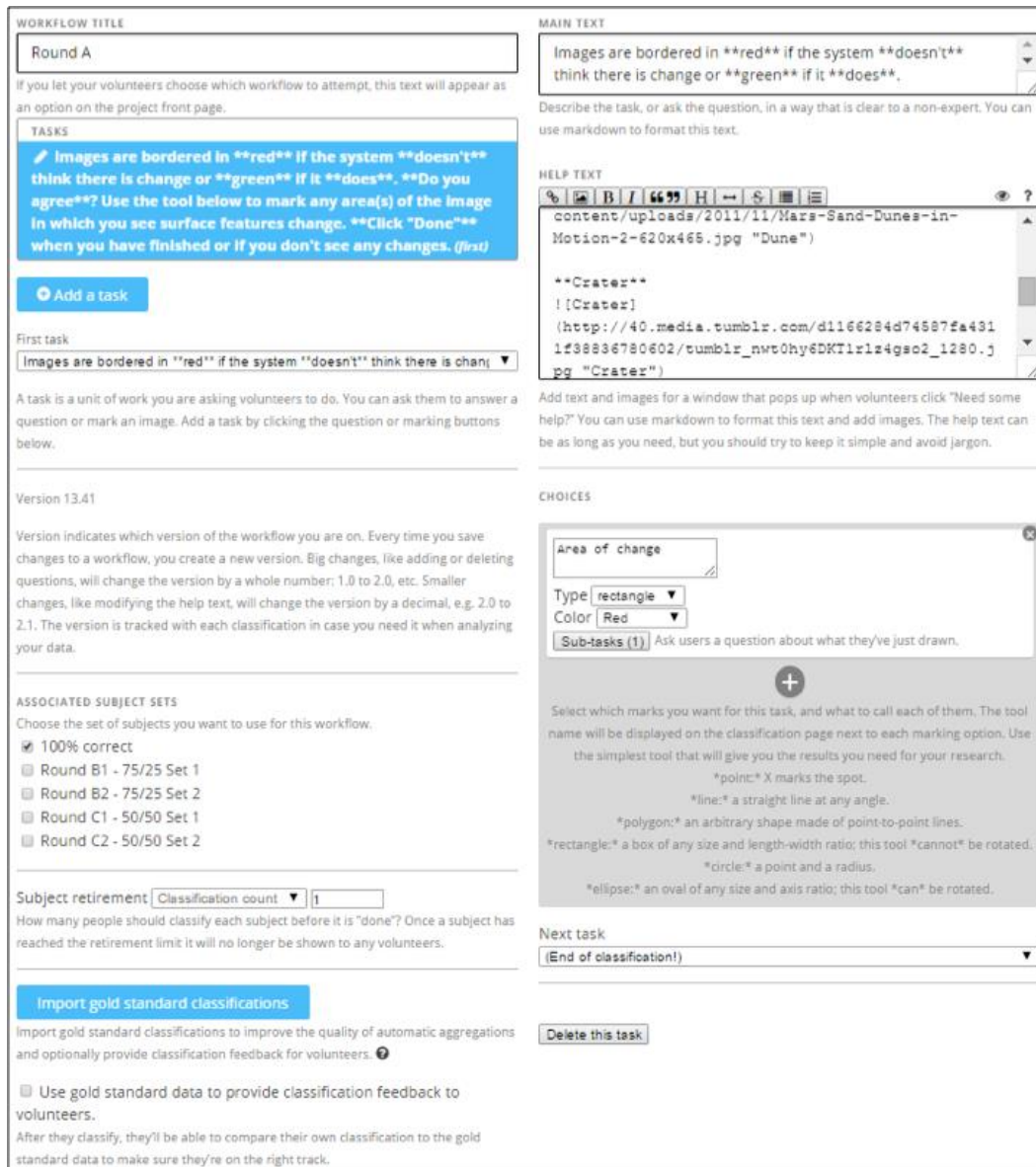
Correct	Image Set	Incorrect
100%	A	0%
75%	B1 (Movement: Gullies & dunes)	25% (New: Streaks & impact craters)
	B2 (New: Streaks & impact craters)	25% (Movement: Gullies & dunes)
50%	C1 (New: Streaks & impact craters)	50% (Movement: Gullies & dunes)
	C2 (Movement: Gullies & dunes)	50% (New: Streaks & impact craters)

3.3.2.2. Participants

24 participants were recruited through email lists, social media posts etc. and were asked to attend the Nottingham Geospatial Institute at a set appointment time. All participants have been educated to a university-degree level, however none have had any formal training directly relating to planetary science. As such, this is representative of the education and experience regarding existing citizen science volunteer communities. Additionally, none have had any experience or have used other planetary citizen science platforms, such as Planet Four. Participants were gifted a £5 Amazon voucher for their participation in the study.

3.3.2.3. Apparatus & Materials

For these experiments, the Zooniverse project builder (www.zooniverse.org/lab/) was used to create a project with a different workflow for each of the conditions, linking from the same homepage but to different subject sets (Figure 3.13).



The screenshot displays the Zooniverse Project Builder interface for configuring a workflow titled "Round A". The interface is divided into several sections:

- WORKFLOW TITLE:** "Round A". Below this, a note states: "If you let your volunteers choose which workflow to attempt, this text will appear as an option on the project front page."
- TASKS:** A blue box contains the task description: "Images are bordered in **red** if the system **doesn't** think there is change or **green** if it **does**. **Do you agree?** Use the tool below to mark any area(s) of the image in which you see surface features change. **Click 'Done'** when you have finished or if you don't see any changes. (text)". Below this is a blue button labeled "Add a task".
- First task:** A dropdown menu shows the task description: "Images are bordered in **red** if the system **doesn't** think there is change". Below this is a note: "A task is a unit of work you are asking volunteers to do. You can ask them to answer a question or mark an image. Add a task by clicking the question or marking buttons below."
- Version 13.41:** A note explains: "Version indicates which version of the workflow you are on. Every time you save changes to a workflow, you create a new version. Big changes, like adding or deleting questions, will change the version by a whole number; 1.0 to 2.0, etc. Smaller changes, like modifying the help text, will change the version by a decimal, e.g. 2.0 to 2.1. The version is tracked with each classification in case you need it when analyzing your data."
- ASSOCIATED SUBJECT SETS:** A note says: "Choose the set of subjects you want to use for this workflow." Below this are four checkboxes:
 - ☒ 100% correct
 - ☐ Round B1 - 75/25 Set 1
 - ☐ Round B2 - 75/25 Set 2
 - ☐ Round C1 - 50/50 Set 1
 - ☐ Round C2 - 50/50 Set 2
- Subject retirement:** A dropdown menu shows "Classification count" and a value of "1". Below this is a note: "How many people should classify each subject before it is 'done'? Once a subject has reached the retirement limit it will no longer be shown to any volunteers."
- Import gold standard classifications:** A blue button labeled "Import gold standard classifications". Below this is a note: "Import gold standard classifications to improve the quality of automatic aggregations and optionally provide classification feedback for volunteers." Below this is a checkbox:
 - ☐ Use gold standard data to provide classification feedback to volunteers.
 A note follows: "After they classify, they'll be able to compare their own classification to the gold standard data to make sure they're on the right track."
- MAIN TEXT:** A text area contains the task description: "Images are bordered in **red** if the system **doesn't** think there is change or **green** if it **does**." Below this is a note: "Describe the task, or ask the question, in a way that is clear to a non-expert. You can use markdown to format this text."
- HELP TEXT:** A text area contains the help text: "content/uploads/2011/11/Mars-Sand-Dunes-in-Motion-2-620x465.jpg 'Dune'" and "Crater" followed by a URL. Below this is a note: "Add text and images for a window that pops up when volunteers click 'Need some help?' You can use markdown to format this text and add images. The help text can be as long as you need, but you should try to keep it simple and avoid jargon."
- CHOICES:** A section for configuring marking tools. It includes a dropdown for "Area of change", a "Type" dropdown set to "rectangle", and a "Color" dropdown set to "Red". Below this is a button labeled "Sub-tasks (1)" with a note: "Ask users a question about what they've just drawn." Below this is a plus sign icon and a note: "Select which marks you want for this task, and what to call each of them. The tool name will be displayed on the classification page next to each marking option. Use the simplest tool that will give you the results you need for your research." Below this are several tool descriptions:
 - *point: X marks the spot.
 - *line: a straight line at any angle.
 - *polygon: an arbitrary shape made of point-to-point lines.
 - *rectangle: a box of any size and length-width ratio; this tool *cannot* be rotated.
 - *circle: a point and a radius.
 - *ellipse: an oval of any size and axis ratio; this tool *can* be rotated.
- Next task:** A dropdown menu shows "(End of classification!)". Below this is a button labeled "Delete this task".

Figure 3.13: Zooniverse.org Project Builder

The experiment used the same 84 HiRISE image pairs used in the prior experiment (sourced from Mattson et al., 2014), of which 14 contained changes on the surface. To ensure that surface changes did not appear in 70 image pairs, we simulated subtle shifts in lighting with open source software

'Imagemagick'; we could not use the same image twice because the transition between images on the screen was impossible to discern (which could make the pairs with no surface change too obvious).

Two important precautions were taken to further reduce confounding effects:

1. The project builder presented images to participants in a random order, to lessen potential learning effects and randomise how many changes each participant saw;
2. The order in which participants used interfaces of different algorithm accuracy was randomised to assuage biases in their performance and views due to the accuracy with which the algorithm had classified the first image set participants saw (be it completely accurate or only half accurate).

3.3.2.4. Experimental Procedure

Each participant completed the experiment in the same room but at different times. On arrival, participants received an explanation of the project and their task, before they signed a paper copy of the information sheet and consent form, as approved by the Faculty of Engineering's Ethics Board. The debrief emphasised that participants should mark changes on the surface and not changes in lighting or image quality, for example, that might occur due to differences in atmospheric or photographic conditions. When the participant indicated that they understood what was required and had no further questions they completed an introductory questionnaire; this captured basic demographic data, which our experience and previous research suggested might support data analysis.

Following a demonstration of the task, participants had the opportunity to complete the task for one image pair to familiarise themselves with the user interface. Figure 3.14 shows the task that confronted participants. Images were bordered in red if the algorithm suggested there was no change in the two images, and green if the algorithm suggested there was a change. Participants, however, were invited to judge the images independently and to mark where on the images they saw change using the rectangular drawing tool provided, coloured red and labelled "Area of change". If they did not see any changes, they clicked on "Done" and moved to the next image pair. If they did see a change and draw a rectangle, however, a window would pop and ask them "What surface feature have you marked?" Underneath the question were four choices, from which they could only select one with its radio button.

