Cartoscope: Designing Effective Interfaces for Motivating Engagement in Crowdsourced Image Labeling

Sofia Eleni Spatharioti

Khoury College of Computer Sciences, Northeastern University
spatharioti.s@husky.neu.edu

Abstract

Crowdsourcing is increasingly becoming a viable solution for rapidly analyzing large amounts of data that require human judgment. Within this area, image labeling tasks are gaining wide adoption across different disciplines, as a popular and cost-effective solution for performing image analysis, with applications ranging from Machine Learning training and citizen science, to disaster response and monitoring, and environmental justice. However, the repetitive and tedious nature of such tasks poses a significant challenge for project creators to retain participants, as they become disengaged early in the process. This can not only lead to less diverse outputs, due to small groups doing most of the work, but also poses a threat for smaller projects to gain traction and collect enough data, or succeed in their awareness and volunteer recruitment efforts.

This thesis will tackle crowd disengagement in image labeling tasks along three main directions. More specifically, the first part of the thesis will focus on designing more effective traditional, online image labeling interfaces, by employing different framing elements, such as task variety, as well as context information and interactive Points of Interest (POI) identification. The second part will explore injecting game mechanics to create more interactive image labeling interfaces, and how can these impact participant engagement and performance. These mechanics will center around building an image matching online game, and will also include elements such as task variety. Finally, the third part of the thesis will examine how introducing co-location in image labeling, via a multi-person tabletop image labeling game toolkit, can lead to deeper engagement and discussions from participants. Within this co-located setting, the interplay of collaboration and competition will be explored in order to identify effective mechanics and rulesets that can achieve high engagement levels across participants with different backgrounds.

These research directions will be addressed by building an open-source crowdsourcing framework called Cartoscope. The framework is comprised of three different image labeling tools that drive this thesis: 1) a classic image labeling interface called Cartoscope Classic, 2) a web image matching game called Tile-o-Scope Grid and 3) an Augmented Reality (AR) tabletop image labeling toolkit called Tile-o-Scope AR. These implementations will be evaluated via user studies, as well as running experiments on crowdsourcing marketplaces such as Amazon Mechanical Turk.

1 Introduction

When faced with a large amount of data that would be either computationally challenging or rely on human subjectivity to process, crowdsourcing is a popular approach to gathering information on solutions provided by humans. In particular, image analysis tasks—such as image labeling or rating—have proven to be particularly amenable to this approach. A number of projects over the past decade have taken crowdsourced approaches to acquiring labels for images, either for the sake of acquiring the labels themselves or as training data for machine learning techniques [48, 58, 79]. Within this area, crowdsourced image rating for mapping—especially in the context of disaster response—has recently arisen. Projects such as SandyMill[1] (a collaboration between the Humanitarian OpenStreetMap Team and FEMA) [50], Tomnod's involvement in GEO-CAN [4], the Ushahidi-Haiti Project [38], and Cropland Capture [70] have all taken approaches to asking crowd workers to rate images for the purposes of creating, improving, or annotating maps. "Space archaeology" is another emerging area for crowdsourcing image analysis, with projects such as Global Xplorer, a platform for analyzing satellite images [86].

Crowdsourcing holds promise for applying human processing to the massive amounts of data that can be generated in the wake of disasters—either through paid work or volunteering. However, analysis of contribution patterns of participants in volunteer crowdsourcing projects generally indicate that most participants contribute little work, and that the bulk of the work is done by a small number of participants [61, 70]. This leads to relying on a few participants to disproportionately carry out the majority of the work, makes it less likely that projects will find such participants and be successful, and limits the amount of data that can be processed. Additionally, Munro et al. [50] propose paid crowdsourcing as a cost-effective alternative when recruiting volunteers is not feasible; this may arise for less "prominent" disasters, which affect smaller areas, receive less media coverage and thus may attract significantly fewer volunteers. Such a comparison can be seen between two events in 2010: the Haiti earthquake, which received more than 3,000 news stories within the first 10 days, and the Pakistan floods, which received 320 broadcast news stories and 730 print news stories in the same timeframe (Brookings-Bern Project on Internal Displacement [6]).

In this thesis, I propose to tackle disengagement in crowdsourced image labeling systems, along three main directions: *traditional*, *gamified* and *co-located* image labeling. More specifically, the first part of this thesis will center around **Cartoscope Classic [11]**, an online image labeling interface with minimal game elements. Cartoscope Classic was in part inspired by Public Lab's SandyMill project [50]. The second part of the thesis will introduce a more gamified approach towards crowdsourced image labeling, driven by the development of **Tile-o-Scope Grid [76]**, an image matching web game, where images are placed on a grid and, similar to the game Dots [54], players are tasked with collecting images by drawing lines to connect images of the same category.

Both Cartoscope Classic and Tile-o-Scope Grid will focus on designing effective mechanisms that utilize task variety in order to boost participant retention and performance. Variety has long been recognized as a key factor in creating motivating work. Hackman and Oldham [19] designed and tested a Job Characteristics Model, aiming to increase internal motivation and performance of employees through effective job design. They identified five core job characteristics, which include skill variety, along with task identity, task significance, autonomy, and job feedback. Lunenburg [41] followed that work, providing an overview of empirical studies on the Job Characteristics Model, as well as applications of the model in the field of management. Other research has examined the tradeoffs between variety and specialization in work. Staats and Gino [69] found that when examining the repetitive tasks of bank workers, specialization improved productivity over short periods of time (i.e., a day), while over longer periods, variety did. Narayanan et al. [51] found that for software maintenance, a balance between specialization and variety led to the highest productivity.

The third part of this thesis will focus on enhancing the image labeling process by incorporating co-located discussion elements, in tandem with collaboration and competition techniques. This is motivated in part in research conducted in the field of citizen science. Social components of projects in this area have been have been shown to significantly correlate with improvement in scientific literacy [56]. Recent work has also shown that feelings of satisfaction of contributing to projects and engagement are linked to social outcomes, such as a broader sense of community [20, 36], as well as increased trust and reconnecting people with each other [53]. Allowing users to discuss data while also performing analysis opens up a new space for social

[1] SandyMill forked an existing open source image sorting application, MapMill, developed by Public Lab and Jeff Warren. In SandyMill, crowd members sorted images taken by Civil Air Patrol of damage from Hurricane Sandy.

conversation, through what Wylie and Albright have described as "Civic Technoscience" [84]. The benefits of engaging volunteers across disciplines to contribute to citizen science has been the emphasis of a wide body of work [8, 17, 24, 63]. This direction will be explored using **Tile-o-Scope AR [75]**, an Augmented Reality (AR) tabletop game toolkit for image labeling.

This thesis will address increasing engagement and performance in crowdsourcing, volunteer motivated, image labeling platforms, by tackling the following research questions:

- *RQ1: How can traditional interfaces be enhanced to motivate engagement in crowdsourced image labeling?*
  - RQ1.1: What different motivations do participants provide for their contributions in traditional image labeling tasks?
  - RQ1.2: How can introducing fixed task variety scheduling schemes in a traditional online image labeling interface positively impact participant engagement ?
  - RQ1.3: What impact, if any, differing framing techniques may have on participant performance, in terms of levels of engagement and output quality?
- *RQ2: How can image matching game mechanics in online image labeling lead to more engaging participant experiences?*
  - RQ2.1: How can variable task variety scheduling schemes positively impact engagement in crowdsourced image labeling?
  - RQ2.2: Can adaptive task variety mechanisms boost participant retention in crowdsourced image labeling?
- *RQ3: Can the interplay between collaboration and competition in a co-locating setting be utilized to engage participants in crowdsourced image labeling?*

While there exists a wide array of applications for image labeling that utilize digital interfaces, we chose to pursue an approach that uses Augmented Reality (AR) for the third direction of this thesis. This was motivated by a recent body of work that has explored the benefits of using AR technology over other digital or non-digital alternatives. In education, AR applications have been shown to be more effective than non-AR applications both in theory [5, 7] and through empirical research [21, 57, 59]. For example, in a study of medical training, students who used AR in their training were better able to transfer their knowledge to real world situations than students who used Virtual Reality (VR) in their training [57]. Additionally, the concepts learned through AR have been shown to be memorized better both in short and long term memory [43, 44, 77, 78]. Most importantly, AR was proven to be more effective on collaboration than other digital or non-digital media [26, 37, 49, 59, 81]. Despite the usability issues that have been associated with some AR research applications, they have been found to be more satisfying [27], fun, and interesting to be replayed, compared to other forms of digital media [25].

A summary of the plan of this thesis is presented in Figure 1, showcasing the three branches based on the research questions, as well as completed and suggested future work. The remainder of this section provides an overview of the three tools, and how each one will be used to address the aforementioned research questions.

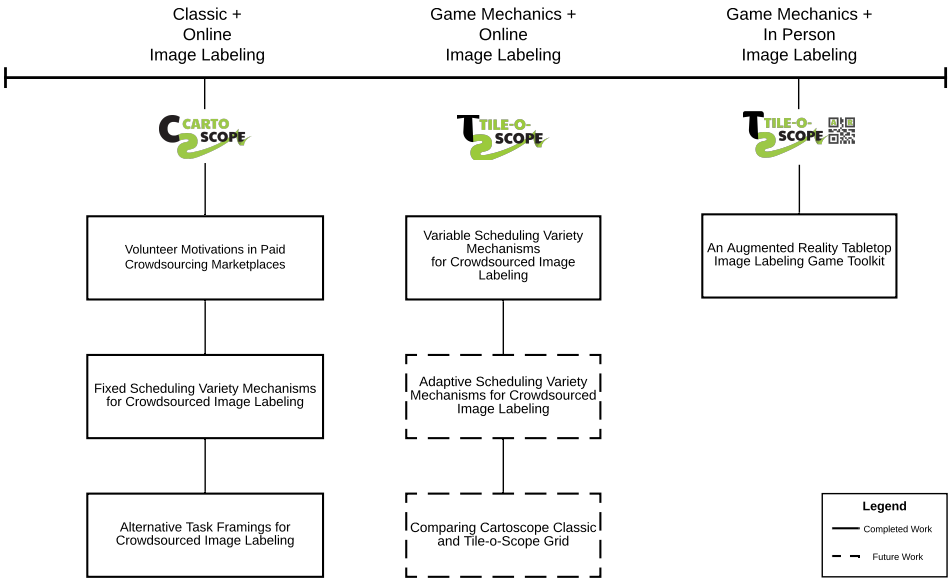
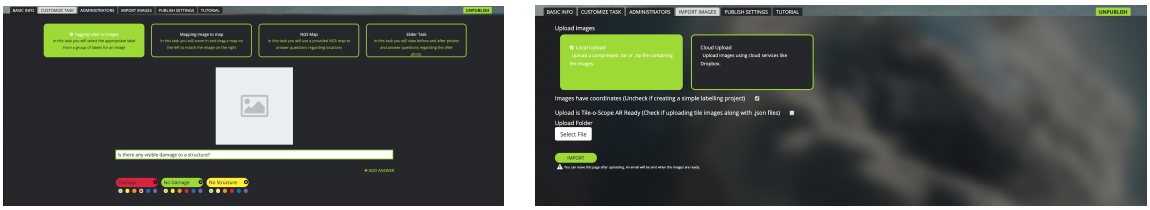


Figure 1: Thesis work plan.

1.1 Interfaces



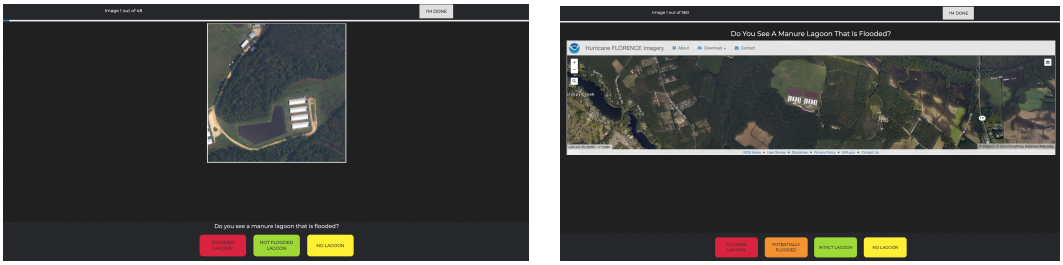
Subtask Selection *Dataset Upload*

Figure 2: Screenshots from the project creation workflow, showing the task customization and dataset upload tabs.

Cartoscope Classic serves as the core hub of this thesis. Smaller organizations will be able to utilize the platform's simple project creation workflow to design projects with their own data, and get real-time visualization and aggregation results. Furthermore, Cartoscope Classic will also be used to create playable sets for the other two interfaces of this research project, Tile-o-Scope Grid and Tile-o-Scope AR. An example of the project creation flow can be seen in Figure 2. Cartoscope Classic currently supports the following types of tasks:

- *Labeling*: Label images by answering questions using provided options. An example can be seen in Figure 3.
- *Mapping*: Perform actions on a map on the left and answer questions about the image on the right.
- *Markers*: Label mapped Points of Interest by coloring their markers using provided options.
- *NGS*: Answer questions about Points of Interest using mapping tools provided by NOAA's National Geodetic Survey (NGS). An example can be seen in Figure 3.
- *Slider*: Label images by comparing before and after versions of the same area.