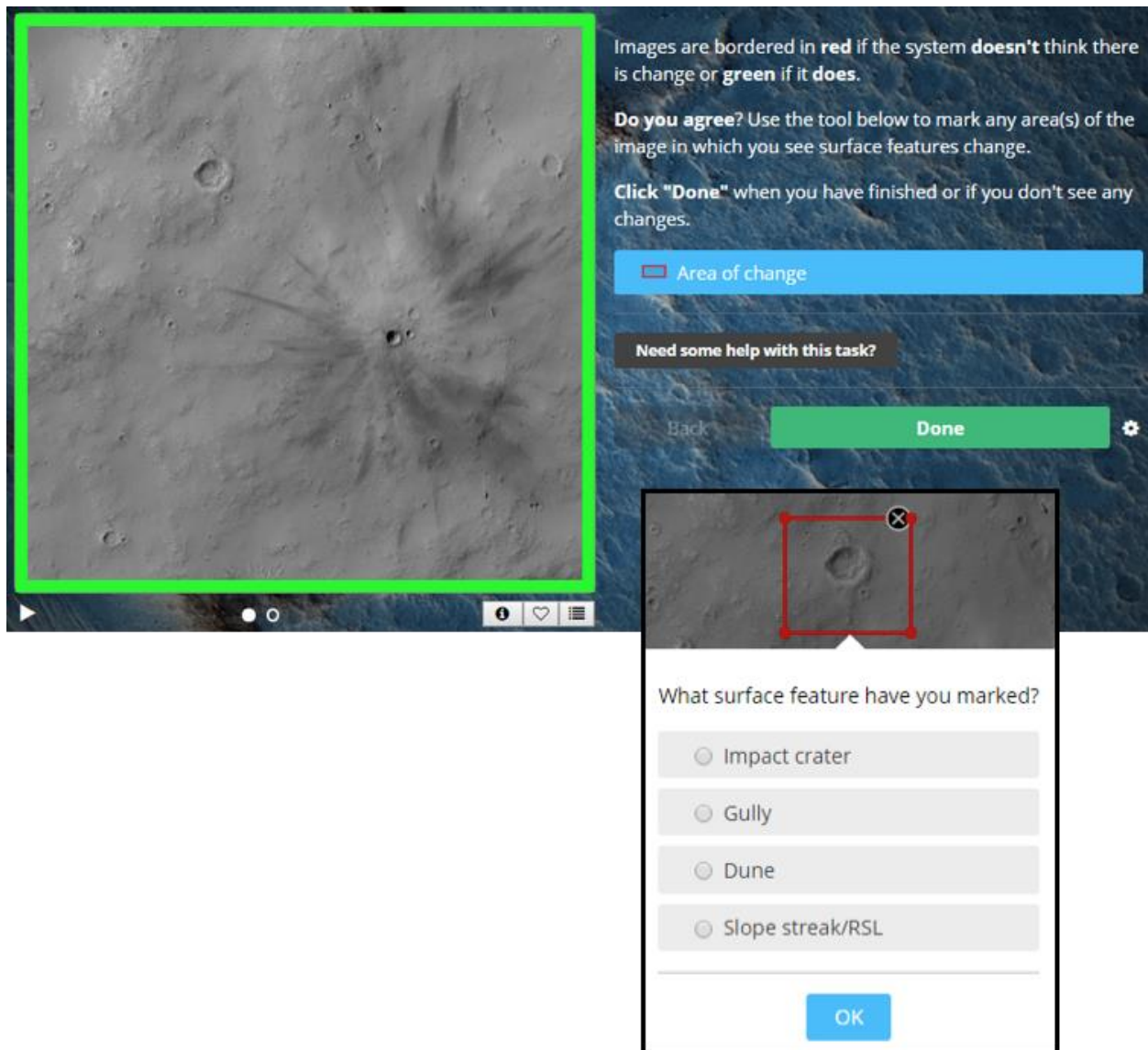


Figure 3.14: Classification pages

Importantly, to reinforce the types of change participants should mark, their attention was drawn to the 'Help' button, which they could click at any point during the experiment to see examples of the four feature types change they mark (Figure 3.15). They then proceeded to work through images at their own pace using the first interface for ten minutes.

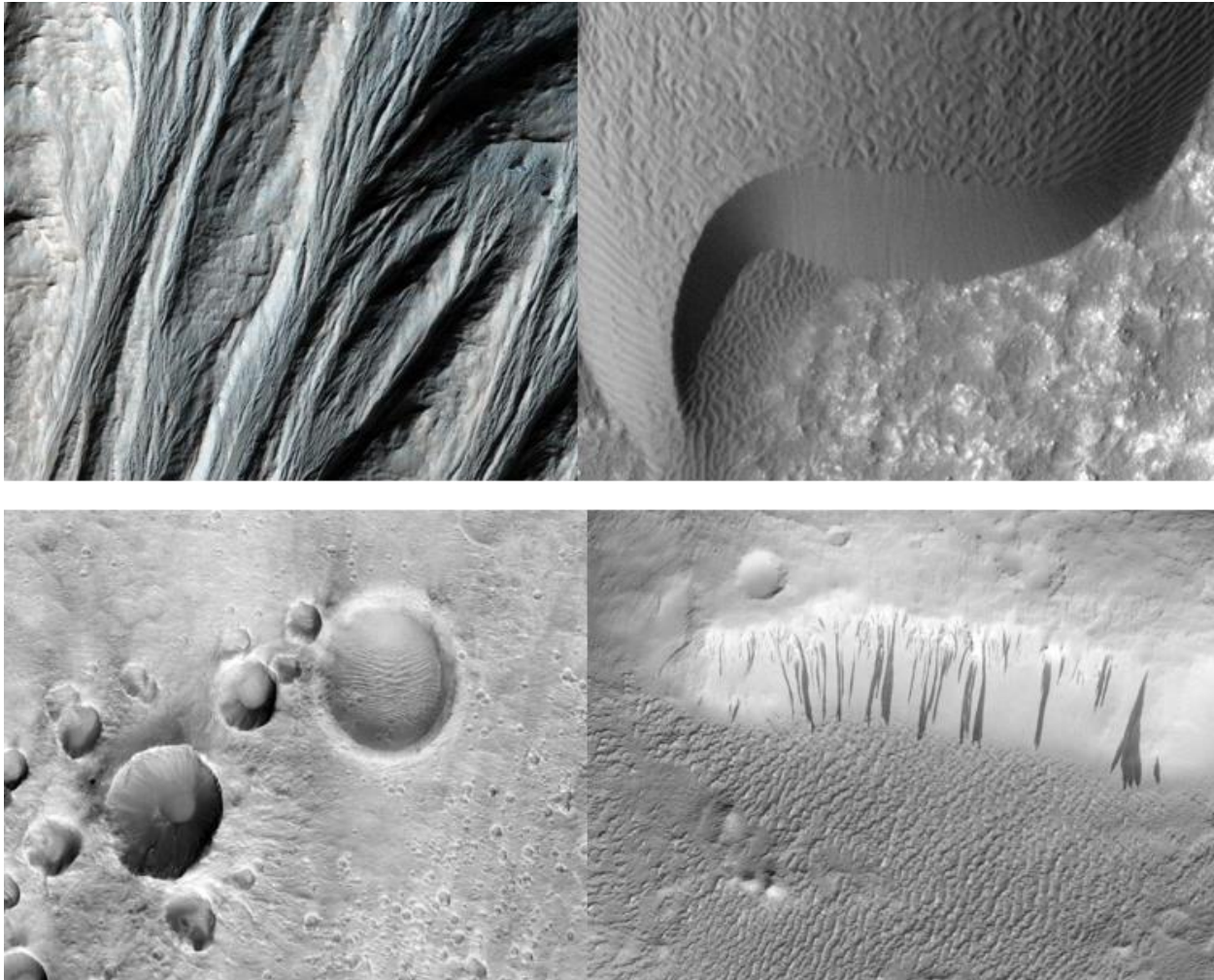


Figure 3.15: Examples of surface changes provided to participants.
Clockwise from top left: Gullies, sand dunes, slope streaks and impact craters

After ten minutes, participants evaluated the website's design, usability and imagery and the task. A survey, hosted on <http://www.onlinesurveys.ac.uk> for data protection, collected Likert scale responses to statements for comparative analysis; participants could add context to their answers by typing in free text boxes. The survey derived measures from the NASA Task Load Index (TLX), used in previous research to measure workload regarding a number of on-screen and HCI type tasks and described earlier in Section 3.1. Additionally a text box allowed participants to explain their answers in order to add context to the findings.

Finally, the survey asked participants to rate their trust in the 'change detection' algorithm result they were shown for each image pair, by responding to the following questions in terms of a 9 point scale where 1 indicated an extremely low extent (not at all) and 9 an extremely high extent (completely):

- To what extent did the computer detect change effectively?
- To what extent could you predict the computer's behaviour with some degree of confidence?
- To what extent was the computer free of errors?
- To what extent do you have a strong belief and trust in the computer to do a particular task effectively for which there may be no proof?
- To what extent do you trust the computer overall?

The qualitative data just described was collected to supplement the quantitative data that the project builder captured of the classifications participants made, available for download as a spread sheet (CSV file). The project builder records each classification made with rich data including: the time that each classification was made (from which we can derive how long participants took to classify each image); the names of images under review (for analysis of false positives and negatives); and the image set under review (for analysis of the impact of the algorithm's accuracy on the performance measures captured).

Participants then repeated this for the two other interfaces. The whole procedure took a maximum of one hour to complete. To mitigate bias caused by learning of the system, the order in which the interfaces were presented was manipulated so that the same number of participants tested the interfaces in the same order. The order in which image pairs were displayed to each participant was also randomised, to prevent bias being caused by image content (images with or without change appearing in the same interface each time etc.).

3.3.3. Experiment C2: Trust in the crowd

3.3.3.1. Experimental Design

A follow-on experiment was performed in which each image pair came with information on a crowd's 1) current consensus (Yes or No: if there was a surface change or not), and 2) strength of consensus (as a percentage) via a button in the corner of the image window. This differentiates this experiment from the previous one in two significant ways: 1) Participants controlled when they sought the crowd's current collective wisdom; they could decide to look at it after they had made their classification, or not consult it at all, in contrast to the coloured borders of Experiment 1 And 2). The information provided them with the strength of the consensus, which may or may not affect their agreement or otherwise with the crowd. Table 3.6. shows a break down of the experimental conditions for each image set in terms of the strength of consensus and their accuracy.

Table 3.6: The Nine Accuracy/Strength of Consensus Conditions for the Crowd Experiment

Image Set	Accuracy (% correct)			Strength of Consensus (%)		
	Change	No Change A	No Change B	Change	No Change A	No Change B
A1	100%	100%	100%	51-65	66-80	81-95
A2				66-80	81-95	51-65
A3				81-95	51-65	66-80
B1	50%	Incorrect	Correct	51-65	66-80	81-95
B2				66-80	81-95	51-65
B3				81-95	51-65	66-80
C1	50%	Correct	Incorrect	51-65	66-80	81-95
C2				66-80	81-95	51-65
C3				81-95	51-65	66-80

3.3.3.2. Participants

36 new volunteers participated in the experiment. As before, they were recruited through email lists, social media posts etc. and were asked to attend the Nottingham Geospatial Institute at a set appointment time. All participants have been educated to a university-degree level, however none have had any formal training directly relating to planetary science. As such, this is representative of the education and experience regarding existing citizen science volunteer communities. Additionally, none have had any experience or have used other planetary citizen science platforms, such as Planet Four. Participants were gifted a £5 Amazon voucher for their participation in the study.

3.3.3.3. Apparatus & Materials

The same platform created through the Zooniverse's Panoptes framework for the previous experiment described in section 5.2 was used. The experiment also used the same imagery as the previous one for consistency and direct comparison of results.

3.3.3.4. Experimental Procedure

Participants tested three versions of the interface for ten minutes over an hour in total, which included time for the introduction and filling in a feedback survey after each one. Each participant completed ten

minutes on one of the three (A) interfaces (see table 3.2) in which the crowd had classified all (change and no change) imagery correctly but to differing levels of consensus. Likewise they spent ten minutes on one of three (B) interfaces in which half of the imagery containing changes, and half the imagery containing no changes, was classified incorrectly to different levels of consensus; for the final ten minutes participants carried out the task with an image set (C) in which the incorrect and correct imagery of the second image set was switched, so that results were due to the algorithm's accuracy.

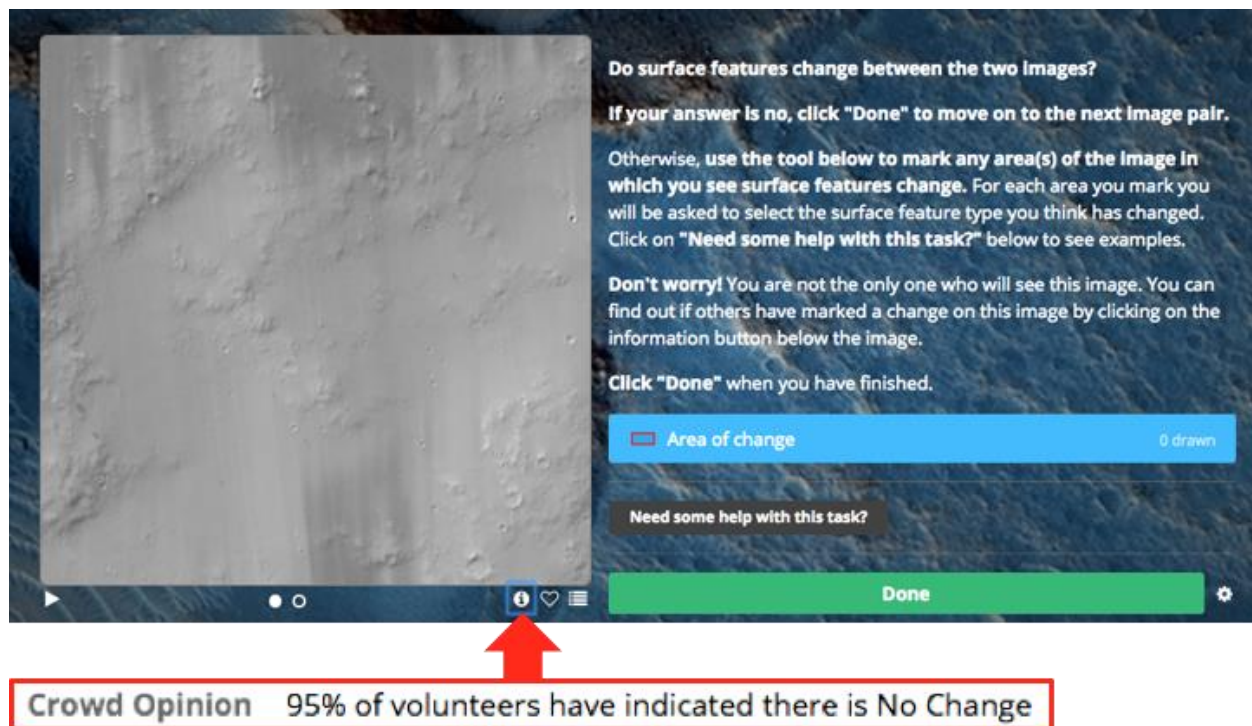


Figure 3.16: Classification page for crowd experiment, highlighting how participants were able to access information about the crowd's current consensus.

At any point while using the interfaces the participant could click the 'information' icon on the screen (indicated by the red arrow in figure 3.16) to see the crowd opinion regarding if change was present, and the strength of the crowds' consensus. Again the order in which the interfaces were presented was varied to prevent learning effects, along with the order of the imagery presented. As with the previous experiment, participants also evaluated the website's design, usability and imagery and the task immediately after using each interface through a survey as previously described.

3.4. Experiment C1 Results: The Computer

NASA-TLX scores were analysed for statistical differences related to the algorithm's accuracy. Participants scored their success and satisfaction significantly lower when the task contained images only 50% accurately identified to contain change/no change (Figure 3.17).

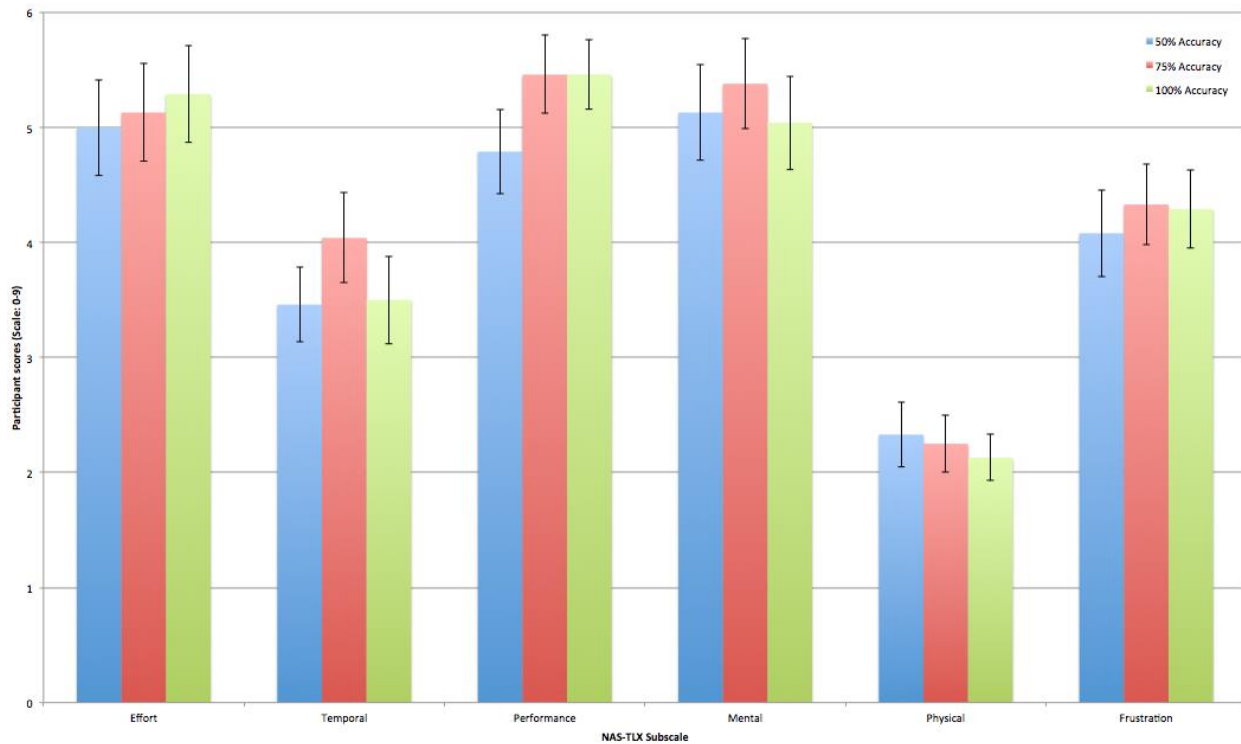


Figure 3.17: NASA-TLX scores for the three levels of algorithm accuracy

Trust Scores

Participants' scores for the final statements, regarding trust, showed significant differences when the colour of the images' border was correct 50% of the time, compared to when it was 100% correct. Participants scored all but the last question (regarding trust in the computer overall) significantly lower when the task used images only 50% accurately classified to contain change/no change (Figure 3.18).

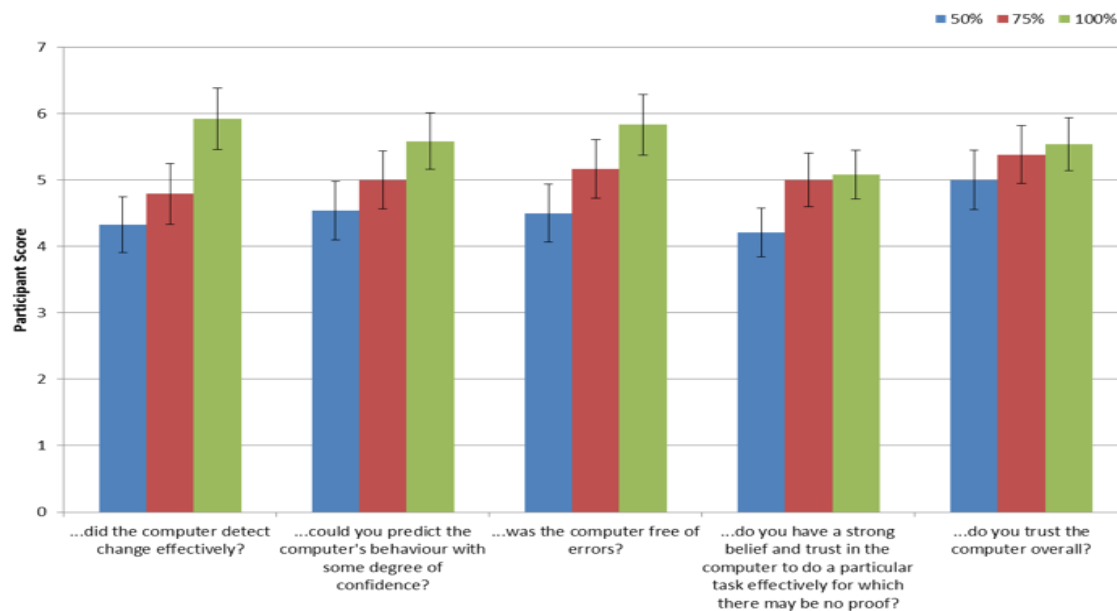


Figure 3.18: Trust scores according to algorithm accuracy

Notable

quantitative

data

- Time taken on images did not vary;
- Participants identified image pairs with no surface changes consistently well, regardless of the algorithm's accuracy;
- Participants identified the feature that has changed more accurately when the change/no change algorithm has performed better, especially features that are generally more difficult to identify (e.g. gullies 0% vs 40%);
- Participants appear to be more vigilant when they use a system that they trust.

Comments

Participants supplemented their scores with comments, which could be added to a text box at the end of each question. The following quotes are indicative of the totality of comments:

"The differences in the shadows in the images makes it difficult to be confident of spotting differences."

"By the end I felt I was just ignoring the red/green borders because I disagreed with them so much."

"Display a degree of certainty or uncertainty with the image to give an idea of how much I need to check. If it's 100% sure that there are no changes, I'm not going to give it as much attention as if it's only 80% sure."

"Choose the speed and number of times the images flash back and forth, as I found it too slow and three times to then click again was annoying."

The most interesting result, however, is that the trust scores showed that participants perceived a difference, even if it was subconscious, and this is illustrated in the next section.