Cartoscope Classic is the main focus of the first reserach question of this thesis. It will first be used to explore motivations in paid crowdsourcing platforms such as Amazon Mechanical Turk, to validate the ability to use such marketplaces for volunteer recruitment (*RQ1.1*). The plethora of task types supported by Cartoscope will be utilized for exploring different mechanisms of task variety for increasing engagement in crowdsourced image labeling (*RQ1.2*). The *Labeling* and *Markers* tasks will be further utilized when exploring possible impacts of introducing contextual geo-location information to participants during the task ((*RQ1.3*). Finally, Cartoscope Classic will be used in comparison with Tile-o-Scope Grid, to identify which use cases can benefit most from using game mechanics in image labeling tasks (*RQ2).*



Tagging *NGS*

Figure 3: Example of the *tagging* and *NGS* subtask types.

Tile-o-Scope Grid was developed to further explore the impact of task variety in image labeling, as it allows for even further task sequence optimization and adaptation, while also combining gamification elements. In Tile-o-Scope Grid, tiles of images are placed on a 2D grid. The purpose of the game is to connect tiles that contain images of the same category by dragging a non-intersecting line connecting neighboring images in order to collect them. Diagonal lines are also allowed. Every level requires a specific amount of tiles to be collected of each category. If a line is valid and matches images of the same category (which can be checked for some images, which have ground truth categories associated with them), then the move is considered correct and the amount to be collected of the category is reduced by the number of tiles in the match. If the line contains images that do not belong to a unique category, then the move is incorrect and players are penalized, by adding items to their collection counts. Once the player has collected (at least) the required amount for all categories, the level is complete and the player can continue to the next. Players can also shuffle the board, in the event that no matches can be accomplished. Visual and audio feedback is provided for both correct and incorrect moves, as well as level completion. An example of the interface can be seen in Figure 4.

Figure 4: Example of the Tile-o-Scope Grid interface, using a dataset about assessing bridge damage after the 2013 Colorado Floods.

Tile-o-Scope Grid's mechanics allow for further customization of the subtask types, by customizing specific aspects of each level, as well as plugging different algorithms for fostering engagement with the tasks. To

this end, this interface will be used to design, deploy and evaluate task variety mechanisms, such as dynamic and adaptive scheduling scheme mechanisms, to tackle our second research question (*RQ2.1*,*RQ2.2*). Tile-o-Scope Grid will also be used in comparison with Cartoscope Classic (*RQ2)*.

Tile-o-Scope AR is a multi-person, co-located image labeling toolkit that can be played using physical tags and a mobile device such as a phone or tablet. The game utilizes augmented reality (AR) technology to read physical tags and show images on the device in their place, which can then be sorted, by pointing the device over the physical tags to reveal the corresponding images on the screen, and grouping similar images into categories.

Tile-o-Scope AR does not enforce a specific set of game rules, but is simply defined by the mechanics of matching images to a category. This offers the potential for bottom-up citizen science game design. While top-down game design refers to a designer making a game for others to play, by providing players with goals and rules that determine gameplay, bottom-up game design provides players with the mechanisms and components, which can allow players determine their own game goals and rules. Bottom-up game design can allow groups to come up with their own rules and themes, depending on a number of factors, such as number of players, skills, purpose and scientific context and application. Our toolkit is targeted towards smaller scale community groups, that can utilize the open-ended nature of the design to play a variety of games. These can range from adaptations of known games, such as Memory, to completely new games that they can craft themselves, tailored to the problem they are trying to solve, using their own or local data.

While the previous two interfaces focus primarily on task variety mechanisms for image labeling, thus closely tied to the first and second research questions of this thesis, Tile-o-Scope AR will be used first and foremost for exploring the interplay between collaboration and competition in co-located image labeling (*RQ3*).

2 Related Work

2.1 Image Labeling Games

One of the most widely successful collaborative game designs for image labeling is the ESP game [79]. Players must provide the same label to an image in order to score points and proceed. This offers a stronger guarantee for the accuracy of the category. A game similar to the ESP game is ARTigo [82], that focuses on providing labels for various pieces of art. Like in the ESP game, two players are presented with the same image and must collaborate to propose the same label in order to be awarded points. ARTigo aims to build a crowdsourced artwork search engine. While the previous two games were designed with the purpose of providing labels to images, BeFaced [73] aims to generate a crowdsourced dataset of facial expressions for training machine learning models. Utilizing mechanics similar to Bejeweled [2], players must connect facial expressions in the game and then perform the expression they are matching to complete the action.

Snapshot Safari [71], offered through Zooniverse, invites participants to contribute to animal preservation efforts across Africa, by labeling images captured by camera traps . Animal classification is also the focus of Forgotten Island [55]. Designed as a story driven adventure game, Forgotten Island enlists players to classify various types of animals, such as insects, in order to progress the story. Similarly related to nature is Cropland Capture [70], where the crowd is tasked with monitoring cropland by labeling aerial photography, with a goal of generating global cropland maps .

Several citizen science image labeling games have been developed in the domain of health. Stall Catchers aims to serve as a detection system for Alzheimer's. Players annotate images from brain scans for the presence of stalls - clogged blood vessels in the brain that have been linked with the disease. MalariaSpot [40] asks the crowd to identify parasites in images from blood smears, based on distinct characteristics, which are linked to Malaria. Finally, Project Discovery [35] enlisted EVE Online players to look at high-resolution images of human cells and categorize protein patterns.

2.2 Utilizing Task Variety for Increasing Engagement

In the crowdsourcing domain, Kittur et al. [29], in their framework for future crowd work, proposed job design to support both organizational performance, but also worker satisfaction. Feedback from workers suggests that motivation is negatively impacted by monotonous tasks. Recent work has demonstrated that

the ordering and continuity of crowdsourced microtasks can impact worker performance and engagement (Lasecki et al. [32]; [9]).

Lasecki et al. [33] examined the effects of contextual interruptions on crowdsourced microtasks (i.e. subtasks), where workers would switch between tasks of landmark locating on a map and image labeling. Their findings showed that switching context between closely related subtasks could slow workers down, but did not examine effects on workers' engagement. Dai et al. [15] found that inserting short "micro-diversions" (such as pages of a graphic novel) into subtask sequences could improve worker engagement. However, these diversions did not result in work being completed. We see our work as a kind of combination of these two approaches: looking towards improving worker engagement through "diversions" that are different types of subtasks. Our work further explores the tradeoffs of various rates of interleaving different types of subtasks.

While we are primarily interested in worker engagement through variety, a wide body of approaches to improving crowdsourcing have been explored, including game mechanics [70, 79], understanding payment schemes [1, 18, 47] and optimizing microtask workflows to reduce the amount of work needed to be done ([85]; [34]; Dai et al. [14]).

One of the interfaces that will be used in this thesis to explore the impact of variety in engagement involves a tile-matching game, called Tile-o-Scope Grid. Image labeling games have long been used for achieving high quality labels and identifying objects in images [79, 80]. Notable examples of citizen science image labeling interfaces include Cropland Capture [70] and Snapshot Safari [71]. A tile-based game closely related to Tile-o-Scope Grid is Befaced [72], which aims to create a crowdsourced facial expression database. Players are tasked with making facial expressions that match the aligned tiles in order to successfully collect them. BeFaced deploys a Dynamic Difficulty Adjustment [22] algorithm to lower certain matching difficulties caused by certain facial expressions, as a means of retaining player engagement.

The high volume of data available makes games highly suitable for deploying RL algorithms. Mandel et al. explore various comparisons of RL algorithms using engagement in an educational game for evaluation of policy performance [46]. RL approaches were further deployed to combat player disengagement in Refraction, an educational game about fractions [45]. Q-learning algorithms were utilized in Q-DeckRec, a recommendation system for Collectible Card Games (CCGs) [13]. Q-DeckRec can be used in CCGs such as Hearthstone to suggest optimal deck builds that may lead, among other things, to increased player engagement. As the state space in CCGs is often too big for maintaining the lookup table required for Q-learning, a Multi-Layer Perceptron (MLP) approach is employed.

2.3 Using Augmented Reality to Foster Collaboration

Museums have been increasingly embracing the use of Augmented Reality as a new means of increasing engagement, interactivity and collaboration in their exhibits. Results from an evaluation in the Tech Museum of Innovation indicate that tangible tokens provide additional opportunities for collaborative problem solving and impact learning through support for tinkering and experimentation [39]. A similar study at the Exploratorium Museum found that using Tangible User Interfaces could encourage group use and foster manipulation [42]. Tile-o-Scope AR also makes use of tangible tokens, in the form of physical tags.

A work closely related to ours is Synflo [52], which aims to create more engaging museum exhibits with a focus on the domain of biology. The physical component in Synflo is Sifteo Cubes, which are interactive both as singular units and when combined together to achieve specific actions. Evaluation of the system revealed peer collaboration patterns among participants. However, SynFlo's design was limited to applications in biology and may not be as easily customizable to serve different citizen science domains. Sifteo cubes create a variety of constraints, such as the requirement of a computer in range and for changing and updating games, short battery life and a considerable investment due to their price. Tile-o-Scope AR is not constrained by any specific hardware requirement beyond a mobile device. Users can change the current dataset being sorted through the main menu, without requiring any additional change on the physical tags. Tile-o-Scope AR does not require any initial investment, as, in its simplest form, can be played simply through the mobile application and prints of the tags.

Another tool that is focused on enhancing museum visitors' experiences in the domain of biology is BakPack [39]. BakPack utilizes BioBricks as the physical component and a multitouch tabletop interface as the digital component. Similar to Tile-o-Scope AR, BakPack aims to increase collaboration by creating tangible experiences that provoke discussions and learning opportunities. However, the technology requirements of

this design make it harder to be adopted by small groups, as both the physical and digital components are not portable.

Citizen Science aims, among other goals, to educate volunteers, while they are contributing to the projects. Hence, it is important for a citizen science toolkit to be able to accommodate educational objectives. Tile-o-scope AR is designed in a way that can be customized to adapt to a variety of settings and purposes including education. Literature review of augmented reality games in education conducted by Koutromanos et al. illustrate evidence of positive outcome in student learning [30] . One notable example is Reliving the Revolution (RtR), which aims at teaching students essential skills such as problem solving, teamwork and civic engagement, by engaging them with a historical event, the Battle of Lexington, in Massachusetts [62]. The tool's core mechanism is based on GPS coordinates, which are used to present relevant information from the area to players, inviting them to investigate evidence and make decisions, based on pre-assigned roles. Although both RtR and Tile-o-Scope AR pose engaging new tools that can be used for education through building problem solving and teamwork skills, the geo-location component for these tools is used differently. Datasets can be geo-located in Tile-o-Scope AR, but there is no requirement for players being physically present to a location. RtR also does not require additional physical components, whereas Tile-o-Scope AR combines mobile devices with physical tags.

An approach that offers similar portability features to Tile-o-Scope AR for education purposes is Pengu [3]. Pengu relies on more affordable tools than tools like Synflo and BakPack, as it can be played using a mobile application and a standard molecular model kit. Players can then use the kit to construct possible molecules, which can then be scanned using the mobile application for validation. While Pengu does not have any location constraints, as seen in some of the tools mentioned, its physical component is specifically designed for use in organic chemistry, thus excluding possible customization to other domains.

In this work, we evaluate our tool using two different datasets, one regarding an environmental disaster following Hurricane Florence and another about identifying geo-located animals. A work closely related to our approach presents an AR game that focuses on teaching students about endangered animals, structured around tangible cubes [25]. Players interact with three cubes to identify and learn more information about animals, in the form of educational videos, and make selections regarding their categorization. Although the physical components in both games are similar, Tile-o-Scope utilizes 12 tags to effectively capture a wider game area that allows players to make more matches and further involve the physical world. The volume of tags is also a deciding factor in the customization of the activity, without constraining players to one specific mechanic. An overview of Augmented Reality Games, highlighting their evolution from entertainment games to more affordable approaches on serious games can also be found in [74].

3 Pior Work: Cartoscope Classic

3.1 Volunteer Motivations in Paid Crowdsourcing Marketplaces [65]

In this thesis, we will be using crowdousrcing platforms such as Amazon Mechanical Turk (*MTurk*) as means of recruiting volunteers. MTurk is a widely popular online marketplace for crowdsourcing. MTurk allows *requesters* to post Human Intelligence Tasks (*HITs*) for *workers* to complete for payment. Although primarily intended for use as a paid crowdsourcing platform, recent work has shown that workers on MTurk are motivated by more than simply money [28] and that aspects such as the framed meaningfulness of a task can impact measures of worker performance [12]. Therefore, we considered MTurk as a means to recruit participants who may not be motivated purely by payment, but also voluntarily assisting in projects.