3.5. Experiment C2 Results: The Crowd

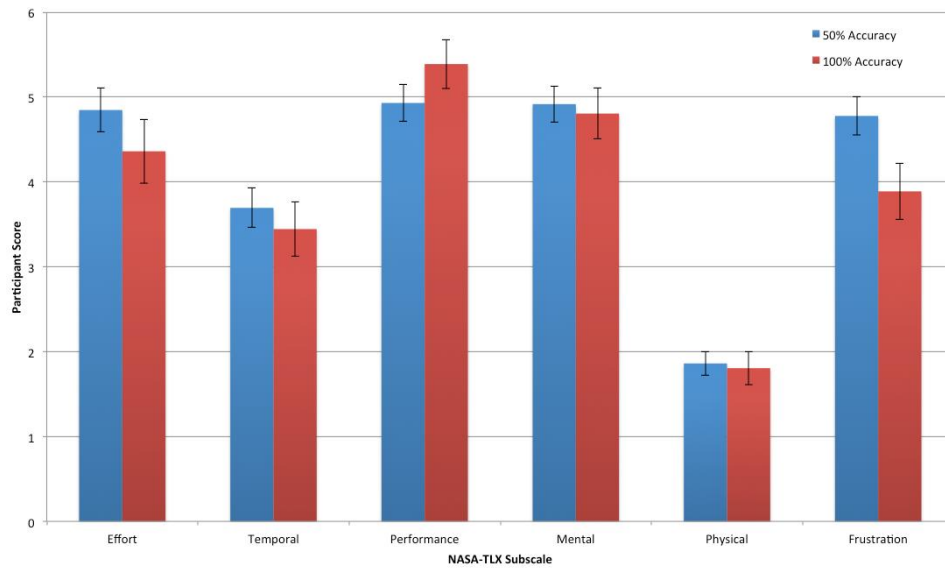


Figure 3.19: Graph of NASA-TLX scores when the crowd was 50% and 100% correct.

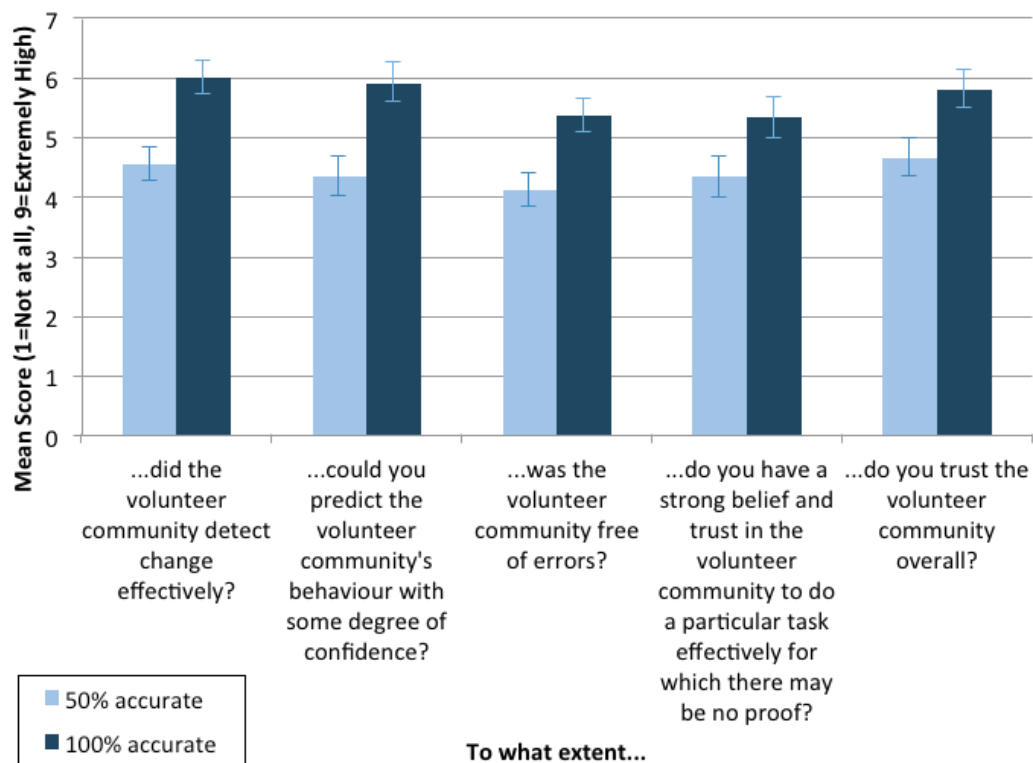


Figure 3.20: Trust in the Crowd

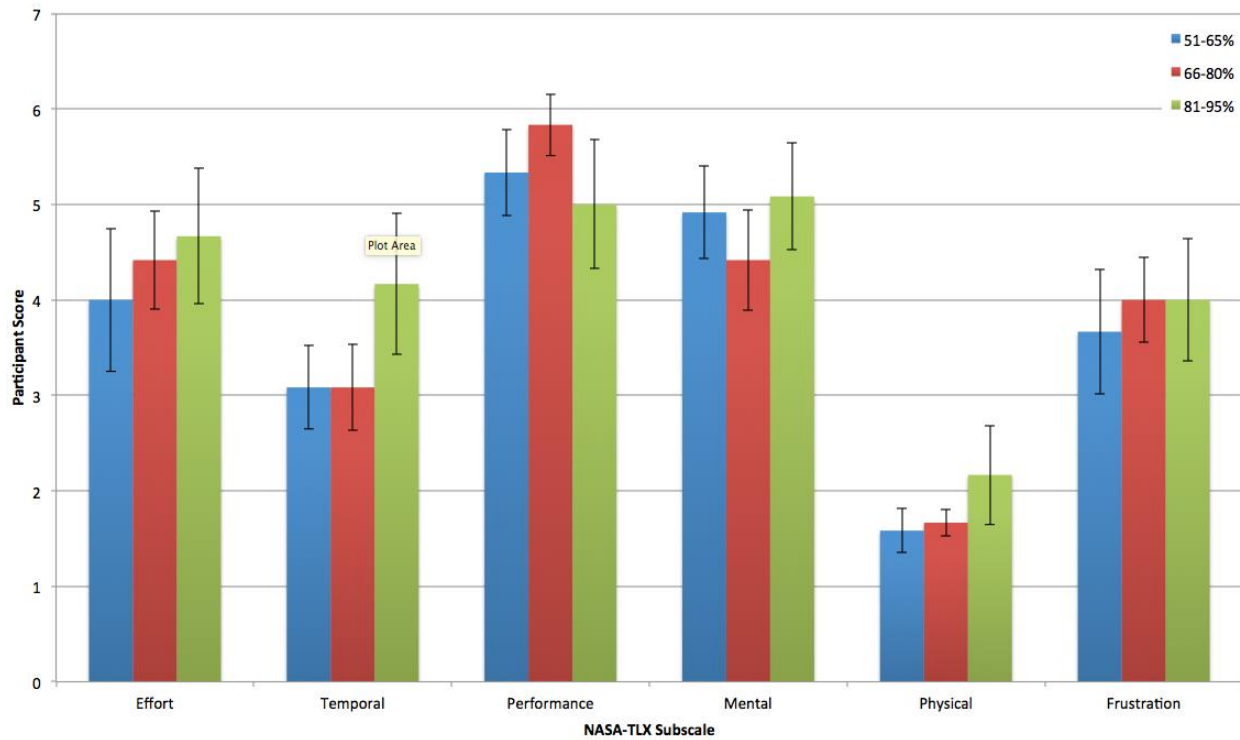


Figure 3.21: Graph of NASA-TLX Scores for the different percentages of crowd agreement

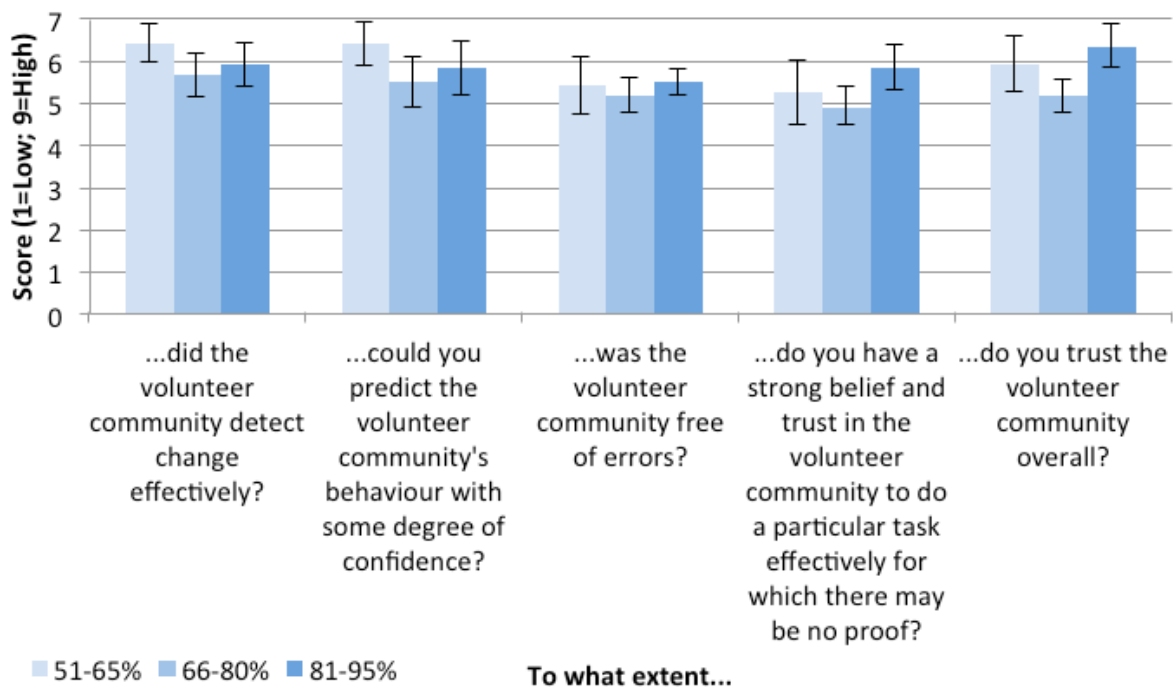


Figure 3.22: Graph of the Scores for Trust Statements for different levels of crowd agreement

Figures 3.19-3.22 present the main qualitative findings. The two results that are most important to point out are that participants reported feeling significantly more insecure, discouraged, irritated and stressed and less trusting of the crowd, when it was 50% accurate compared to a 100% accurate crowd (Figures 3.21 and 3.22). Figure 3.23 also shows that the lower the strength of consensus of a crowd's agreement the more tired participants felt. These findings merit deeper analysis and investigation.

Participants expressed uncertainty about their performance when the crowd's judgement was only 50%, as the following quotes illustrate:

"I felt completely incapable of determining whether the community was right or not. I personally think they were very wrong a great deal of the time, but the percentages were so overwhelming that I can't claim I'm certain that I'm right."

"Often off putting to know that the majority of people had put there was a change and you couldn't find one (and visa versa)"

These comments contrast with those made when the crowd's judgement was 100% correct; participants appeared to trust the crowd more and considered it more carefully when forming their own judgement:

"Used the community opinion as an pointer, checking my own decision, my faith being in numbers rather than skill."

"I did not allow the crowd to sway my decision. I trusted my own judgment rather than the majority where there was one."

"When I checked, the percentages were in the 50s to low 60s, meaning that even if the community was correct overall, a lot of people were very wrong."