In order to gain insight into the motivations of workers on MTurk, and the interplay of paid versus volunteer work, we ran a HIT on MTurk using an initial prototype for Cartoscope Classic. We presented participants with a sequence of aerial photos of the Colorado Floods from 2013 and asked them to identify images containing damage. We posted a HIT on MTurk that used a *required work payment scheme*: workers were paid a fixed amount to rate *at least* some minimum required number of images; workers could then voluntarily continue rating images if they desired. The layout of the image rating page can be seen in Figure 5.

We carried out a 3x2 between subjects experiment design, using the following factors and levels.

- Amount of `required work`:

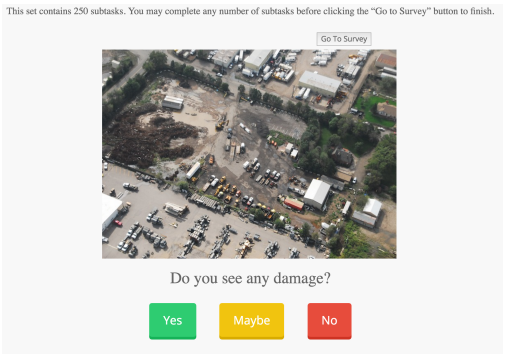
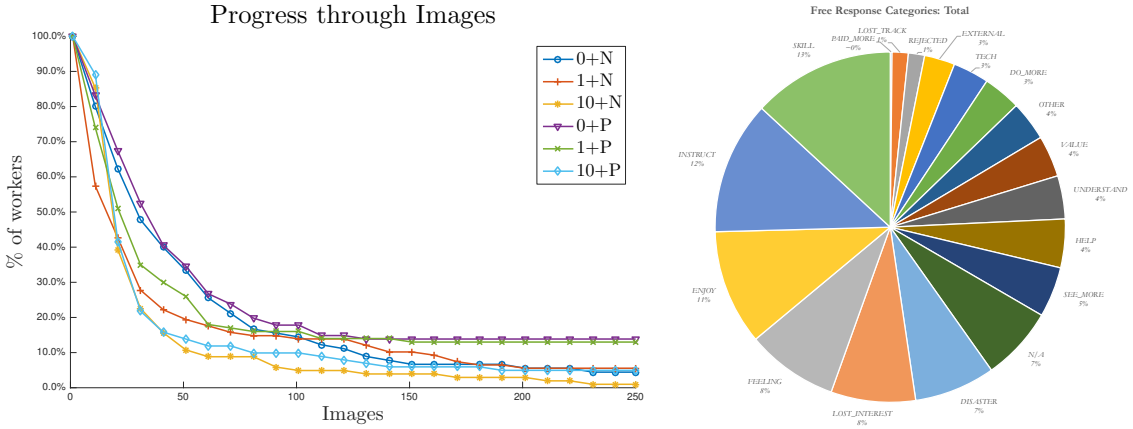


Figure 5: Screenshot of the image rating page used for our task. Instructions are shown along the top. In this condition, progress through the total set and progress through the required work are not available to workers.



(a) (b)

Figure 6: (a) Worker retention chart by image count. The x-axis shows the image count and the y-axis shows the percentage of workers in a condition that rated at least that many images. The rapid falloff after completing work in the `10+` conditions, and the large percentage of workers in the `0+P` and `1+P` conditions who competed all available images, can be clearly seen. (b) Breakdown of response categories for responses to the free-response motivational survey question.

  - `0+`: Participants could rate as many images as they wanted (even none) and finish rating at any time by proceeding to the survey. This examined requiring no work.
  - `1+`: Participants had to rate at least one image before proceeding to the survey. This examined requiring the smallest amount of work possible, to see if this would crowd-out motivation.
  - `10+`: Participants had to rate at least 10 images before proceeding to the survey. This examined requiring an order of magnitude more more work than `1+`, but still a relatively small amount.

- Presence of `progress feedback`:
  - `N`: No feedback on the current progress of ratings provided was present.
  - `P`: Feedback was given to the participants in the form of progress counts, showing how many images they have rated, as well as how many ratings were required (if any) and the total size of the set.

Other than the variations described here, workers received identical tasks. This resulted in 6 conditions, which we refer to using `0+N`, `1+N`, `10+N`, `0+P`, `1+P`, `10+P`.

Our results can be summarized as follows:

◇ *When less work was required, workers did more work, were more likely to indicate stopping due to a loss of interest, and were less likely to indicate stopping due to the instructions.*

Workers in the `0+` conditions did significantly more work beyond their requirements than the `1+` and `10+` conditions (*Extra Image Count*), and overall provided significantly more ratings (*Image Count*), and spent more time on the work (*Total Time*). On the contrary, workers in the `10+` conditions were significantly less likely to provide more work than required than workers in lower requirement conditions (*More than Required*), as well as less likely to complete the entire set (*Finished*).

Requiring work also impacted self-reported worker motivation. Analyzing workers' reported motivations for their performance, we observed that workers in the `10+` conditions were more likely to report complying with instructions as the primary reason for providing the amount of ratings they did, with 21.7% of the responses in these conditions belonging to the *INSTRUCT* category. In contrast, workers in the 0+ conditions were more likely to express interest up to the point when they decided to finish the task, with 13.1% of the responses in these conditions belonging to the *LOST-INTEREST* category. Compared with the previous quantitative measures, these results lend support to the presence of the motivation *crowding-out* effect.

◇ *Providing progress feedback resulted in more work done and improved perceived clarity of the task.*

Workers in conditions with progress feedback were significantly more likely to complete the entire image set of 250 images (*Finished*), as well as spend more time on the task (*Total Time*). These results indicate that providing progress feedback increased worker engagement in the task.

Giving workers progress feedback also reduced the amount of losing track of what they had done (*LOST-TRACK*) and fear of work being rejected (*REJECTED*).

◇ *Workers provided a variety of motivations other than payment for their participation.*

Our open coding approach revealed a variety of different motivations for workers performing beyond their requirements. Figure 6b depicts the breakdown of categories in responses to the first survey question.

The task was well received by workers, with 10.7% of the workers' free-response answers falling into the *ENJOY* category and another 4.5% clearly stating wanting to help the project as a primary reason for rating more images. Another 8.5% of workers reported that they felt the amount they had reached was adequate to their opinion and felt ready to finish the HIT (*FEELING*). The above responses indicate intrinsic motivations regarding the worker's personality and abilities. We also identified another motivation for workers, which has to do with the nature of the task, as 7.5% of responses fell into the *DISASTER* category. This group of workers were interested in analyzing disasters.

Notably, there were still some responses indicating potential misinterpretation of the instructions, including workers concerned about their work being rejected, even though we explicitly stated work would not be rejected, or expecting more payment for more work even though none was offered. This may be due to worker expectation based on the MTurk platform. This would indicate that clarity of instructions and setting clear expectations is of great importance.

3.2 Fixed Scheduling Variety Mechanisms for Crowdsourced Image Labeling [64].

In this work, we explored the use of *variety* to improve crowdsourcing worker engagement in image labeling tasks. We focus primarily on *behavioral engagement*—that is, the observable persistence and effort one puts toward a task [60]. We developed a task where workers were invited to complete a sequence of image related subtasks, specifically, *label subtasks*, a simple subtask where workers selected a label for an image from a set, and *map subtasks*, a more complex subtask where workers were asked to locate an image on a map and indicate if they were successful. When considering the *task complexity* [83] of the subtasks, we view the map subtask as more complex as it contains, for example, both more "information cues" (the image and the map) and "required acts" (navigating the map and indicating success). We considered the baseline task to be a sequence of only label subtasks. To examine the effects of variety—in terms of subtasks with different complexity—on performance, we varied the frequency with which the more complex map subtasks were served to workers during their progress through completing label subtasks. We posted a Human Intelligence Task (HIT) on MTurk that paid workers a fixed amount to complete any number of subtasks, with no minimum amount of work required, and answer a post-survey about their experience. Workers were randomly assigned into one of 6 conditions with different proportions of subtasks. In the `AllL` and `AllM` *uniform conditions*,

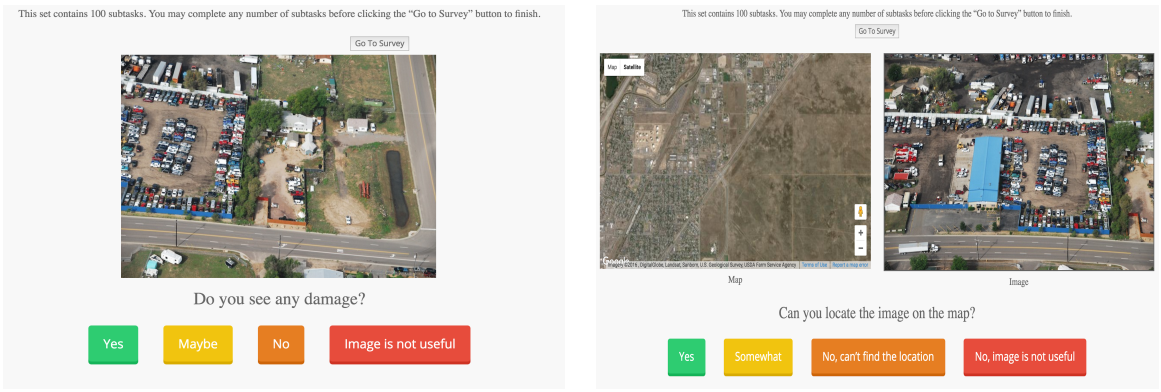


Figure 7: Example screenshots of the label (left) and map (right) subtasks.

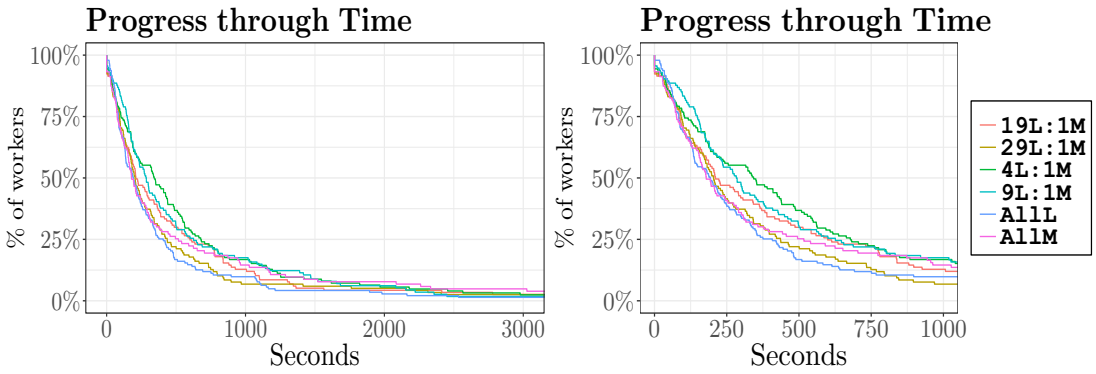


Figure 8: Survival chart for worker retention over time in the different conditions. The x-axis shows progress through time and the y-axis shows the percentage of workers retained up to that point; left shows up to 3000 seconds and right shows detail up to 1000 seconds. Note that early on, `4L:1M` and `9L:1M` retain more workers than conditions that serve map subtasks more or less frequently.

workers were served all label or all map subtasks, respectively. In the `4L:1M`, `9L:1M`, `19L:1M`, and `29L:1M` *variety conditions*, workers were served one map subtask at a regular interval of label subtasks (e.g., in `4L:1M` one map subtask was served after every 4 label subtasks).

A summary of the variables and statistical comparisons can be found in the Appendix. Figure 8 shows a survival chart, displaying worker retention over time for each of the conditions. We summarize our findings in the following points:

The map subtask was generally more complex than the label subtask. Looking at the `AllL` against `AllM` comparison allows us to compare worker behavior on each of the subtask types themselves. The increased *Subtask Time* workers spent working on a single map subtask is consistent with the increased time expected with increased complexity [10]. Additionally, among workers in the `AllM` condition, there was decreased *Subtask Agreement* and *Subtask Count*, along with increased in *Abandonment* of the HIT, compared to the `AllL` condition. Of note, though not statistically significant in all cases, is the observation that the `AllM` condition had the highest rate of abandonment and indication of reporting issues understanding the instructions. We do not consider this a surprising insight, rather supporting our initial assumption about the difference in complexity between subtask types.

Using only one type of subtask or the other did not impact total time spent on the task. Also looking at the `AllL` against `AllM` comparison, there was no significant difference in *Total Time* between the uniform subtask conditions, indicating that neither individual subtask type was inherently more engaging, with respect to time spent, than the other.