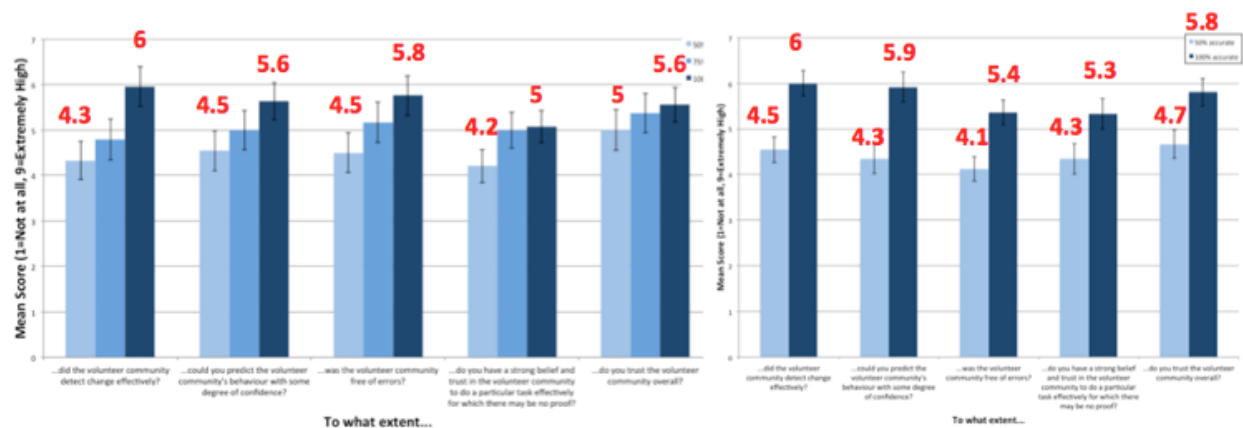


Figure 3.23: Comparison of trust scores for the two experiments

3.6. Discussion

The most interesting finding from these two experiments was the difference in participants' attitude towards an inaccurate crowd compared to an inaccurate algorithm. They were much more forgiving of

an inaccurate algorithm but became more sceptical of the crowd more quickly. This finding might be viewed as contrary to previous findings that humans would be more trusting of flawed human judgements than statistically less flawed algorithms (e.g., Dietvorst, Simmons & Cade, 2015). However, on reflection it is to be noted that we compared the machine to the crowd represented merely by numerical strength of judgement, not – if we can entertain the semantics of this distinction – as such a collection of humans. From the subjective comments made by participants, it appears that they instead equated the failure of the crowd with online phenomena such as trolling. One speculative explanation of these findings is that humans are normally trusted more on the basis that good intentions can be imputed to them; on this occasion this might have backfired as *nefarious* motives were applied to a crowd that could not be tied to a computer algorithm.

The first implication relates directly to the project under study and for the provision of metadata in the final Citizen Science project. It also hints at the care we must take if the change detection algorithm is used to filter images; the proportion of image pairs that show change should be as high as possible maintain a desirable level of quality and quantity of volunteered data. Follow on studies might remove the information all together, in a more realistic simulation of a live project, so that any influence on volunteered judgements arising from the algorithm's accuracy would be subconscious, and to provide a control for what happens without the disclosure of others' judgements.

The finding that participants can rapidly, in a single trial, come to a fairly accurate and consistent view (measured here in terms of subjective ratings of trust) of the performance of an algorithm or crowd suggests care must be taken in terms of the task data pushed to participants. If participants enter the workflow after an algorithm that produces perfect performance, our data suggest that they will quickly realise this and will recognise the redundancy of their efforts (even though we may have other reasons to retain their involvement). Equally, where performance is too low, participants are probably best left uninformed of crowd performance as the evidence from the present study suggests this leads to a high level of frustration (noting that this advisory information actually has no specific impact on a participant completing the task itself; this remains the same in all circumstances). In summary, it is important to balance issues of under- and over-trust (see Lee & See, 2004) together with the general effects of vigilance decrement discussed earlier in this report; these variables can all be addressed by monitoring the hit rate in a project and being prepared to intervene to modify it if required.

4. MSSL Workshop

4.1. Introduction

As an integral part of iMars outreach activities, in early June of 2016, the Europlanet training workshop on "3D facilities available at the UK NASA RPIF" to introduce early-career scientists to the range of software tools available for the generation of 3D products. This include both DTMs and terrain corrected orthorectified images (ORIs), using the "NASA-USGS SOCET+ISIS" and the more recent UCL modification of the open source "UCL-NASA AMES stereo pipeline" that was developed within the EU FP7 iMars

project. Ways of data handling and data fusion as well as tools for digitising geological and geomorphological features from the resultant ORIs and DTMs within the COTS ArcGIS and the open source QGIS were also introduced. WebGIS tools developed within iMars WP5, for selecting appropriate Mars orbital data, was introduced. Training was provided in the use of tools on the iMars webGIS developed for the display of change detection (WP6), and 3D flyovers (featuring examples developed for our own outreach in iMars WP8).

Additionally, attendees participated in an experiment of the ‘Mars in Motion’ Citizen Science project. The dual aims of this were for planetary scientists to contribute data that could be used for comparison with: 1) the volunteered data in the live project, and also 2) the results of the automated change detection algorithm from WP6 in order to refine its success at detecting surface changes, over and above other differences between images.

4.2. Method

22 planetary scientists participated in the experiment, comprising PhD students and post-doctoral researchers funded by Europlanets to attend a workshop on 3D data. Europlanets 2020 Research Infrastructure is a European Commission Horizon 2020 project to integrate and support planetary science activities across Europe. The requirements for participation in the workshop ensured that participants had the necessary planetary imagery expertise to provide “expert” data. Participants again signed an information sheet and consent form, which were approved by the Faculty of Engineering’s Ethics Board independently of those used previously. This guaranteed they understood what we were asking them to do, why, and their permission to use their classifications for the purposes described. Participants could leave and/or request for their data to be removed at any time.

The experiment was described to them and they registered on www.zooniverse.org during half an hour before a lunch break. This had two benefits: first it gave attendees time to consider their participation and ask questions about it during the lunch break, and second it enabled a punctual start after lunch, without distraction from the task at hand (Figure 4.1, Figure 4.2 and Figure 4.3). The design of the task was informed by the experiments reported in previous sections in several ways. These included the types of features they were asked to annotate (Section 2.2), the detection of multiple rather than single feature types (Section 3.1) and the automatic mechanism for flicking between the images (Section 3.2); the speed of flickering between the images was pre-determined by Panoptes and was not changed because participants in the experiments detailed in Section 3.2 did not report any problems.