Interleaving subtask types did not observably impact speed or quality through distractions. When comparing

AllL against each of the conditions with subtask variety, no significant pairwise comparisons were found among *Label Time* or *Label Agreement*. This indicates that workers remained focused on the task and the quality of their work was not negatively impacted by the introduction of subtask type switches.
In some conditions, inserting map subtasks into a label subtask sequence increased the time spent on the task. Again comparing AllL against the subtask variety conditions, workers in the 4L:1M and 9L:1M conditions spent significantly more *voluntary* time on the task, when compared to the AllL condition. However, the 4L:1M condition was combined with a significant reduction in label count, while the 9L:1M condition did not show a negative impact on label count. Also of note, though not statistically significant in all cases, is the observation that uniform subtask conditions (AllL and AllM) had the lowest *Total Time* among all conditions.
There was little, if any, observable impact on workers' subjective experience between conditions. There were no significant differences in the various survey responses among conditions. Even *Understand*, which was significant in the omnibus test, did not show significant differences in the post-hoc tests. This indicates that we did not observe a difference in workers' subjective experience and opinions regarding *self-reported* engagement, concern about getting approved, or the amount of work worth doing for the payment, and so forth. Despite this, we did observe differences in workers' actual behavior. However, it is of note that the top two selected survey responses indicated wanting to help the project, feeling engaged, and wanting to see more images, and the lowest were not understanding the instructions and desiring more pay, across all conditions. This provides some additional support to workers participating voluntarily even thought they were receiving payment.

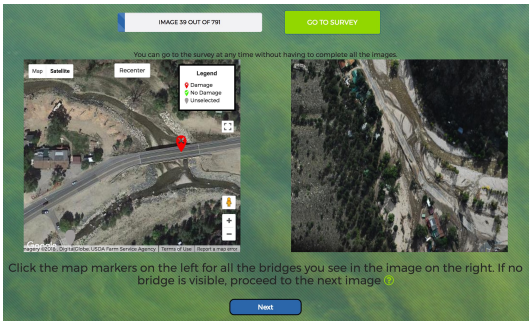
We believe these observations indicate that the comparison between AllL and 9L:1M is of particular interest. When compared with a sequence of simple subtasks (labels), interleaving a more complex subtask (maps) at a small interval increased the total time *voluntarily* spent on the HIT without observably negatively impacting other performance variables *of those workers who completed the HIT*. Figure 8 highlights the increased portion of workers in variety conditions with maps frequently interleaved who are retained early on. Examining other variety conditions indicates that the interval at which the map subtasks are interleaved matters. As the median *Subtask Count* was generally around 20 for conditions with many labels, it is possible that in the conditions with more space between the map subtasks, workers simply did not encounter enough maps to affect our measures, or most workers finished before even reaching a map.

3.3 Designing Contextual Geo-Location Enhanced Image Labeling Tasks [66,67]

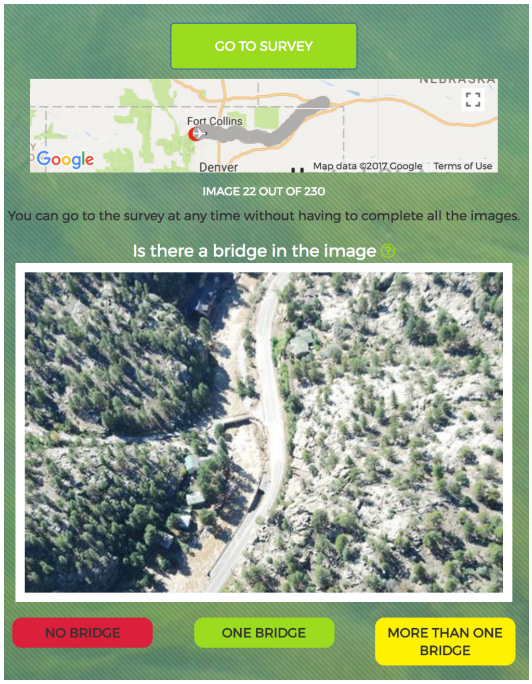
In the first part of this work, we explored providing two specific types of context to participants: first, varying the *order* in which they encountered images, and second, varying the presence of *map context* during the task. Showing images in their original sequence offered visual overlap to participants, creating a context continuity. We also considered map context, by visualizing the path of the images on a map, along with the current image location, with previously submitted labels annotated accordingly, creating a progress context. We were interested in whether the loss of context would have a negative impact on performance in a crowdsourcing setting. To this end, we posted a Human Intelligence Task (HIT) on Amazon Mechanical Turk, asking workers to label a series of images taken from a disaster scenario involving floods in the State of Colorado.

We found that removing context from the task interface did not negatively impact worker performance, including output rate, number of labels provided, and label quality (in terms of accuracy with ground truth). Showing images at random was as effective as preserving order and context, which indicates that other types of ordering, such as decision-theoretic means (Dai et al. [14]) may be better suited when designing a task for disaster response. Further, we found that including a map showing image locations and progress had a negative impact on worker completion of the task. This work contributes an empirical study of how designing a task to include context may impact behavior and performance in crowdsourced aerial imagery analysis, with a focus on disaster response applications.

In the second part of this work, we focused on designing an interactive crowdsourcing application for searching, identifying and labeling Points of Interest (POIs) using aerial photography data, by visualizing verified geo-location information on maps. Participants were shown aerial images and asked to click on the marked locations of any structures from the images on a map, in order to select and label them for damage. This design allows multiple structures captured by the same image source to be labeled at the same



POI Identification



Context

Figure 9: Interfaces for identifying POIs and providing geo-located context. (left) The main task, for the disaster dataset. The image source is visible on the right and the interactive map is on the left. Participants were asked to identify bridge structures in the image and search for the relevant marker on the map. They would then have to click on the marker to label the level of damage of the bridge, according to the image. (right) Showing contextual information as progress feedback for the *Flight Path* condition.

time, potentially minimizing the amount of images needed to acquire damage reports on the structures. We further explored and compared applications of our design in non-disaster settings. We found that crowd workers performed with varying levels of success, depending on the type of application. Specifically, workers showcased higher levels of accuracy in the non-disaster application, while their behavior in the disaster application indicated a higher level of difficulty for that case. For the labeling process, we selected two types of datasets, a *Disaster Dataset*, with a goal of identifying bridge damage after the Colorado Floods of 2013 and a *Non-disaster Dataset*, about identifying the condition of tennis courts in the Boston area.

4 Prior Work: Tile-o-Scope Grid

4.1 Variable Scheduling Variety Mechanisms for Crowdsourced Image Labeling [68].

Contrary to our previous work on fixed scheduling schemes, in this work we were interested in exploring Reinforcement Learning (RL) algorithms to serve task difficulty sequences. We designed a difficulty sequencing approach based on the reinforcement learning Q-learning algorithm, to generate sequences of level difficulties for players. These levels served as our different subtasks for the variety mechanism. We then compared our Q-learning approach against both uniform random and greedy sequencing methods, using an image matching game we developed called Tile-o-Scope Grid. Our Q-learning approach utilized a history of (up to) the last 3 difficulties encountered, plus a terminal state X for quitting. We used weighted tile collection as the reward, to indicate that tiles collected at higher difficulties are worth more than ones in the easiest, and used a Weighted Random Selection, using the squared value of action, to allow a small level of exploration, while maintaining strong preference for higher valued actions. Players were recruited by running Human

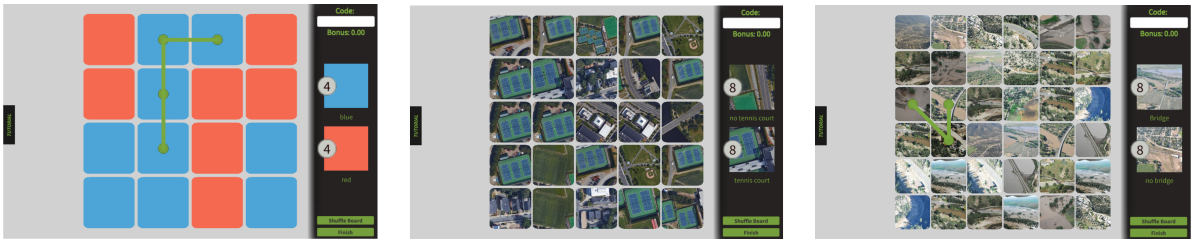


Figure 10: Example of the game interface, showing, from top to bottom, the difficulties E (easy), M (medium; Imagery ©2019 Google, Map data ©2019 Google) and H (hard; Images from U.S. Geological Survey).

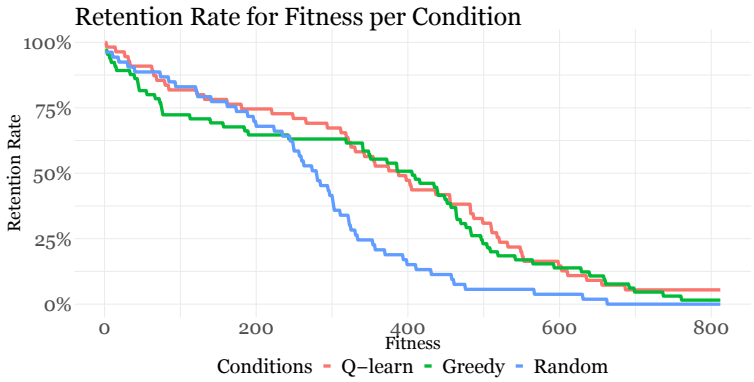


Figure 11: Retention rate of players over fitness.

Intelligence Tasks (HITs) on Amazon Mechanical Turk.

For our evaluation HIT, workers were randomly assigned to one of the following conditions:

- *Q-learn* (Q): Sequences generated using the Q-learning based algorithm.
- *Greedy* (G): Sequences that contain only the highest valued difficulty (H) based on the weight values.
- *Random* (R): Sequences generated by selecting a difficulty uniform randomly, using the 3 available difficulties. The difference between *Random* and *Q-learn* ordering is in the weights for randomly selecting a difficulty.

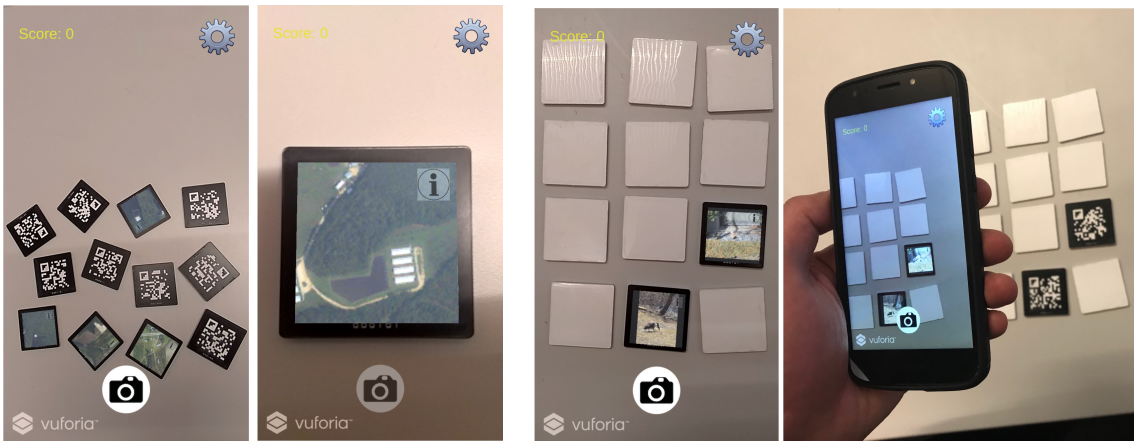
A summary of results can be found in the Appendix. The fitness metric corresponds to the sum of tiles collected from all categories, multiplied by the relevant weight of that category. For example, if a player collected 10, 20 and 30 tiles from E, M and H categories respectively, the fitness value would be $10 \times 0.0 + 20 \times 1.0 + 30 \times 1.2 = 56$. This metric is an indicator of meaningful labels provided, as it takes into consideration the importance of the level's difficulty.

Our post-hoc pairwise comparison analysis revealed that the *Q-learn* condition outperformed the *Random* condition both in terms of weighted tiles collected, as observed in the fitness metric, as well as the number of moves and total time spent playing the game. While players in the *Random* condition completed the most levels, that can likely be attributed to the presence of more E levels than in the other conditions.

When comparing *Q-learn* to the *Greedy* approach, which serves only levels of the highest value, i.e H difficulty, we found that players completed significantly fewer levels, spent significantly more time per level and attempted moves of significantly smaller length in the *Greedy* condition.

The fitness of the *Q-learn* to the *Greedy* approaches were not found to be significantly different. However, we note that *Q-learn* had the highest mean fitness, and higher early fitness in the retention curve (discussed below), which may indicate directions for further exploration of this approach.

Additionally, of the 3 conditions, *Greedy* offers the least amount of variety, as it essentially repeats the same category. As a result, players in this condition only encounter and contribute to one dataset, providing



(a) (b)

Figure 12: (a) Screenshot from the game, using the *Florence* dataset. Physical tags are read using the mobile application and converted into images. (b) Playing Memory using the toolkit.

the least diverse label output. On the other hand, the *Random* and *Q-learn* conditions are able to leverage the game design to combine multiple datasets, which means that players are exposed to different types of labeling tasks and end up contributing to more datasets.