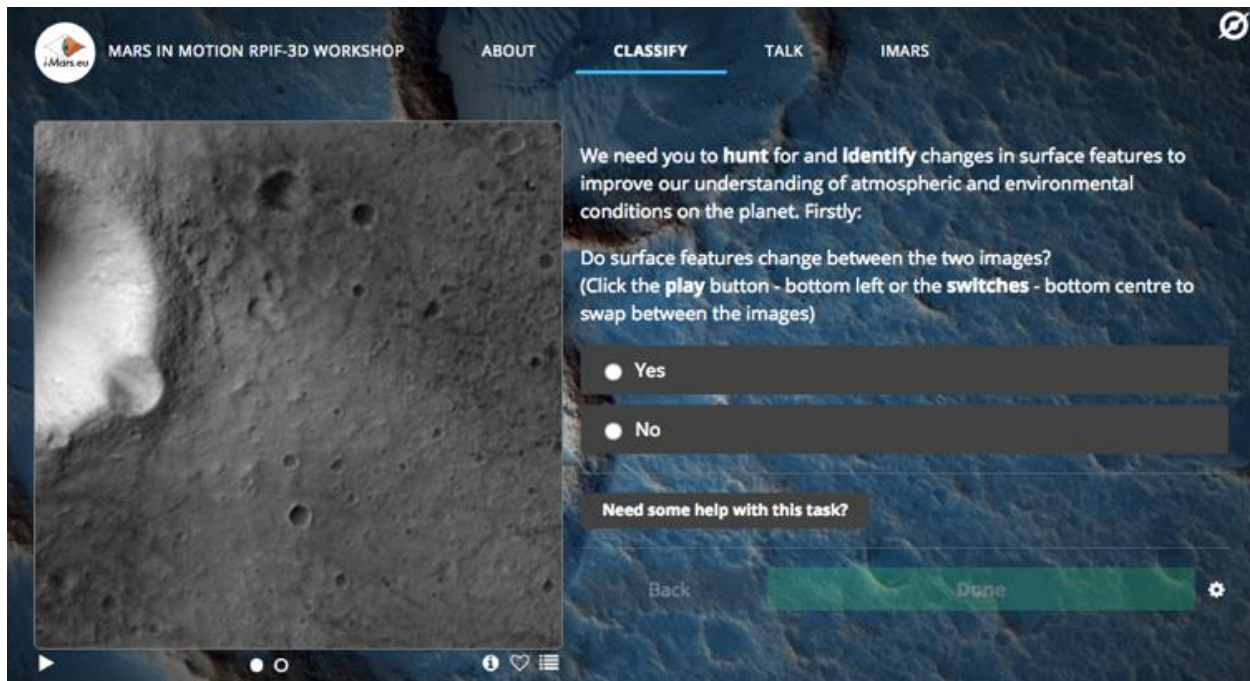


Figure 4.1: The task for workshop participants

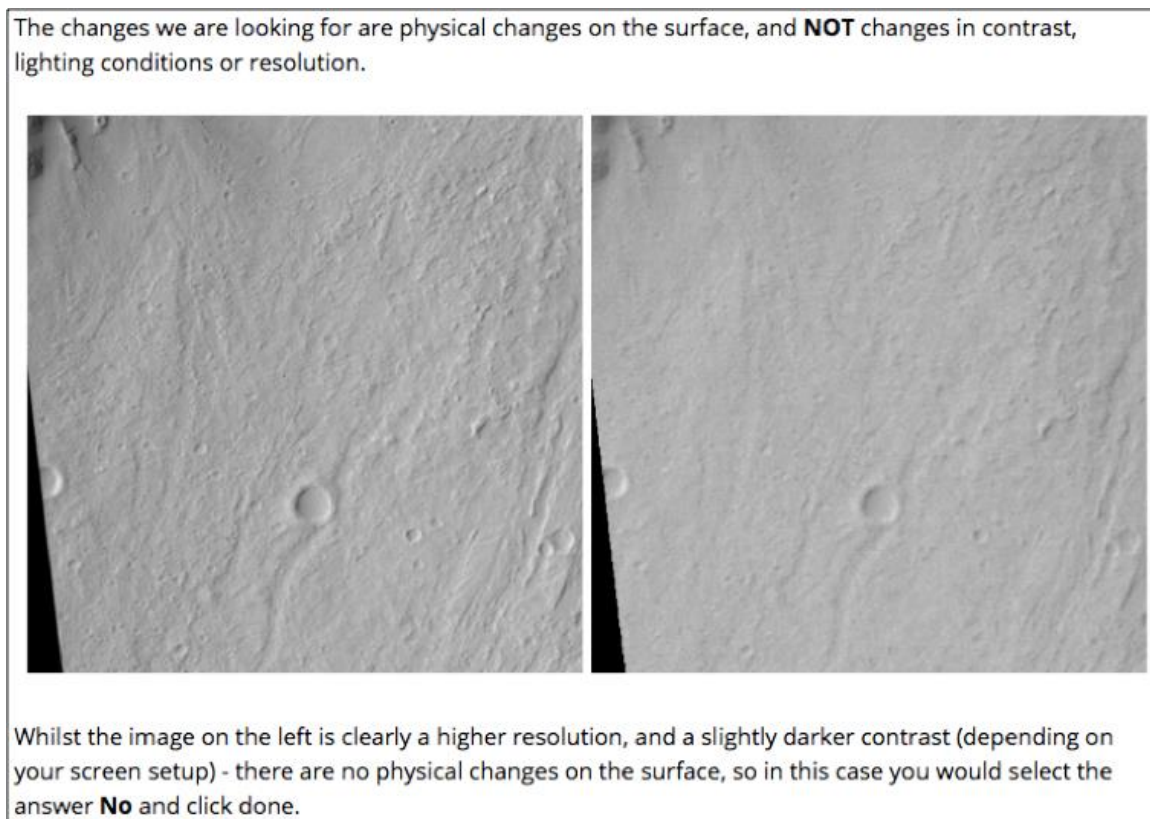


Figure 4.2: The 'help' text provided to workshop participants

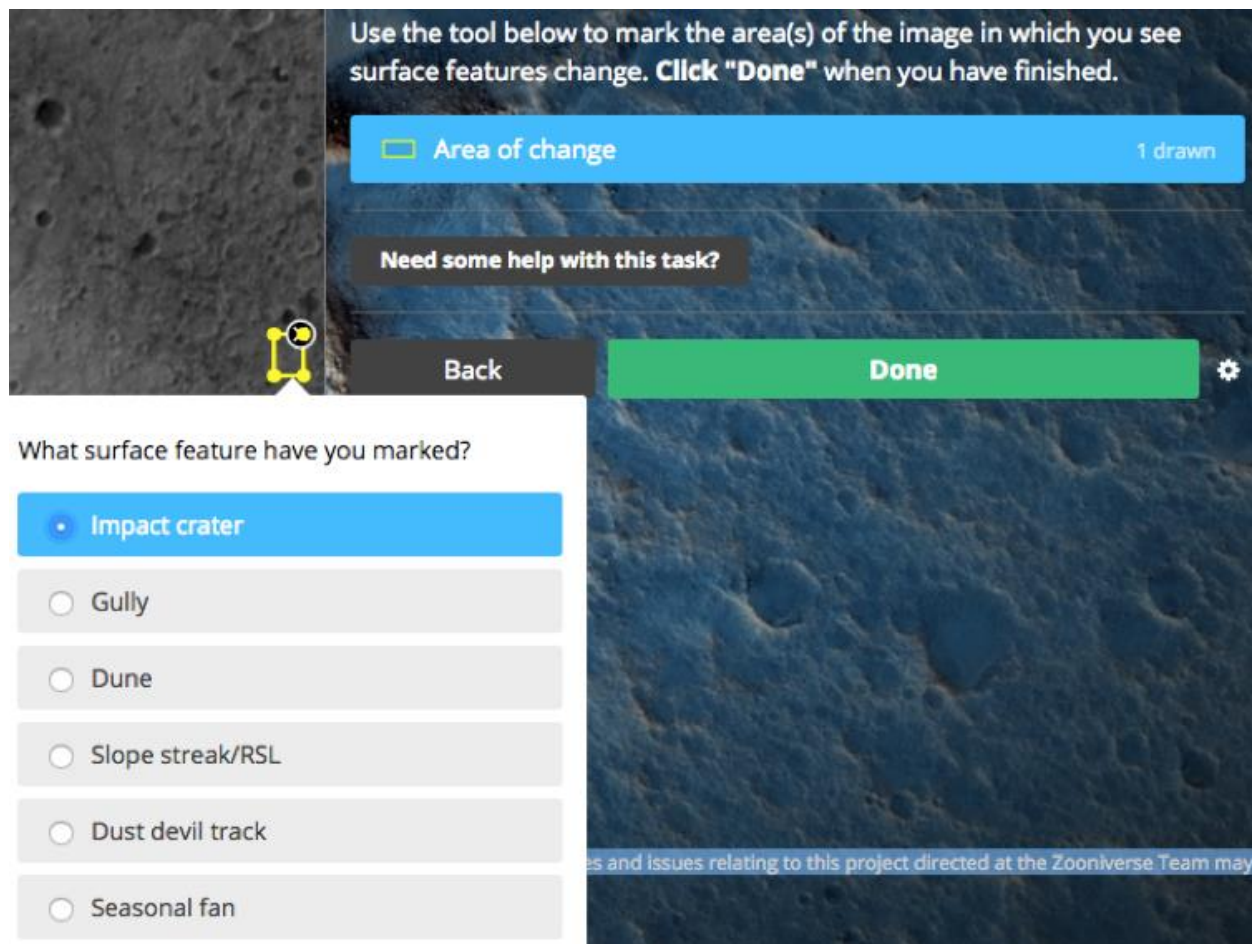


Figure 4.3: The Subtask of Feature Identification. Notice changes to the feature list in Figure 6.2.

4.3. Results

In total 1,877 classifications were recorded within the time allotted to Citizen Science in the two-day workshop, of which 1553 (82.7%) reported no change and 324 (17.3%) indicated change. Of these 324 classifications, 142 were singletons (only identified once) and 182 were repeated. This is reassuring in terms of the parameters of the task itself as it suggests a good level of events commensurate with keeping participants interested and motivated. Table 4.1 presents the number of images in which each feature type had changed, and also the number of individual changes annotated for each feature type; this type of information can inform the development of the live project as it indicates the performance of the current automated change detection algorithm and which features have seen the most change.

The TLX scores for this group of experts show the same ordinal relationship of task workload factors as in prior studies in this report suggesting that familiarity/expertise with the subject matter had no real impact. This is worth noting particularly with regard to the relative relationship between frustration, mental effort and other factors and hints that the pattern may remain stable amongst citizen scientists even as they get more experienced with the task at hand.

Table 4.1: Workshop classification data

Feature	Number of images	Number of unique features
Impact crater	31	34
Gully	11	15
Dune	38	48
Slope streak/ Recurring Slope Lineae	90	141
Dust Devil Track	138	192
Seasonal Fan	19	26

NASA-TLX

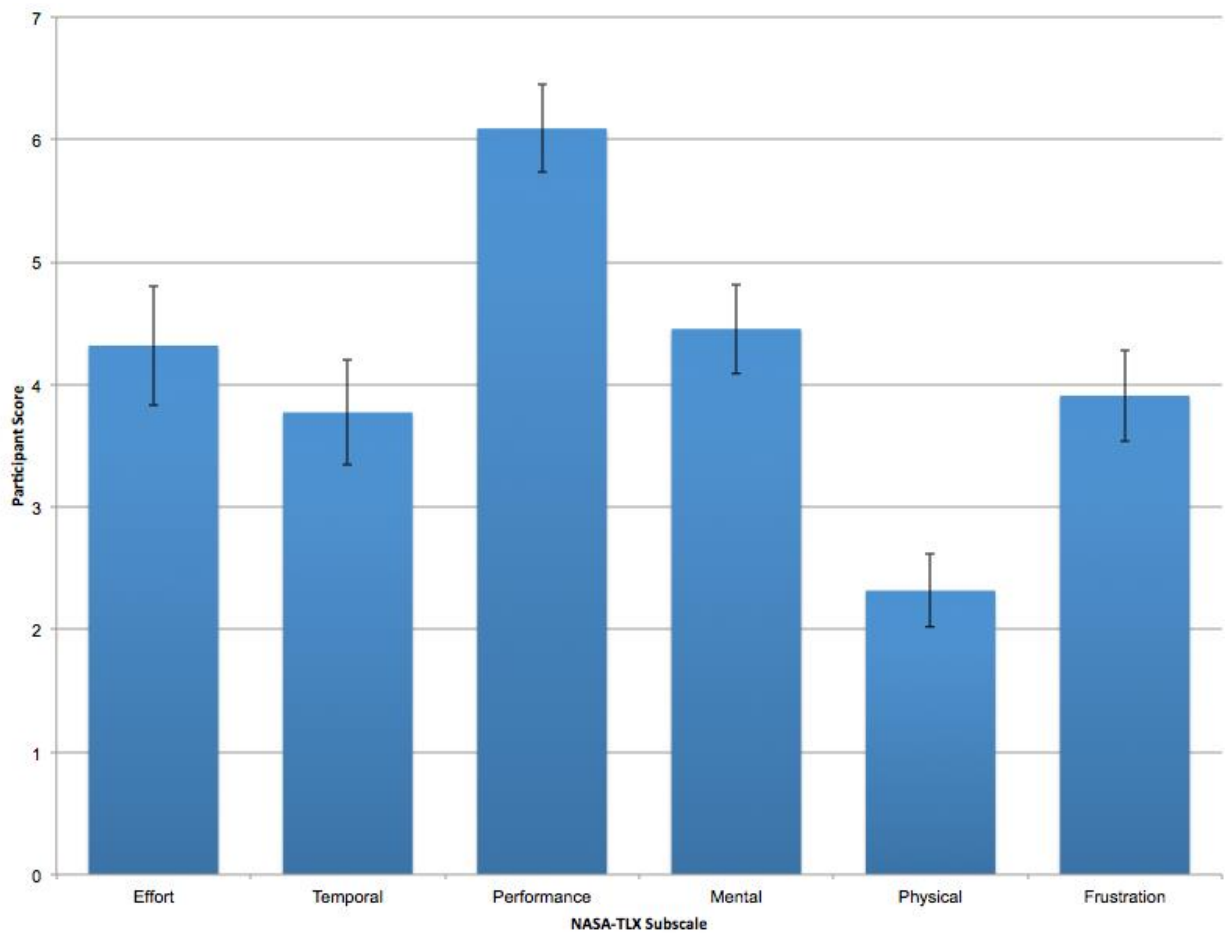


Figure 4.4: NASA-TLX Scores Given by Workshop Participants

Feedback

Participant free text comments echoed some themes that arose in previous experiments but also produced some new food for thought. We grouped them into thematic concerns as per Table 4.2 below.

Table 4.2: Workshop participants' feedback

Theme	Example
More control over the flickering between the pairs of images	<p><i>"A click on the image to change pictures...would be helpful as you can keep your cursor where it is on a location and see the difference between images."</i></p> <p><i>"Maybe I would like more "blinks" between images."</i></p> <p><i>"A slightly longer interval between image changes...would be useful."</i></p>
Uncertainty due to the quality and contrast	<p><i>"Brightness correction option would be great."</i></p> <p><i>"Some images are so bright as to be useless, and others are very thin strips that you can't see features on."</i></p> <p><i>"It would be helpful to have an option to say when the resolution/contrast is too low to use an image or when you simply cannot tell because of movement in the images or the image quality."</i></p> <p><i>"An option to select when image data is too poor to identify change is really important to prevent false negatives."</i></p>
Further training that volunteers	<p><i>"I think more examples for each type of change would make it easier to identify changes...it was quite difficult and therefore became stressful as I wanted to ensure I was finding all the changes and categorising them properly."</i></p> <p><i>"A tutorial may be beneficial to show examples of different surface feature differences."</i></p> <p><i>"There should be a tutorial to show the effect of parallax and shadows on image distortion."</i></p>
Participants proposed functions for the website	<p><i>Would be nice to have some kind of globe in one corner indicating roughly where the image location is on Mars, perhaps as a feature that</i></p>

Theme	Example
	<p><i>could be enabled, keeping the user engaged.”</i></p> <p><i>“It would be beneficial to have list of features ... on the first page of the image (before you select yes/no) so you have a reference of what to look for.”</i></p> <p><i>“Some form of scale to indicate the difference in the sun’s angle incident on the Martian surface and amount of distortion due to parallax would be useful especially in identifying dune migration as otherwise it can be quite deceptive.”</i></p>

5. Conclusions and implications for future work

The findings of the present deliverable are summarized in Table 5.1 below in terms of the questions that were addressed, the findings and their relevance to the development of the project. The work reported here has informed (a) the viability of collection of information about different surface features; (b) the task design behind the interface (what we ask people to do); (c) explored the impacts of working-with-algorithms vs. working-with-crowds; and (d) sought general formative feedback from a group of planetary science expert users with a view to optimising the final public deployment. Aside from this we have explored some fundamental primary science issues including an inconsistency in the scientific literature (single vs. multiple feature detection trade-offs in terms of accuracy vs. vigilance decrement) and explored timely issues of human-algorithm and human-crowd interaction. A particularly striking finding in this area was how rapidly naive participants are able to learn about and assess the relative performance of the “algorithm”.