The retention rate of players for fitness across all conditions can be found in Figure 11. Both the *Random* and *Greedy* conditions experience an earlier drop-off in the percent of players above a given fitness, although *Greedy* recovers later. Players in the *Random* condition end up contributing significantly less, while the behavior of players in the *Greedy* condition is in line with findings in the literature that indicate that most participants quit early on, with a smaller subset ending up contributing bulk of the work. In contrast, the *Q-learn* condition is, visually, able to provide a smoother drop-off of participants, suggesting a potentially better approach for retaining engagement in the task.

5 Prior Work: Tile-o-Scope AR

5.1 An Augmented Reality Tabletop Toolkit for Image Labeling

While the first part of this thesis focuses on task variety mechanisms for increasing retention in citizen science projects, the second part will examine how to enhance the image labeling experience, by incorporating co-located discussion elements. This is motivated in part in research conducted in the field of citizen science. Social components of projects in this area have been have been shown to significantly correlate with improvement in scientific literacy [56]. Recent work has also shown that feelings of satisfaction of contributing to projects and engagement are linked to social outcomes, such as a broader sense of community [20,36], as well as increased trust and reconnecting people with each other [53]. Allowing users to discuss data while also performing analysis opens up a new space for social conversation, through what Wylie and Albright have described as "Civic Technoscience" [84]. The benefits of engaging volunteers across disciplines to contribute to citizen science has been the emphasis of a wide body of work [8,17,24,63].

To this end, we focused on the following research questions, within the scope of crowdsourced image labeling tasks:

- **RQ3.1:** How can image labeling tasks be enhanced by introducing a co-location augmented reality game toolkit, such as by fostering discussions among participants?
- **RQ3.2:** How can customizable rulesets engage participants of different backgrounds using the same toolkit?

To approach these two research questions, we designed a multi-person, co-located image labeling toolkit that can be played using physical tags and a mobile device such as a phone or tablet, called Tile-o-Scope AR. The game utilizes augmented reality (AR) technology to read physical tags, and show images on the device in their place, which can then be sorted, by pointing the device over the physical tags to reveal the corresponding images on the screen, and grouping similar images into categories.

An overview of the interface can be seen in Figure 12. First, players are prompted to choose a dataset for image sorting and are then provided with the available categories at the start of the game. The categories are always available by clicking on the gear icon. Players have to point the camera over the tiles, in order to scan the tags, which are then replaced with the actual images on the mobile device screen. In order to mitigate potential cognition overload due to the amount of tags present, we limited the amount of images that can be shown at the same time to four. Once the player has identified 2 to 4 images that they believe belong to the same category, they can use the camera button to capture these images. A confirmation dialog then appears, containing their selection, to ensure no images have been selected by accident. Upon confirmation, the selection is logged in the system, and a text prompt informs the player whether the selection was a correct match or not, accompanied by appropriate audio feedback.

Image labelling from image matching can be achieved by adding some images with ground truth (i.e. whose correct category is already known) in the playable dataset. Therefore, an *incorrect match* would occur if players attempt to match images with ground truth that belong to separate categories. A *correct match* would occur if players matched images with ground truth that belong to the same category. Finally, a *valid match* would occur if players matched a combination of ground truth images that belong to the same category, with one or more images that have no ground truth. We can therefore identify a unique candidate category for all images in the selection whose category we don't know.

To evaluate our design, conducted two studies. In the first study, we recruited 14 student participants (5 male, 9 female) from the university for a small-scale qualitative user study. Participants were put into groups of 2 or 4 people who sat either next to to each other (groups of 2) or across (groups of 4). In total, there were 2 groups of 4 and 3 groups of 2 people. Participants were offered a $15 Amazon Gift Card for participating. We aimed for groups including various disciplines, from Computer Science (6 people) and Social Sciences (8 people), and with varying familiarity and experience in games and applications. After a briefing and a small tutorial on how to use the toolkit, we first asked participants to play a simple game (Memory) and then to brainstorm new games and designs using the core mechanics of the tool. The total duration of the study was 1 hour for every group.

All sessions were transcribed and then sorted using an affinity diagram technique [31]. Three raters independently highlighted observations from all sessions, which were then collaboratively organized into distinct clusters to form the resulting affinity diagram. Affinity diagrams are often used when sorting complex data, as this technique facilitates identifying unbiased themes. Our findings are summarized in the following categories:

Collaborative In Person Image Sorting. All of the groups started collaborating on the game almost immediately, exchanging ideas about the images they were seeing and sharing observations about which images they thought belonged to which categories, even if this action would effectively help their opponents. One group in particular evolved rapidly from a competition among four players to a collaboration as one, in order to categorize all images, often pausing to discuss what they did wrong in an incorrect match and offer opinions on the image content. One participant noted that the thing she liked the most was "talking to the group and comparing images."

Regarding using the tool for building community, one participant noted that "I think that it would have pretty strong potential for that." Participants across different groups commended the toolkit's ability to initiate conversations by getting people together in a room and helping them understand the problem better. When discussing this notion with one of the larger groups, participants also suggested that they could see neighbors coming together to play this game and use this as a monitoring tool of facilities among the neighborhood.

A number of participants across different groups expressed motivation due to contributing to a meaningful project ("I liked that I was helping someone,"), ("The scientific layer to it is super interesting,"). Several people also commented on the potential of using this kind of activity for educational purposes ("I thought it was interesting to find pictures that looked like they were not related to. To figure out they are the same animal. [...] With datasets that are more varied, where it is not very clear these two things match, I would

enjoy learning about that. This is a good way to learn by that,"). In a session involving a dataset monitoring CAFOs in North Carolina, one group talked about how they felt they were learning about manure lagoons and how to identify issues such as flooded structures while playing the game.

Grassroots Game Design. Participants came up with a variety of games, ranging from simple adaptations of the Memory game or of existing games, to completely novel ideas. Notable suggestions of existing games were Go Fish, Candy Crush and Mahjong Solitaire, and a variation of CAPTCHA that requires flipping non-relevant tiles positioned on a grid and capturing the rest as one category. The above suggestions can be readily played using Tile-o-Scope AR, requiring no software modifications. Adaptations for Memory included rolling n-sided dice that would dictate which of the n categories should be matched next, flipping a coin for a chance to chain matches or lose everything, and increasing the difficulty in various ways, such as making tiles visible for a certain time limit. While these games require an external element (dice, coins, timers) with our current implementation, such utilities could be incorporated in a future version with minimal software modifications. Although the toolkit was designed with multi-person activities in mind, some participants suggested using it as a single player activity as well, to play games when traveling, or to relax.

Participants were able to go beyond adapting existing games, by brainstorming completely new ideas and discussing in detail the newly formulated rules, even giving these new games their own novel names. Some games suggested were "Convince-a-Match", a game involving subterfuge and negotiation tactics, where players must convince or fool other players into making correct or incorrect matches, and "BattleTile", where participants must describe images to other players. Story-driven games were also proposed ("A puzzle or story game where you start with one animal and it wants you to do something for another animal and you have to find that animal,").

Adaptability to Other Image Labeling Projects. Participants in both dataset conditions were able to identify and suggest other possible use cases for our toolkit, regarding related image labeling projects. In the *Florence* dataset, one discussion among participants in one of the bigger sized groups revealed potential for use of the toolkit in disaster response and damage assessment.When discussing neighborhood collaboration on the topic of hog farms, another participant in the *Florence* dataset also pointed out the potential use of the toolkit on a project about identifying and reporting violations. In the *Animals* dataset, one participant suggested using the toolkit for a project regarding predator-prey relationships and animal conservation ("You are a zookeeper and you have to put animals in appropriate places so they don't conflict or attack each other"). These suggestions highlight the toolkit's ability to simultaneously support multiple different image labeling projects, without requiring any software modifications. Community organizers would only need to upload an image set, which can be either their own, or any publicly available or crowdsourced one, and then develop an activity that best suits their needs.

Interactions with Physical Component. The combination of physical and digital components was generally well received ("I liked that it was a virtual game but also something you touch with your hand"). Participants enjoyed being able to interact with physical tiles with comments such as ("I liked the tiles, they make a nice sound [...] It felt satisfying slamming them together", *Florence*), ("I like playing with the tiles. Physical things help me relax.", *Animals*). Another interesting observation was about the use of the collected tiles from different participants. In groups of bigger sizes, we found that some participants used their stacks of earned tiles as an added element of interaction within the group, in friendly banter to demonstrate that they were good at the task. The stack formations were also used as a motivator from other members of the group to improve their performance and to help the group collect all tiles.

Multiple participants praised being able to customize the physical component as well with sentences like ("I like that they are tiles and that they are a little bit heavy. Tiles are so versatile. [...] If you want to have an impromptu playing session, you can just print out a number of these and you would be ok.") Some participants even considered building a deck by mixing and matching different physical tiles and playing them when traveling.

To explore how customizable rulesets can be utilized to engage participants from varying backgrounds (RQ3.2), we conducted a second, mixed study, where we invited groups from differing research fields to perform three different image labeling activities with the toolkit and answer some questions about their experience at the end of each activity. After the end of the play-session, we asked participants to rate the three activities, as well as answer a set of open-ended questions about their overall experience.

We invited participants to play the following activities:

- **Sorting (no-game condition):** Participants were asked to sort images into the available categories, by using the available matching or labeling mechanics. They were given the option to work in any way they chose, i.e. either as a team, in subgroups, or individually. Used 12 tiles.
- **Memory (competitive game condition):** Participants were asked to play a competitive game of Memory, similar to the first study. Used 12 tiles.
- **TrekStack (collaborative game condition):** Tiles are placed on a grid and players must make appropriate matches, pushing with a "hand" tile, in order to move the tile on top of which the pawn sits, to reach the goal. A visualization of the rules can be found in Figure **??**. The grid used for the board was 3×3. Used 10 tiles.

We conducted three studies, with groups of different backgrounds. Each group consisted of four participants. Each participant was given their own mobile device. The order with which each group encountered the three activities was randomized. The groups were as follows:

- *Game Designers (`GD`)*: Graduate students with a background in Game Design. The average age of the group was 27 (two female, one male, one chose not to disclose). Study was conducted on campus. Some of the participants knew each other prior to the study.
- *Sociology (`SC`)*: Faculty from the Social Sciences department. The average age of the group was 53 (three female, one male). Study was conducted on campus. All participants in this group knew each other as colleagues prior to the study.
- *Environmental Health & Justice (`EH`)*: Researchers from an environmental health fellowship program attending a bi-yearly conference off campus. The average age of the group was 38 (two female, two male). Compared to the previous two groups, this study was conducted during dinner at a restaurant. This allowed us to explore how the toolkit can be used outside of a lab setting. A thirty minute break occurred between the first activity and the second, during which participants had dinner. The researchers all knew each other well through the fellowship program and studying environmental health and justice issues.

Our findings revealed that each group had different activity preferences. In particular, `GD` group enjoyed the collaborative game (TrekStack) the most, while `EH` preferred Memory, and `SC` rated the Sorting activity highest, followed by Memory. There was no apparent relationship between the order in which a group played a game and reported enjoyment levels, as two of the groups reported as most enjoyable the second activity they played, and one reported the first. This indicates that groups with different backgrounds enjoy different kinds of activities and that no single pre-defined ruleset was able to be equally enjoyable by all groups. Our findings support the need for a toolkit that can easily support a variety of different activities, in order to maximize engagement levels across participants with different backgrounds. Moreover, every group approached each activity in a different manner. A summary of comparisons for each group can be found in Table 10 in the Appendix.

6 Future Work: Tile-o-Scope Grid

6.1 Proposed Future Work

6.1.0.1 Adaptive Task Variety for Crowdsourced Image Labeling:

While our work so far both in the fixed and variable scheduling schemes has revealed promising direction for utilizing task variety towards retaining participants in image labeling tasks, our implementation takes into consideration only historical data, and the schedules are generated only at the beginning of the task. However, recent work on using Reinforcement Learning on games has examined purely adaptive approaches that work in an online fashion, taking into consideration both current and past data, as well as the actions and progress of a player during the game. One such example is Q-DeckRec, which uses Q-learning to build more efficient decks in Collectible Card Games (CCGs) [13].

Figure 13: A mockup of the communication intervention, showing placement and reading of messages through cairns.

Moreover, our task variety scheduling mechanisms thus far have utilized only contextually related types of tasks. However, an alternative approach of introducing variety can be achieved by utilizing non contextually related types of tasks, that can be considered "interventions", in between image labeling tasks that are of interest. Our motivation in exploring the use of interventions is in line with recent work in image labeling literature. One such example is the work by Dai et al. [15], that utilized "micro-diversions" in between subtask sequences. These items were not contextually related to the contents of the main subtasks, i.e. asking participants to read pages of a graphic novel. Their results indicated that such micro-diversions could improve worker engagement.