Table 5.1 Summary of experiments

Study	Question	Summary findings	iMars implications
Detection of single or multiple feature types?	How many feature types should users search for? Trade-off between vigilance decrement (specialisation = fewer targets per session) and sensitivity/specialisation.	1. Roughly equivalent performance overall but... 2. Multiple features leads to improving performance over time, specialisation into declining performance	1. Suggests user engagement better supported by allowing users to detect multiple features 2. Monitor deployed system for hit-rate and miss-rate by feature type (cross-ref section

Study	Question	Summary findings	iMars implications
			2), consider intervening if some turn out to be especially challenging and feed back in as a single-feature tasks.
Optimisation of change detection method	Comparison of manual flicker, automatic flicker and side-by-side with Martian imagery.	<ol style="list-style-type: none"> 1. Automatic flicker quickest and most accurate method 2. Manual flicker and side-by-side comparable in terms of time but manual flicker better performance (~5%) 3. Above pattern holds across all feature types with the exception of impact craters where manual flicker was better 4. Some interaction effects in feature detection difficulty but broadly speaking detection method has more influence than differences between features. 	<ol style="list-style-type: none"> 1. Use automatic flicker as default method. 2. However, TLX and qualitative feedback suggest that some users find this aversive so include manual flicker option if it can be integrated into the same interface. 3. Few arguments for side-by-side especially as it would lead to creation of two separate interfaces in a common project.
Trust in the crowd vs. trust in the machine	<ol style="list-style-type: none"> 1. One pathway to integration with data mining/machine learning techniques is to present participants with images to “check”, how does this impact performance and engagement? 2. Are these effects the same if users know they are checking “crowd” 	<ol style="list-style-type: none"> 1. Trust is related to the performance of both crowd and algorithm 2. Naive participants were able to rapidly (within a single session) judge the relative accuracy of an algorithm/crowd. 3. Participants are more forgiving of algorithms than crowds themselves. 	<ol style="list-style-type: none"> 1. When interleaving human and machine performance care must be given to issues of under and over-trust that may impact and bias overall performance. The overall action to be taken is to be careful to maintain a

Study	Question	Summary findings	iMars implications
	<p>outputs?</p> <p>This speaks to wider issues regarding biasing and weighting effects in task completion.</p>		<p>fairly consistent hit rate and to consider leavening this with either hits or misses (changes present or not present) to avoid complacency in over-trust (“our work here is redundant, the computer is always right”) and under-trust (“the stimuli we are being fed are garbage”).</p>
Workshop/test deployment	<p>1. Deployment of candidate release platform with planetary science experts – general feedback sought in qualitative and quantitative performance form.</p>	<p>1, Control over flicker</p> <p>2. Uncertainty regarding quality and contrast</p> <p>3. Further training of volunteers</p> <p>4. Additional functionality (esp. Sun angle and parallax indicator).</p>	<p>1. TBC, but to be acted on as reasonably practicable.</p>

The next steps this work leads to are primarily in terms of ongoing prototyping of the iMars citizen science project.

- The form of the crucial change detection interface will contain both automatic and manual flicker options as we have established this will lead to higher performance and the manual flicker format (or the option to switch to it) will mitigate concerns over temporal demand and the feeling some participants had that it made the task feel rushed or visually irritating.
- Design the initial version of the prototype so that participants can annotate multiple feature types (in view of the general idea that single feature detection, while easier to explain and initially producing higher performance, sees a relatively fast ‘drop off’ in performance and is at scale a for less efficient use of time). Based on analysis of viability these features are: Impact Crater, Gully, Dune, Slope Streak/RSL, Dust Devil Track and Season Fan using a bounding box. We will however explore further how we can monitor detection of feature types, a first test of which was done in the trial workshop reported here in the event some features become clearly under-detected or where reporting about them shows a high level of conflict. Should this occur,

we will review whether it is necessary to issue a variant of the citizen science task specialised to single feature detection.

- The workshop also led to some formative suggestions to be explored to enhance the experiment. We will evaluate these in terms of their technical viability: control over automatic flicker rate, whether extra training is required within the site in terms of trials with feedback (overall performance using the platform suggests this may not be a great problem but we should still be concerned with this in terms of user experience and satisfaction) and finally, whether it is possible to extract from image metadata information about sun angle and parallax to inform the observer. In the event it is, it will be worthwhile to test whether this actually affects (improves or even impairs) performance. We note that in the absence of this information, participants still appeared to perform well and were able to make allowance for these sources of variability in the image pairs. Indeed, it is this flexibility (and the ability to note one's own act of flexibility) that is the point of involving human input.
- With caveats and potential improvements noted, the workshop also demonstrated that the online system was functional under simultaneous use, the tasks required of participants were at a baseline level 'doable' and generally found to be interesting enough to sustain engagement, and that useful scientific data could be collected and exchanged with Work Package 6 to train algorithms. In view of this, our plan more generally is to increase the number of testers/users in advance of the point of public release.

6. Outputs & Publications

6.1. Publications

Houghton, R.J., Wardlaw, J., Sprinks, J., Giordano, M., Bamford, S., Marsh, S. 2016. Martian Factors: A systems ergonomics approach to citizen science. Human Factors in Complex Systems, Nottingham UK.

Sprinks, J., Wardlaw, J., Houghton, R.J., Bamford, S., Marsh, S. 2016. Mars in Motion: An online Citizen Science platform looking for changes on the surface of Mars, DPS 48/EPSC 11 Meeting, Pasadena, USA

Wardlaw, J., Sprinks, J., Houghton, R.J., Bamford, S., Marsh, S. 2016. Better the Martian you know? Trust in the crowd vs. trust in the machine when using a Martian Citizen Science platform, DPS 48/EPSC 11 Meeting, Pasadena, USA

6.2. Workshops & Demonstrations

Europlanet training workshop: 3D facilities available at the UK NASA RPIF. 7-9th June, 2016, UCL Mullard Space Science Laboratory, Surrey, UK. Demonstration of beta version of 'Mars in Motion' citizen science platform (described in section 7)

7. Relevant links

Citizen Science Alliance website and application, available at: www.citizensciencealliance.org

Panoptes 'Project Builder' interface, available at: <https://www.zooniverse.org/lab> - requires registration with Zooniverse platform

Mars in Motion Citizen Science platform (used for MSSL Workshop) available at: www.zooniverse.org/projects/imarsnottingham/mars-in-motion-rpif-3d-workshop

8. References

Balme, M. & the ISSI team, n.d. Northern plains of Mars: Origins, evolution and response to climate change. <http://www.issibern.ch/teams/plainsofmars/> accessed 17.10.16

Becerra, P., 2014. Transient bright "halos" on the south polar residual cap of mars: Implications for mass balance. LPI Contributions 1791, 1013.

Bexton, W.H., Heron, W., Scott, R.H., 1954. Effects of decreased variation in the sensory environment. Canadian Journal of Psychology, 8, 70–76.

Bourke, M., McGaley-Towle, Z., 2014. Why do sand furrow distributions vary in the North Polar latitudes on Mars? Presented at the EGU General Assembly Conference Abstracts, p. 13626.

Bowyer, A., Lintott, C., Hines, G., Allen, C., Paget, E. 2015. Panoptes: A project building tool for citizen science. Proceedings of the AAAI Conference on Human Computation and Crowd Sourcing (HComp'15), San Diego, CA.

Bridges, N.T., Ayoub, F., Avouac, J-P, Leprince, S., Lucas, A., Mattson, S. 2012. High sand fluxes and abrasion rates on mars determined from HiRISE Images. 43rd Lunar and Planetary Science Conference, 1322.

Broadbent, D.E., Gregory, M., 1965. Effects of noise and of signal rate upon vigilance analysed by means of decision theory. Human Factors, 7, 155–162. doi:10.1177/001872086500700207

Brunetti, M.T., Guzzetti, F., Cardinali, M., Fiorucci, F., Santangelo, M., Mancinelli, P., Komatsu, G., Borselli, L., 2014. Analysis of a new geomorphological inventory of landslides in Valles Marineris, Mars. Earth and Planetary Science Letters 405, 156–168. doi:10.1016/j.epsl.2014.08.025

Byrne, S., Dundas, C.M., Kennedy, M.R., Mellon, M.T., McEwen, A.S., Cull, S.C., Daubar, I.J., Shean, D.E., Seelos, K.D., Murchie, S.L., Cantor, B.A., Arvidson, R.E., Edgett, K.S., Reufer, A., Thomas, N., Harrison, T.N., Posiolova, L.V., Seelos, F.P., 2009. Distribution of Mid-Latitude Ground Ice on Mars from New Impact Craters. Science 325, 1674–1676. doi:10.1126/science.1175307

Chojnacki, M., Burr, D.M., Moersch, J.E., 2014. Valles Marineris dune fields as compared with other martian populations: Diversity of dune compositions, morphologies, and thermophysical properties. *Icarus*, Third Planetary Dunes Systems 230, 96–142. doi:10.1016/j.icarus.2013.08.018

Chojnacki, M., Johnson, J.R., Moersch, J.E., Fenton, L.K., Michaels, T.I., Bell III, J.F., n.d. Persistent aeolian activity at Endeavour crater, Meridiani Planum, Mars; new observations from orbit and the surface. *Icarus*. doi:10.1016/j.icarus.2014.04.044

Chojnacki, M., McEwen, A., Dundas, C., Mattson, S., Ojha, L., Byrne, S., Wray, J., 2014. Geologic Context of Recurring Slope Lineae in Coprates Chasma. Presented at the Lunar and Planetary Science Conference, p. 2701.

Conci, M., Müller, H.J., 2014. Global scene layout modulates contextual learning in change detection. *Frontiers in Psychology* 5. doi:10.3389/fpsyg.2014.00089

Conway, S., Balme, M., Murray, J., Towner, M., 2014. Comparing the topographic long profiles of gullies on Earth and Mars. Presented at the EGU General Assembly Conference Abstracts, p. 15122.

Crosta, G.B., Utili, S., De Blasio, F.V., Castellanza, R., 2014. Reassessing rock mass properties and slope instability triggering conditions in Valles Marineris, Mars. *Earth and Planetary Science Letters* 388, 329–342. doi:10.1016/j.epsl.2013.11.053

Daubar, I.J., Atwood-Stone, C., Byrne, S., McEwen, A.S., Russell, P.S., 2014. The morphology of small fresh craters on Mars and the Moon. *Journal of Geophysical Research: Planets*. JE004671. doi:10.1002/2014JE004671

Daubar, I.J., McEwen, A.S., Byrne, S., Kennedy, M.R., Ivanov, B., 2013. The current martian cratering rate. *Icarus* 225, 506–516.

de Vet, S.J., Merrison, J.P., Mittelmeijer-Hazeleger, M.C., van Loon, E.E., Cammeraat, L.H., 2014. Effects of rolling on wind-induced detachment thresholds of volcanic glass on Mars. *Planetary and Space Science* 103, 205–218. doi:10.1016/j.pss.2014.07.012

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, 012015. doi:10.1088/1755-1315/17/1/012015

Dietworst, B.J., Simmons, J.P. & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114-126.