We thus propose to extend our current Q-learning based implementation in two ways. First, in order to generate more player tailored subtask sequences, our new implementation will take into consideration both historical and current data, as well as the specific progress of the player. Instead of generating a sequence at the start of the game, the subtasks sequence for each player will be updated in intervals during the entire task, to ensure adaptability both in terms of all players but also the individual. Second, we propose to broaden the pool of available actions, by introducing a new, non-label related action, called "leave a cairn", where participants will be able to leave short messages for other users, and then observe their message being place on the top of the cairn formation, consisting of the most recent messages left from other participants. A mockup is shown in Figure **??**. The introduction of this new type of subtask serves a dual purpose, as not only can it be used to increase the variety of subtasks encountered in a non-label related way, but also the available cairn messages can be chosen in a way that reflect useful feedback from players about their experience and progress in the game. Therefore, the type of message players choose to leave for others can also be used as information in the Q-learning algorithm.

Our proposed algorithm will be implemented on Tile-o-Scope Grid, although it can be generalized to other platforms, such as Cartoscope Classic as well. Extending our previous work, our algorithm will generate policies for either serving subsequent levels to players, ranging from easy, medium to hard, or, to serve a "leave cairn" intervention. Instead of generating a sequence once at the beginning of the game, Tile-o-Scope Grid will query the backend server for updated sequences at specific intervals during the game, to ensure players are on the most up to date path. The performance of the specific user will also be encoded in the state, in the form of three separate encodings, depending on the user accuracy thus far. While our previous work allowed for storing the Q-Table in memory, by increasing both the action space size by adding two new actions, and also considering current data as well, this may not be feasible. To this end, we will explore other techniques, such as using Multi-Layer Perceptron (MLP), similar to work done by Chen et al. [13].

To evaluate our algorithm, we will recruit participants through Amazon Mechanical Turk to play as many levels in Tile-o-Scope Gridas they want. To measure our algorithm's performance, we will use quantitative metrics such as tile count, level count, fitness (i.e. weighted tile collection), time spent, etc. Participants will be randomly sorted into the following conditions:

- *Random* (R): Sequences generated by selecting a difficulty uniform randomly, using the 3 available difficulties.

- *Greedy* (G): Sequences that contain only the highest valued difficulty level, which will be determined by comparing all available levels.
- *Q-learn+Cairns* (QC): Sequences generated using the extension of our Q-learning based algorithm, including the cairns interventions.

6.1.1 Comparing Cartoscope Classic to Tile-o-Scope Grid:

Cartoscope Classic Tile-o-Scope Grid

Figure 14: Screenshots of the image labeling task for Cartoscope Classic and Tile-o-Scope Grid. An example of the animal identification task on Cartoscope Classic can be seen on the left, while an example of the tennis court identification task on Tile-o-Scope Grid can be seen on the right.

Thus far, our results indicate that the two online image labeling interfaces we have been developed have a strong potential towards engaging participants in crowdsourcing marketplaces such as Amazon Mechanical Turk. Cartoscope Classic has been used to explore payment schemes and worker motivations on MTurk, along with fixed scheduling schemes and contextual information elements as part of the interface design, while Tile-o-Scope Grid has been used to examine more dynamic and adaptive scheduling schemes for increasing engagement.

However, the underlying image labeling mechanism for these interfaces is significantly different. Cartoscope Classic follows the traditional approach for image labeling, where images are labeled serially, by asking the user to take one action on each image at a time. Moreover, in Cartoscope Classic, users have full access to the image at hand, with an additional ability to zoom, which makes more information available before making a decision on a label. We therefore expect that labels coming from Cartoscope Classic will be of higher quality but lower quantity, due to the nature of performing one image label at a time. On the other hand, Tile-o-Scope Grid achieves image labeling by using an image matching game mechanism, which allows users assign many images to one category at the same time. However, as all images are placed on a grid, this means fewer space for each individual image, which may pose problems for more difficult datasets. We therefore expect that labels coming from Tile-o-Scope Grid will be of higher quantityt but perhaps lower quality.

We thus propose to conduct a comparison between the two image labeling interfaces, in order to determine under which conditions one may be able to perform better than the other. To do that, we will recruit participants through Amazon Mechanical Turk, who will be randomly sorted into performing image labeling either on Cartoscope Classic or Tile-o-Scope Grid. To measure performance, we will use the same quantitative metrics across both interfaces, such as total number of images labeled, total time spent on the entire task, accuracy and crowd agreement. To measure engagement, we will include a questionnaire at the end of the task for both interfaces, which will focus on the Intrinsic Motivation Inventory's (IMI) enjoyment subscale [16]. We will be using the image labeling task on Cartoscope Classic and a 4×4 grid arrangement on Tile-o-Scope Grid.

We hypothesize that the nature of the task will be a factor in how well the crowd performs on either platform, which is in part motivated by participant performance so far in both interfaces. For example, a task that requires a certain amount of searching in the image to identify specific minor patterns or Points

of Interest will be more challenging to perform on Tile-o-Scope Grid than Cartoscope Classic. On the other hand, a task where a decision about a label can be made by taking into consideration the image as a whole will be easier and faster to perform on Tile-o-Scope Grid but perhaps perceived as less enjoyable and mundane on Cartoscope Classic. We therefore propose to compare the two interfaces using the following tasks:

- *Identifying Tennis Courts*: This task will be developed using images from Google Maps around the Boston area. We consider this the least challenging task and we hypothesize that accuracy levels will be equally high for both interfaces, but that enjoyment levels may be higher for Tile-o-Scope Grid.
- *Identifying Reptiles and Mammals*: This task will be developed using images sourced from the iNaturalist API [23]. Participants will be asked to label various animals into the Amphibia or Reptilia category. An example of the task on Cartoscope Classic can be seen in Figure 14. We consider this a task of medium difficulty, as well as an example of a common citizen science task.
- *Identifying Bridge Damage*: This task will be developed using images from the Colorado Floods of 2013 captured by Civil Air Patrol. As these images are taken from a certain altitude and may not always have a focus on a specific bridge, we hypothesize that participants on Cartoscope Classic will achieve higher accuracy levels than in Tile-o-Scope Grid. We consider this the most challenging task, as well as an application of a disaster scenario.

6.2 Broader Impacts

Work on this direction falls under the main Cartoscope research project, which aims to generate open-sourced tools for community led image labeling. These include both the core Cartoscope image labeling and project creation platform, as well as the Tile-o-Scope Grid image matching web-based game. These tools can be used and adapted by crowdsourcing communities in their image analysis efforts. Increased performance in image labeling tasks, achieved by deploying our task variety mechanisms, can be utilized by organizations to gather more data and achieve deeper engagement with their audience. Moreover, our mechanisms can inform the design of more effective algorithms for image labeling in other crowdsourcing platforms, beyond the ones developed under the Cartoscope umbrella.

7 Timeline

Time	Task
Spring 2020	• Adaptive Task Variety for Crowdsourced Image Labeling • Adaptive Task Variety for Crowdsourced Image Labeling • Comparing Cartoscope Classic and Tile-o-Scope Grid • Adaptive Task Variety for Crowdsourced Image Labeling: Paper Submission
Summer 2020	• Comparing Cartoscope Classic and Tile-o-Scope Grid: Paper Submission • Thesis Defense.

Table 1: Projected Timeline Until Graduation

References

[1] Oguz Ali Acar and Jan van den Ende. 2011. Motivation, reward size and contribution in idea crowdsourcing. *DIME-DRUID ACADEMY. Comwell Rebild Bakker, Aalborg, Denmark* (2011).

[2] Electronic Arts. 2019. Bejeweled. (June 2019). https://www.ea.com/games/bejeweled

[3] Marc Baaden, Olivier Delalande, Nicolas Ferey, Samuela Pasquali, Jérôme Waldispühl, and Antoine Taly. 2018. Ten simple rules to create a serious game, illustrated with examples from structural biology. *PLOS Computational Biology* 14, 3 (March 2018), e1005955. DOI:http://dx.doi.org/10.1371/journal.pcbi.1005955

[4] Luke Barrington, Shubharoop Ghosh, Marjorie Greene, Shay Har-Noy, Jay Berger, Stuart Gill, Albert Yu-Min Lin, and Charles Huyck. 2012. Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics* 54, 6 (Jan. 2012). DOI:http://dx.doi.org/10.4401/ag-5324

[5] Mark Billinghurst. 2002. Augmented reality in education. *New horizons for learning* 12, 5 (2002), 1–5.

[6] Brookings-Bern Project on Internal Displacement. 2011. A Year of Living Dangerously: A Review of Natural Disasters in 2010. http://www.refworld.org/docid/4dabde142.html. (April 2011).

[7] Keith R Bujak, Iulian Radu, Richard Catrambone, Blair Macintyre, Ruby Zheng, and Gary Golubski. 2013. A psychological perspective on augmented reality in the mathematics classroom. *Computers & Education* 68 (2013), 536–544.

[8] Wouter Buytaert, Zed Zulkafli, Sam Grainger, Luis Acosta, Tilashwork C. Alemie, Johan Bastiaensen, Bert De Bièvre, Jagat Bhusal, Julian Clark, Art Dewulf, Marc Foggin, David M. Hannah, Christian Hergarten, Aiganysh Isaeva, Timothy Karpouzoglou, Bhopal Pandeya, Deepak Paudel, Keshav Sharma, Tammo Steenhuis, Seifu Tilahun, Gert Van Hecken, and Munavar Zhumanova. 2014. Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science* 2 (2014). DOI:http://dx.doi.org/10.3389/feart.2014.00026

[9] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. 2016. Chain reactions: the impact of order on microtask chains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[10] Donald J. Campbell and Karl F. Gingrich. 1986. The interactive effects of task complexity and participation on task performance: a field experiment. *Organizational Behavior and Human Decision Processes* 38, 2 (Oct. 1986), 162–180. DOI:http://dx.doi.org/10.1016/0749-5978(86)90014-2

[11] Cartosco.pe. 2019. https://cartosco.pe. (2019). https://cartosco.pe Accessed: 2019-04-06.

[12] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (June 2013), 123–133. DOI: http://dx.doi.org/10.1016/j.jebo.2013.03.003

[13] Zhengxing Chen, Christopher Amato, Truong-Huy D. Nguyen, Seth Cooper, Yizhou Sun, and Magy Seif El-Nasr. 2018. Q-DeckRec: a fast deck recommendation system for collectible card games. In *2018 IEEE Conference on Computational Intelligence and Games*. 1–8.

[14] Peng Dai, Mausam, and Daniel S. Weld. 2010. Decision-theoretic control of crowd-sourced workflows. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.

[15] Peng Dai, Jeffrey M. Rzeszotarski, Praveen Paritosh, and Ed H. Chi. 2015. And now for something completely different: improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, Vancouver, BC, Canada, 628–638. DOI:http://dx.doi.org/10.1145/2675133.2675260

[16] Edward Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. DOI:http://dx.doi.org/10.1007/978-1-4899-2271-7

[17] Sascha Dickel, Christoph Schneider, Carolin Thiem, and Klara-Aylin Wenten. 2019. Engineering Publics: The Different Modes of Civic Technoscience. *Science & Technology Studies* (May 2019), 8–23. DOI: http://dx.doi.org/10.23987/sts.59587

[18] Yihan Gao and Aditya Parameswaran. 2014. Finish them!: pricing algorithms for human computation. *Proceedings of the VLDB Endowment* 7, 14 (Oct. 2014), 1965–1976. DOI:http://dx.doi.org/10.14778/2733085.2733101

[19] J. Richard Hackman and Greg R. Oldham. 1976. Motivation through the design of work: test of a theory. *Organizational Behavior and Human Performance* 16, 2 (Aug. 1976), 250–279. DOI:http://dx.doi.org/10.1016/0030-5073(76)90016-7

[20] Benjamin K Haywood. 2016. Beyond data points and research contributions: the personal meaning and value associated with public participation in scientific research. *International Journal of Science Education, Part B* 6, 3 (2016), 239–262.

[21] NR Hedley. 2003. Empirical evidence for advanced geographic visualization interface use. In *International cartographic congress, Durban, South Africa*. Citeseer.

[22] Robin Hunicke. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. 429–433.

[23] iNaturalist.org. 2019. https://www.inaturalist.org. (2019). Accessed: 2019-04-06.

[24] Kirk Jalbert, Samantha Malone Rubright, and Karen Edelstein. 2017. The Civic Informatics of FracTracker Alliance: Working with Communities to Understand the Unconventional Oil and Gas Industry. *Engaging Science, Technology, and Society* 3, 0 (Sept. 2017), 528–559. DOI:http://dx.doi.org/10.17351/ests2017.128

[25] Carmen M. Juan, Giacomo Toffetti, Francisco Abad, and Juan Cano. 2010. Tangible Cubes Used As the User Interface in an Augmented Reality Game for Edutainment. In *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies (ICALT '10)*. IEEE Computer Society, Washington, DC, USA, 599–603. DOI:http://dx.doi.org/10.1109/ICALT.2010.170

[26] Amy M Kamarainen, Shari Metcalf, Tina Grotzer, Allison Browne, Diana Mazzuca, M Shane Tutwiler, and Chris Dede. 2013. EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education* 68 (2013), 545–556.

[27] Hannes Kaufmann and Andreas Dünser. 2007. Summary of usability evaluations of an educational augmented reality application. In *International conference on virtual reality*. Springer, 660–669.

[28] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker motivation in crowdsourcing – a study on Mechanical Turk. In *Proceedings of the Americas Conference on Information Systems*.

[29] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. DOI:http://dx.doi.org/10.1145/2441776.2441923

[30] George Koutromanos, Alivisos Sofos, and Lucy Avraamidou. 2015. The use of augmented reality games in education: a review of the literature. *Educational Media International* 52, 4 (2015), 253–271. DOI: http://dx.doi.org/10.1080/09523987.2015.1125988

[31] Mike Kuniavsky. 2003. *Observing the User Experience: A Practitioner's Guide to User Research*. Elsevier. Google-Books-ID: 1tE4Skp9pI8C.

[32] Walter S. Lasecki, Adam Marcus, Jeffrey M. Rzeszotarski, and Jeffrey P. Bigham. 2014. *Using microtask continuity to improve crowdsourcing*. Technical Report CMU-HCII-14-100. School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. http://reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-14-100.pdf

[33] Walter S. Lasecki, Jeffrey M. Rzeszotarski, Adam Marcus, and Jeffrey P. Bigham. 2015. The effects of sequence and delay on crowd work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1375–1378. DOI:http://dx.doi.org/10.1145/2702123.2702594

[34] Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with Amazon Mechanical Turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Edinburgh, United Kingdom, 1546–1556. http://dl.acm.org/citation.cfm?id=2145432.2145597

[35] Hjalti Leifsson and Jóhann Örn Bjarkason. 2015. *Project Discovery-Advancing scientific research by implementing citizen science in EVE Online*. Ph.D. Dissertation.

[36] Eva J Lewandowski and Karen S Oberhauser. 2017. Butterfly citizen scientists in the United States increase their engagement in conservation. *Biological Conservation* 208 (2017), 106–112.

[37] Nai Li, Yuan Xun Gu, Leanne Chang, and Henry Been-Lirn Duh. 2011. Influences of AR-supported simulation on learning effectiveness in face-to-face collaborative learning for physics. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*. IEEE, 320–322.

[38] Sophia B. Liu. 2014. Crisis crowdsourcing framework: designing strategic configurations of crowdsourcing for the emergency management domain. *Computer Supported Cooperative Work* 23, 4-6 (July 2014), 389–443. DOI:http://dx.doi.org/10.1007/s10606-014-9204-3

[39] Anna Loparev, Lauren Westendorf, Margaret Flemings, Jennifer Cho, Romie Littrell, Anja Scholze, and Orit Shaer. 2017. BacPack: Exploring the Role of Tangibles in a Museum Exhibit for Bio-Design. In *Proceedings of the Tenth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '17*. ACM Press, Yokohama, Japan, 111–120. DOI:http://dx.doi.org/10.1145/3024969.3025000

[40] Miguel Angel Luengo-Oroz, Asier Arranz, and John Frean. 2012. Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears. *Journal of Medical Internet Research* 14, 6 (2012), e167. DOI:http://dx.doi.org/10.2196/jmir.2338

[41] Fred C. Lunenburg. 2011. Motivating by enriching jobs to make them more interesting and challenging. *International Journal of Management, Business, and Administration* 15, 1 (2011), 1–11.

[42] Joyce Ma, Lisa Sindorf, Isaac Liao, and Jennifer Frazier. 2015. Using a Tangible Versus a Multi-touch Graphical User Interface to Support Data Exploration at a Museum Exhibit. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '15)*. ACM, New York, NY, USA, 33–40. DOI:http://dx.doi.org/10.1145/2677199.2680555 event-place: Stanford, California, USA.

[43] Nickolas D Macchiarella, Dahai Liu, Sathya N Gangadharan, Dennis A Vincenzi, and Anthony E Majoros. 2005. Augmented reality as a training medium for aviation/aerospace application. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications Sage CA: Los Angeles, CA, 2174–2178.

[44] Nickolas D Macchiarella and Dennis A Vincenzi. 2004. Augmented reality in a learning paradigm for flight aerospace maintenance training. In *The 23rd Digital Avionics Systems Conference (IEEE Cat. No. 04CH37576)*, Vol. 1. IEEE, 5–D.

[45] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. 2016. Offline evaluation of online reinforcement learning algorithms. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 1926–1933.

[46] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. 2014. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. 1077–1084.

[47] Winter Mason and Duncan J. Watts. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, Paris, France, 77–85. DOI:http://dx.doi.org/10.1145/1600150.1600175

[48] Danny Mitry, Tunde Peto, Shabina Hayat, James E. Morgan, Kay-Tee Khaw, and Paul J. Foster. 2013. Crowdsourcing as a novel technique for retinal fundus photography classification: analysis of images in the EPIC Norfolk cohort on behalf of the UKBiobank Eye and Vision Consortium. *PLOS ONE* 8, 8 (2013), e71154.

[49] Ann Morrison, Antti Oulasvirta, Peter Peltonen, Saija Lemmela, Giulio Jacucci, Gerhard Reitmayr, Jaana Näsänen, and Antti Juustila. 2009. Like bees around the hive: a comparative study of a mobile augmented reality map. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1889–1898.

[50] Robert Munro, Tyler Schnoebelen, and Schuyler Erle. 2013. Quality analysis after action report for the crowdsourced aerial imagery assessment following Hurricane Sandy. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*.

[51] Sriram Narayanan, Sridhar Balasubramanian, and Jayashankar M. Swaminathan. 2009. A matter of balance: specialization, task variety, and individual learning in a software maintenance environment. *Management Science* 55, 11 (Nov. 2009), 1861–1876. DOI:http://dx.doi.org/10.1287/mnsc.1090.1057

[52] Johanna Okerlund, Evan Segreto, Casey Grote, Lauren Westendorf, Anja Scholze, Romie Littrell, and Orit Shaer. 2016. SynFlo: A Tangible Museum Exhibit for Exploring Bio-Design. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '16)*. ACM, New York, NY, USA, 141–149. DOI:http://dx.doi.org/10.1145/2839462.2839488 event-place: Eindhoven, Netherlands.

[53] Maria Peter, Tim Diekötter, and Kerstin Kremer. 2019. Participant outcomes of biodiversity citizen science projects: A systematic literature review. *Sustainability* 11, 10 (2019), 2780.

[54] Playdots, Inc. 2013. *Dots*. Game [Mobile]. (30 April 2013). Playdots, Inc., New York City, USA.

[55] Nathan Prestopnik and Dania Souid. 2013. Forgotten island: a story-driven citizen science adventure. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. Association for Computing Machinery, Paris, France, 2643–2646. DOI:http://dx.doi.org/10.1145/2468356.2479484

[56] C Aaron Price and Hee-Sun Lee. 2013. Changes in participants' scientific attitudes and epistemological beliefs during an astronomical citizen science project. *Journal of Research in Science Teaching* 50, 7 (2013), 773–801.

[57] John Quarles, Samsun Lampotang, Ira Fischler, Paul Fishwick, and Benjamin Lok. 2008. A mixed reality approach for merging abstract and concrete knowledge. In *2008 IEEE Virtual Reality Conference*. IEEE, 27–34.

[58] M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. 2010. Galaxy Zoo: exploring the motivations of citizen science volunteers. *Astronomy Education Review* 9, 1 (Dec. 2010). DOI:http://dx.doi.org/10.3847/AER2009036 arXiv: 0909.2925.

[59] Iulian Radu. 2014. Augmented reality in education: a meta-review and cross-media analysis. *Personal and Ubiquitous Computing* 18, 6 (Aug. 2014), 1533–1543. DOI:http://dx.doi.org/10.1007/s00779-013-0747-y

[60] Johnmarshall Reeve. 2015. *Understanding Motivation and Emotion (Sixth edition)*. Wiley, Hoboken, New Jersey.

[61] Henry Sauermann and Chiara Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences* 112, 3 (Jan. 2015), 679–684. DOI: http://dx.doi.org/10.1073/pnas.1408907112

[62] Karen Schrier. 2006. Using Augmented Reality Games to Teach 21st Century Skills. In *ACM SIGGRAPH 2006 Educators Program (SIGGRAPH '06)*. ACM, New York, NY, USA. DOI:http://dx.doi.org/10.1145/1179295.1179311 event-place: Boston, Massachusetts.

[63] Enric Senabre, Núria Ferran Ferrer, and Josep Perelló. 2018. Participatory design of citizen science experiments. (Jan. 2018). http://diposit.ub.edu/dspace/handle/2445/119189

[64] Sofia Eleni Spatharioti and Seth Cooper. 2017. On variety, complexity, and engagement in crowdsourced disaster response tasks. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*. Albi, France, 489–498. http://idl.iscram.org/files/sofiaelenispatharioti/2017/2037_SofiaEleniSpatharioti+SethCooper2017.pdf

[65] Sofia Eleni Spatharioti, Rebecca Govoni, Jennifer S. Carrera, Sara Wylie, and Seth Cooper. 2017. A required work payment scheme for crowdsourced disaster response: worker performance and motivations. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*. Albi, France, 475–488. http://idl.iscram.org/files/sofiaelenispatharioti/2017/2036_SofiaEleniSpatharioti_etal2017.pdf

[66] Sofia Eleni Spatharioti, Sara Wylie, and Seth Cooper. 2018a. Does Flight Path Context Matter? Impact on Worker Performance in Crowdsourced Aerial Imagery Analysis. (2018), 8.

[67] Sofia Eleni Spatharioti, Sara Wylie, and Seth Cooper. 2018b. Identifying and Assessing Points of Interest through Crowdsourced Image Analysis. In *ISCRAM 2018 Conference Proceedings – 15th International Conference on Information Systems for Crisis Response and Management*, Kees Boersma and Brian Tomaszeski (Eds.). Rochester Institute of Technology, Rochester, NY (USA), 1123–1125. http://idl.iscram.org/files/sofiaelenispatharioti/2018/2186_SofiaEleniSpatharioti_etal2018.pdf

[68] Sofia Eleni Spatharioti, Sara Wylie, and Seth Cooper. 2019. Using Q-Learning for Sequencing Level Difficulties in a Citizen Science Matching Game. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (CHI PLAY '19 Extended Abstracts)*. ACM, New York, NY, USA, 679–686. DOI:http://dx.doi.org/10.1145/3341215.3356299 event-place: Barcelona, Spain.

[69] Bradley R. Staats and Francesca Gino. 2012. Specialization and variety in repetitive tasks: evidence from a Japanese bank. *Management Science* 58, 6 (June 2012), 1141–1159. DOI:http://dx.doi.org/10.1287/mnsc.1110.1482

[70] Tobias Sturn, Michael Wimmer, Carl Salk, Christoph Perger, Linda See, and Steffen Fritz. 2015. Cropland Capture – a game for improving global cropland maps. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

[71] Alexandra Burchard Swanson. 2014. Living with lions: spatiotemporal aspects of coexistence in savanna carnivores. (July 2014). http://conservancy.umn.edu/handle/11299/167642

[72] Chek Tien Tan, Daniel Rosser, and Natalie Harrold. 2013. Crowdsourcing facial expressions using popular gameplay. In *SIGGRAPH Asia 2013 Technical Briefs*. Article 26, 4 pages. DOI:http://dx.doi.org/10.1145/2542355.2542388

[73] Chek Tien Tan, Hemanta Sapkota, and Daniel Rosser. 2014. BeFaced: a casual game to crowdsource facial expressions in the wild. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. Association for Computing Machinery, Toronto, Ontario, Canada, 491–494. DOI:http://dx.doi.org/10.1145/2559206.2574773

[74] Chek Tien Tan and Donny Soh. 2010. Augmented Reality Games: A Review. (2010).

[75] Tile-o-Scope AR. 2019. http://cartosco.pe/ar#/home_ar. (2019). http://cartosco.pe/ar#/home_ar Accessed: 2019-04-06.

[76] Tile-o-Scope Grid. 2019. http://cartosco.pe/Tiles. (2019). http://cartosco.pe/Tiles Accessed: 2019-04-06.

[77] R Brian Valimont, Dennis A Vincenzi, Sathya N Gangadharan, and AE Majoros. 2002. The effectiveness of augmented reality as a facilitator of information acquisition. In *Proceedings. The 21st Digital Avionics Systems Conference*, Vol. 2. IEEE, 7C5–7C5.

[78] Dennis A Vincenzi, Brian Valimont, Nickolas Macchiarella, Chris Opalenik, Sathya N Gangadharan, and Anthony E Majoros. 2003. The effectiveness of cognitive elaboration using augmented reality as a training and learning paradigm. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. SAGE Publications Sage CA: Los Angeles, CA, 2054–2058.