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, CIKM '10*. ACM, New York, NY, USA, pp. 749–758. doi:10.1145/1871437.1871534

Doring, C., Eichhorn, A., Wang, X., Kruse, R., 2006. Improved Classification of Surface Defects for Quality Control of Car Body Panels, in: 2006 IEEE International Conference on Fuzzy Systems. Presented at the 2006 IEEE International Conference on Fuzzy Systems, pp. 1476–1481. doi:10.1109/FUZZY.2006.1681903

Drury, C.G., Addison, J.L., 1973. An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance. *Ergonomics* 16, 159–169. doi:10.1080/00140137308924492

Dundas, C.M., Byrne, S., McEwen, A.S., Mellon, M.T., Kennedy, M.R., Daubar, I.J., Saper, L., 2014. HiRISE observations of new impact craters exposing Martian ground ice. *Journal of Geophysical Research: Planets* 119, 2013JE004482. doi:10.1002/2013JE004482

Dundas, C.M., Diniega, S., McEwen, A.S., n.d. Long-term monitoring of martian gully formation and evolution with MRO/HiRISE. *Icarus*. doi:10.1016/j.icarus.2014.05.013

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

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* 42, 1035–1039. doi:10.1130/G35798.1

Fenton, L.K., Michaels, T.I., Chojnacki, M., Beyer, R.A., 2014. Inverse maximum gross bedform-normal transport 2: Application to a dune field in Ganges Chasma, Mars and comparison with HiRISE repeat imagery and MRAMS. *Icarus, Third Planetary Dunes Systems* 230, 47–63. doi:10.1016/j.icarus.2013.07.009

Goodman, J., 2014. The Wages of Sinistrality: Handedness, brain structure, and human capital accumulation. *Journal of Economic Perspectives* 28, 193–212. doi:10.1257/jep.28.4.193

Grimm, R.E., Harrison, K.P., Stillman, D.E., 2014. Water budgets of martian recurring slope lineae. *Icarus* 233, 316–327. doi:10.1016/j.icarus.2013.11.013

Harris, D.B., Chaney, F.B., 1969. *Human Factors in Quality Assurance*. John Wiley & Sons.

Harris, D.H., 1968. Effect of defect rate on inspection accuracy. *Journal of Applied Psychology* 52, 377–379. doi:10.1037/h0026241

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

Hart, G. (2006). NASA Task Load Index (NASA-TLX): 20 Years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(19), 904–908.

Hayden, D.S., Chien, S., Thompson, D.R., Castaño, R., 2012. Using Clustering and Metric Learning to Improve Science Return of Remote Sensed Imagery. *ACM Transactions on Intelligent System Technology*. 3, 51:1–51:19. doi:10.1145/2168752.2168765

Hayward, R.K., Fenton, L.K., Titus, T.N., 2014. Mars Global Digital Dune Database (MGD3): Global dune distribution and wind pattern observations. *Icarus, Third Planetary Dunes Systems* 230, 38–46. doi:10.1016/j.icarus.2013.04.011

Hobbs, S.W., Paull, D.J., Clarke, J.D.A., 2014. A comparison of semiarid and subhumid terrestrial gullies with gullies on Mars: Implications for Martian gully erosion. *Geomorphology* 204, 344–365. doi:10.1016/j.geomorph.2013.08.018

Johnsson, A., Reiss, D., Hauber, E., Hiesinger, H., Zanetti, M., 2014. Evidence for very recent melt-water and debris flow activity in gullies in a young mid-latitude crater on Mars. *Icarus* 235, 37–54. doi:10.1016/j.icarus.2014.03.005

Johnson, M.B., Zimbelman, J.R. 2014. Documentation of sand ripple patterns and recent surface winds on martian Dunes. 45th Lunar and Planetary Science Conference, 1518.

Kristofferson, M.W., Groen, M., Kristofferson, A.B., 1973. When visual search functions look like item recognition functions. *Perception & Psychophysics* 14, 186–192. doi:10.3758/BF03198632

Lee, J., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 1, 50-80.

Lorenz, R.D., 2013. Dust devil populations : Comparing in-situ measurements with imaging and tracks. *Mars Atmosphere: Modelling and observation*, 5th international workshop, 1, 1406.

Lorenz, R.D., Reiss, D., 2015. Solar panel clearing events, dust devil tracks, and in-situ vortex detections on Mars. *Icarus* 248, 162–164. doi:10.1016/j.icarus.2014.10.034

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

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

Marshall, P.J., Verma, A., More, A., Davis, C.P., More, S., Kapadia, A., Parrish, M., Snyder, C., Wilcox, J., Baeten, E., Macmillan, C., Cornen, C., Baumer, M., Simpson, E., Lintott, C.J., Miller, D., Paget, E., Simpson, R., Smith, A.M., Küng, R., Saha, P., Collett, T.E., Tecza, M., 2015b. Space Warps: I. Crowd-sourcing the Discovery of Gravitational Lenses. *Monthly Notices of the Royal Astronomical Society* 455, 1171–1190. doi:10.1093/mnras/stv2009

Mattson, S., McEwen, A., Kirk, R., Howington-Kraus, E., Chojnacki, M., Runyon, K., Cremonese, G., Re, C., 2014. Martian Landscapes in Motion. *EGU General Assembly Conference Abstracts*, p. 10153.

Mattson, S., Kilgallon, A., Byrne, S., McEwen, A.S., Herkenhoff, K., Okubo, C., Putzig, N.E., Russel, P. 2014. Meter-scale pits in Mars' north polar layered deposits. 45th Lunar and Planetary Society Conference, 2431.

McEwen, A., Byrne, S., Chevrier, V., Chojnacki, M., Dundas, C., Masse, M., Mattson, S., Ojha, L., Pommerol, A., Toigo, A., Wray, J., 2014. Recurring Slope Lineae and Future Exploration of Mars. Presented at the EGU General Assembly Conference Abstracts, p. 8851.

McEwen, A.S., Ojha, L., Dundas, C.M., Mattson, S.S., Byrne, S., Wray, J.J., Cull, S.C., Murchie, S.L., Thomas, N., Gulick, V.C., 2011. Seasonal Flows on Warm Martian Slopes. *Science* 333, 740–743. doi:10.1126/science.1204816

Megaw, E.D., Alexander, C.J., Richardson, J., 1979. Fault mix and inspection performance. *International Journal of Production Research* 17, 181–191. doi:10.1080/00207547908919606

Meuter, R.F.I., Lacherez, P.F., 2016. When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening. *Human Factors* 58, 218–228. doi:10.1177/0018720815616306

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 81–97. doi:10.1037/h0043158

More, A., Verma, A., Marshall, P.J., More, S., Baeten, E., Wilcox, J., Macmillan, C., Cornen, C., Kapadia, A., Parrish, M., Snyder, C., Davis, C.P., Gavazzi, R., Lintott, C.J., Simpson, R., Miller, D., Smith, A.M., Paget, E., Saha, P., Küng, R., Collett, T.E., 2015. Space Warps II. New Gravitational Lens Candidates from the CFHTLS Discovered through Citizen Science. *Monthly Notices of the Royal Astronomical Society* 455, 1191–1210. doi:10.1093/mnras/stv1965

Neisser, U., Novick, R., Lazar, R., 1963. Searching for ten targets simultaneously. *Perceptual & Motor Skills* 17, 955–961. doi:10.2466/pms.1963.17.3.955

Ojha, L., McEwen, A., Dundas, C., Byrne, S., Mattson, S., Wray, J., Masse, M., Schaefer, E., 2014. HiRISE observations of Recurring Slope Lineae (RSL) during southern summer on Mars. *Icarus* 231, 365–376. doi:10.1016/j.icarus.2013.12.021

Pashler, H., 1987. Familiarity and visual change detection. *Perception & Psychophysics* 44, 369–378. doi:10.3758/BF03210419

Peirce, J.W., 2008. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2, 10. doi:10.3389/neuro.11.010.2008

Peirce, J.W., 2007. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods* 162, 8–13. doi:10.1016/j.jneumeth.2006.11.017

Raack, J., Reiss, D., Appéré, T., Vincendon, M., Ruesch, O., Hiesinger, H., 2014. Present-day seasonal gully activity in a south polar pit (Sisyphi Cavi) on Mars. *Icarus*. doi:10.1016/j.icarus.2014.03.040

- Rao, P., Bowling, S.R., Khasawneh, M.T., Gramopadhye, A.K., Melloy, B.J., 2006. Impact of training standard complexity on inspection performance. *Human Factors in Manufacturing* 16, 109–132. doi:10.1002/hfm.20045
- Reiss, D., Hoekzema, N.M., Stenzel, O.J., 2014a. Dust deflation by dust devils on Mars derived from optical depth measurements using the shadow method in HiRISE images. *Planetary and Space Science* 93–94, 54–64. doi:10.1016/j.pss.2014.01.016
- Reiss, D., Spiga, A., Erkeling, G., 2014b. The horizontal motion of dust devils on Mars derived from CRISM and CTX/HiRISE observations. *Icarus* 227, 8–20. doi:10.1016/j.icarus.2013.08.028
- Rensink, R.A., O'Regan, J.K., Clark, J.J., 1997. To See or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8, 368–373. doi:10.1111/j.1467-9280.1997.tb00427.x
- Russell, P.S., Byrne, S., Dawson, L.C. 2004. Active powder avalanches on the steep north polar scarps of Mars – 4 years of HiRISE observation. 45th Lunar and Planetary Science Conference, 2688.
- Sawyer, B.D., Finomore, V.S., Funke, G.J., Mancuso, V.F., Funke, M.E., Matthews, G., Warm, J.S., 2014. Cyber vigilance effects of signal probability and event rate. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1771–1775. doi:10.1177/1541931214581369.
- S Mattson, A.K., Kilgallon, A., Byrne, S., McEwan, A.S., Herkenhoff, K. Okubo, C., Putzig, N.E., Russell, P. 2014. Meter-scale pits in Mars' north polar layered deposits. 45th Lunar and Planetary Science Conference, 2431.
- 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, Image Processing and Analysis in Space Science* 53, 1822–1833. doi:10.1016/j.asr.2013.05.012
- Stillman, D.E., Michaels, T.I., Grimm, R.E., Harrison, K.P., 2014. New observations of martian southern mid-latitude recurring slope lineae (RSL) imply formation by freshwater subsurface flows. *Icarus* 233, 328–341. doi:10.1016/j.icarus.2014.01.017
- Su, J.-Y., Konz, S., 1981. Evaluation of Three Methods for Inspection of Multiple Defects/Item. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 25, 627–630. doi:10.1177/1071181381025001164
- Warm, J.S., Dember, W.N., Hancock, P.A., 1996. Vigilance and workload in automated systems. ResearchGate.
- Weiss, D.K., Head, J.W., 2014. Ejecta mobility of layered ejecta craters on Mars: Assessing the influence of snow and ice deposits. *Icarus* 233, 131–146. doi:10.1016/j.icarus.2014.01.038
- Wickens, T.D., 2001. *Elementary Signal Detection Theory*. Oxford University Press, Cary, NC, USA.
- Wiener, E.L., Curry, R.E., Faustina, M.L., 1984. Vigilance and task load: In search of the inverted U. *Human Factors*, 26, 215–222. doi:10.1177/001872088402600208

Williams, J.-P., Pathare, A.V., Aharonson, O., 2014. The production of small primary craters on Mars and the Moon. *Icarus* 235, 23–36. doi:10.1016/j.icarus.2014.03.011