[79] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vienna, Austria, 319–326. DOI: http://dx.doi.org/10.1145/985692.985733

[80] Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 55–64. DOI: http://dx.doi.org/10.1145/1124772.1124782

[81] Hung-Yuan Wang, Tzung-Jin Lin, Chin-Chung Tsai, Henry Been-Lirn Duh, and Jye-Chong Liang. 2012. An investigation of students' sequential learning behavioral patterns in mobile CSCL learning systems. In *2012 IEEE 12th International Conference on Advanced Learning Technologies*. IEEE, 53–57.

[82] Christoph Wieser, François Bry, Alexandre Bérard, and Richard Lagrange. 2013. ARTigo: Building an Artwork Search Engine With Games and Higher-Order Latent Semantic Analysis. In *First AAAI Conference on Human Computation and Crowdsourcing*. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7634

[83] Robert E. Wood. 1986. Task complexity: definition of the construct. *Organizational Behavior and Human Decision Processes* 37, 1 (Feb. 1986), 60–82. DOI:http://dx.doi.org/10.1016/0749-5978(86)90044-0

[84] Sara Wylie and Len Albright. 2014. WellWatch: reflections on designing digital media for multi-sited para-ethnography. *Journal of Political Ecology* 21, 1 (2014), 321–348.

[85] Yan Yan, Rómer Rosales, Glenn M. Fung, and Jennifer G. Dy. 2011. Active learning from crowds. In *Proceedings of the 28th International Conference on Machine Learning*. 1161–1168.

[86] Donna Yates. 2018. Crowdsourcing Antiquities Crime Fighting: A Review of GlobalXplorer. *Advances in Archaeological Practice* 6, 2 (2018), 173–178. DOI:http://dx.doi.org/10.1017/aap.2018.8

Appendices

A Volunteer Motivations in Paid Crowdsourcing Marketplaces

Category	Guideline	Example Response
DISASTER	Were interested in disasters.	*"Live in tornado alley. Interesting to see if anything looked like tornado damage."*
DO-MORE	Wanted to do as much as possible or finish the whole set. Also wanted to be thorough or do more than expected.	*"Wanted to finish the whole thing"*
ENJOY	Enjoyed the task, found it interesting and fun.	*"It was fun and interesting."*
EXTERNAL	Mentioned external obligations that did not allow them to continue, such as going to work.	*"Short on time and had to get to work."*
FEELING	Felt that what they did was the right amount or, wanted to do a certain number or, felt ready to move on to survey.	*"I thought I had reached 50 images submitted."*
HELP	Wanted to help the project.	*"I tried to complete as many as possible to contribute to the study."*
INSTRUCT	Said they did what the instructions said or the minimum.	*"because it asked to do at least 10 images"*
LOST-INTEREST	Mentioned getting bored or the task was becoming repetitive.	*"Was enjoying the task and stopped when it became monotonous"*
LOST-TRACK	Lost count of how many images they had rated or forgot to go to the survey.	*"wasn't thinking of the amount until I saw the survey button"*
PAID-MORE	Thought they would get paid more.	*"I completed 10 tasks looking for a bonus."*
REJECTED	Thought their submission would get rejected or not approved, wanted to make sure they got paid.	*"I tried to complete enough images to ensure i will get paid"*
SEE-MORE	Wanted to see more images.	*"Wanted to look at more"*
SKILL	Wanted to get better at the task, or thought they were good at it. Also concerned about accuracy of their work.	*"I felt that I got the general gist of the types of imagery I would see."*
TECH	Mentioned technical reasons.	*"I completed the number of images I did because the software had a slight delay for each image that was loaded."*
UNDERSTAND	Didn't understand the instructions or how many they were supposed to do. Thought the instructions were unclear.	*"I was unclear if I needed to do them all or if just by looking at one I could go on to the survey."*
VALUE	Considered the amount they did appropriate for payment, wanted to do enough work for how much they were getting paid.	*"Well honestly as much as I enjoy aerial images and high res airborne imagery there is only a certain amount of time I'm going to spend for $0.50."*
OTHER	Other reasons.	*"Because I felt the survey would be indepth and the focus of this HIT."*
N/A	Blank responses, numbers, not addressing the question.	*"Observed"*

Table 2: Summary of the categories for workers' text responses, along with exemplar responses, as a result of applying an open coding scheme. Each category also corresponds to a *survey free-response* variable.

Variable	Description
Image Count	The total number of images rated.
Extra Image Count	The number of images rated beyond those required.
Total Time	Total time spent rating images, in seconds; the time between seeing the first image and moving on to the survey. We identified and excluded breaks of more than 5 minutes when rating an image.
More than Required	Whether the worker rated more images than the required amount.
Finished	Whether the worker rated the whole set.
Abandoned	Whether the worker accepted the HIT but abandoned it before completing the survey. Note that for this variable, all workers who accepted the HIT were included.
Agreement	Agreement with consensus. As ground truth ratings for the images were not previously known, we calculated agreement as the percentage of images for which a worker selected the consensus rating.

Table 3: Summary of the *performance* variables, based on workers' performance during the rating part of task.

Variable	Description
Checked-Understand	Whether the "I did not understand the instructions" box was checked.
Checked-Paid-More	Whether the "I thought I would get paid more" box was checked.
Checked-Rejected	Whether the "I thought my submission would not get approved" box was checked.

Table 4: Summary of the *survey checkbox* variables, based on workers' selection of the survey checkboxes.

Variable	RWrk	Fdbk	RWrk × Fdbk	0+N	1+N	10+N	0+P	1+P	10+P
Workers				90	108	102	101	100	101
Image Count	$p < .001$			29	15	17	33	21	15
Extra Image Count	$p < .001$			29	14	7	33	29	5
Total Time	$p < .001$	$p = .011$		201s	127s	148s	252s	188s	156s
More than Required	$p < .001$			100.0%	96.3%	85.3%	100.0%	100.0%	89.1%
Finished	$p = .001$	$p < .001$		4.4%	5.6%	0.9%	13.9%	13.0%	4.9%
Checked-Paid-More			$p = .04$	23.3%	26.0%	13.7%	13.9%	27.0%	23.8%
INSTRUCT	$p < .001$			6.7%	9.3%	19.6%	3.9%	10.0%	23.8%
LOST-INTEREST	$p = .002$			15.6%	9.3%	4.9%	10.9%	4.0%	2.9%
LOST-TRACK		$p = .011$		1.1%	2.8%	3.9%	0.9%	0.0%	0.0%
REJECTED		$p = .011$		1.1%	5.6%	0.9%	0.9%	0.0%	0.0%

Table 5: Summary of results for variables with statistically significant differences from omnibus tests. Values are based on workers who completed the HIT; worker counts for each condition are given in the top row). Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Aligned Rank Transform and Boolean variables with logistic regression.

Variable	`0+ - 1+`	`0+ - 10+`	`1+ - 10+`
Image Count	31 - 19 $p < .001, r = 0.23$	31 - 16 $p < .001, r = 0.25$	
Extra Image Count	31 - 18 $p < .001, r = 0.27$	31 - 6 $p < .001, r = 0.56$	18 - 6 $p < .001, r = 0.33$
Total Time	234s - 153s $p = .002, r = 0.20$	234s - 154s $p = .005, r = 0.18$	
More than Required		100.0% - 87.2% $p < .001, \phi = 0.25$	98.1% - 87.2% $p < .001, \phi = 0.20$
Finished		9.4% - 3.0% $p = .04, \phi = 0.13$	9.1% - 3.0% $p = .047, \phi = 0.12$
INSTRUCT		5.2% - 21.7% $p < .001, \phi = 0.23$	9.6% - 21.7% $p = .004, \phi = 0.16$
LOST-INTEREST		13.1% - 3.9% $p = .006, \phi = 0.16$	

Table 6: Summary of statistically significant post-hoc comparisons for `required work` main effects. *Finished* had no such comparisons. Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Wilcoxon rank-sum test and Boolean variables with the chi-squared test, applying the Bonferroni correction.

Variable	`N - P`
Total Time	151s - 189s $p = .019, r = 0.11$
Finished	3.7% - 10.6% $p = .020, \phi = 0.13$
LOST-TRACK	2.7% - 0.3% $p = .043, \phi = 0.18$
REJECTED	2.7% - 0.3% $p = .043, \phi = 0.18$

Table 7: Summary of statistically significant post-hoc comparisons for `progress feedback` main effects. *Total Time* had no such comparisons. Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Wilcoxon rank-sum test and Boolean variables with the chi-squared test, applying the Bonferroni correction.

B Fixed Scheduling Variety Mechanisms for Crowdsourced Image Labeling

Variable	`AllL - AllM`	`AllL - 4L:1M`	`AllL - 9L:1M`	`AllL - 19L:1M`	`AllL - 29L:1M`
Subtask Count $p < .001$	**20 - 6** $p < .001, r = 0.55$	*20 - 15* $p = .052, r = 0.18$	20 - 22	20 - 17	20 - 20
Label Count $p < .001$	—	**20 - 12** $p < .001, r = 0.29$	20 - 19	20 - 16	20 - 19
Subtask Agreement $p = .002$	**75% - 67%** $p = .042, r = 0.20$	75% - 73%	75% - 76%	75% - 75%	75% - 80%
Label Agreement *n.s.*	—	73% - 75%	73% - 78%	73% - 76%	73% - 80%
Subtask Time $p < .001$	**8s - 26s** $p < .001, r = 0.70$	**8s - 16s** $p < .001, r = 0.51$	**8s - 13s** $p < .001, r = 0.39$	**8s - 11s** $p < .001, r = 0.28$	8s - 8s
Label Time $p = .05$	—	8s - 10s	8s - 9s	8s - 9s	8s - 7s
Total Time $p = .021$	193s - 179s	**193s - 343s** $p = .033, r = 0.19$	**193s - 285s** $p = .021, r = 0.21$	193s - 211s	193s - 205s
Abandonment $p = .014$	**8% - 22%** $p = .007, \phi = 0.03$	8% - 15%	*8% - 18%* $p = .072, \phi = 0.15$	*8% - 18%* $p = .082, \phi = 0.15$	8% - 12%
Understand $p = .036$	6% - 12%	6% - 7%	6% - 4%	6% - 2%	6% - 7%

Table 8: Summary of data and statistical comparisons. p-values given in the first column are for omnibus tests; in other columns, for post-hoc tests. Significant post-hoc comparisons highlighted in bold, borderline significant post-hoc comparisons highlighted in italics. Numerical variables (including *Subtask Agreement* and *Label Agreement*) are given as medians and Boolean variables as percentages.

C Variable Scheduling Variety Mechanisms for Crowdsourced Image Labeling

	Q	G	R
N	55	65	53
# Levels* ($\mathbf{Q–G^{***}}$, $\mathbf{Q–R^{***}}$, $\mathbf{G–R^{***}}$)			
median	9	8	12
# Tiles			
median	363	338	305
# Moves* ($\mathbf{Q–R^{*}}$, $\mathbf{G–R^{*}}$)			
median	195	218	134
Time (s)* (*Q–R*, *G–R*)			
median	992	1057	624
Avg Level Time (s)*** ($\mathbf{Q–G^{*}}$, $\mathbf{Q–R^{***}}$, $\mathbf{G–R^{***}}$)			
median	115	176	62
Fitness* ($\mathbf{Q–R^{*}}$, *G–R*)			
median	388	406	280
mean	376	344	267
Avg Move Length (# Tiles)*** ($\mathbf{Q–G^{***}}$, $\mathbf{Q–R^{***}}$, $\mathbf{G–R^{***}}$)			
median	3.1	2.2	3.9

***:$p < 0.001$; **:$p < 0.01$; *:$p < 0.05$

Table 9: Summary of performance for players per condition in the ordering comparison HIT. The *fitness* metric corresponds to the sum of tiles collected of each difficulty, multiplied by the respective difficulty weight. **Bold** indicates $p < 0.05$ and *italics* indicates $p < 0.1$ for omnibus and post-hoc pairwise tests. (X–Y) indicates comparison between X and Y conditions.

D An Augmented Reality Tabletop Toolkit for Image Labeling

		Game Designers	Sociology	Environmental Health & Justice
Sorting	**Team Play**	✓	✗	✓
	Image Discussion	Some	None	Strong
	Group Engagement	Strong	None	Strong
	Group Votes	✓	✗	✓
	Finished	✓	✓	✗
	Labor	Division	Overlap	Division
	Piles	✓	✗	✓
Memory	**Team Play**	✗	✓	✓
	Image Discussion	None	Strong	Strong
	Group Engagement	None	Some	Strong
	Group Votes	✗	✗	✓
	Finished	✓	✗	✓
	Competition	Strong	Moderate	None
	Opponent Assistance	None	Strong	N/A
TrekStack	**Team Play**	✓	✓	✓
	Image Discussion	Strong	Some	Strong
	Group Engagement	Strong	Some	Strong
	Group Votes	✓	✓	✓
	Finished	✓	✗	✗
	Labor	Division	Overlap	Division
	Strategy	Strong	None	Moderate

Table 10: Comparison of group observations, organized by activity condition. All groups had differing approaches to each game, as well as varying levels of successfully completing each activity. **Bold** indicates categories that were present in all games, while *italics* indicates activity-specific observations.