Gamers, Citizen Scientists, and Data: Exploring Participant Contributions in two Games with a Purpose

Nathan Prestopnik
Department of Computer Science
Ithaca College
Ithaca, NY 14850
nprestopnik@ithaca.edu

Kevin Crowston
School of Information Studies
Syracuse University
Syracuse, NY 13244
crowston@syr.edu

Jun Wang
School of Information Studies
Syracuse University
Syracuse, NY 13244
junwang4@gmail.com

Two key problems for crowd-sourcing systems are motivating contributions from participants and ensuring the quality of these contributions. Games have been suggested as a motivational approach to encourage contribution, but attracting participation through game play rather than intrinsic interest raises concerns about the quality of the contributions provided. These concerns are particularly important in the context of citizen science projects, when the contributions are data to be used for scientific research.

To assess the validity of concerns about the effects of gaming on data quality, we compare the quality of data obtained from two citizen science games, one a "gamified" version of a species classification task and one a fantasy game that used the classification task only as a way to advance in the game play. Surprisingly, though we did observe cheating in the fantasy game, data quality (i.e., classification accuracy) from participants in the two games was not significantly different. As well, data from short-time contributors was also at a usable level of accuracy. Finally, learning did not seem to affect data quality in our context.

These findings suggest that various approaches to gamification can be useful for motivating contributions to citizen science projects.

Key Words: Games, gamification, crowdsourcing, citizen science, data quality

1. INTRODUCTION

In this paper, we examine the interplay of motivation and quality of contribution in the context of crowd sourced systems. Crowd sourcing can be a powerful mechanism for rapidly generating high-quality outputs through distributing work across many different contributors. In this current research, we explore one specific form of crowd sourcing, citizen science.

In citizen science projects, members of the general public are recruited to contribute to scientific investigations. Citizen science initiatives have been undertaken to address a wide variety of goals, including educational outreach, community action, support for conservation or natural resource management, collecting data from the physical environment or analyzing data for research purposes. Many citizen science projects rely on computer systems through which participants undertake scientific data collection or analysis, making them examples of social computing (Cohn, 2008; Wiggins & Crowston, 2011).

Because many participants are not trained scientists and have limited scientific knowledge, a frequent concern about citizen science projects is the quality of the data participants generate (raw or analyzed) and the suitability of this data for the science goals of the project. For citizen science, "data quality" is a complex construct that encompasses validity, reliability, and ultimately, the usefulness of data (Orr, 1998; Pipino, Lee, & Wang, 2002; Prestopnik & Crowston, 2011; Wang & Strong, 1996).

Contrary to these concerns, previous studies have reported favorably on citizen science data quality. For example, Galloway et al. (2006) compared novice field observations to expert observations, finding that observations between the two groups were comparable with only minor differences. Delaney et al. (2008) checked data quality in a marine invasive species project, finding that participants were 95% accurate in their observations. However, their study did find that motivation had an impact on the final data set, with some participants failing to finish because of the tedious nature of the tasks.

This last finding is notable because citizen science projects often rely on the inherent appeal of the topic to attract and motivate participants. For example, "charismatic" sciences like bird watching, astronomy, and conservation all have enthusiastic communities of interest, and a number of successful citizen science projects have grown up around these topics. While the intrinsic motivation of science is undeniably powerful, citizen science projects that rely on this motivation to attract contributions face limits on their available pools of participants, namely those who share the particular scientific interest. Less charismatic topics of inquiry that lack a large natural base of users could therefore benefit from alternative mechanisms for motivating participants.

Purposeful games have the potential to become one such motivational mechanism. Games are recognized for their potential to motivate and engage participants in human computation tasks (e.g. Deterding, Dixon, Khaled, & Nacke, 2011; Law & von Ahn, 2009; McGonigal, 2007, 2011; von Ahn, 2006; von Ahn & Dabbish, 2008) and so seem to offer great potential for increasing the pool of contributors to citizen science projects and their motivation to contribute.

However, in citizen science projects that incorporate games, concerns about data quality are heightened. Designing gamified systems involves creative tradeoffs, where playful interactive elements compete for primacy against outcome objectives. Systems designed to maximize engagement and fun may do so at the cost of reduced data validity, reliability, and usefulness. Players who are engrossed in a game may find themselves concentrating only on the fun elements of a game, ignoring, neglecting, or even cheating on embedded science tasks. On the other hand, games that are designed to prevent such behaviors may improve data quality but impose difficult, boring, or even unpleasant constraints upon their users, making them less fun for players and leaving them unable to attract many participants.

The interrelated issues of game-driven participant engagement and citizen science data quality are of interest to game designers, HCI researchers, and those involved with citizen science. It is important for these various constituencies to understand how citizen scientists produce data using games, how accurate that data can be, how different approaches to "gamification" can influence player motivation and data quality, and innate player attitudes and interests can mediate participation and data quality. In this paper, we address these questions.

2. THEORY: GAMIFICATION AND GAMES WITH A PURPOSE

2.1 Gamification, Diegesis, and Rewards

The goal of most so-called "gamification" is to use certain enjoyable features of games to make non-game activities more fun than they would otherwise be (Deterding, Dixon, et al., 2011; Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011). Often, the term gamification refers to the use of things like badges and points to place a "game layer" on top of real-world activities, especially in corporate, governmental, or educational settings. However, this usage is heavily contested by game designers and scholars, with some going so far as to criticize these approaches as "exploitationware" (Bogost, 2011). As Bogost (2011) and others have pointed out, points, badges, rewards, scores, and ranks do not really engage players, that is, they are not core game mechanics themselves. Rather, these are just metrics by which really meaningful interactions – the play experiences that truly compel and delight players – are measured and progress is recorded. To remove meaningful aspects of play and retain only these measurement devices is to produce something that is not really a game at all (Bogost, 2011; Deterding, Dixon, et al., 2011; Deterding, Sicart, et al., 2011; Salen & Zimmerman, 2004).

To conceptualize different rewards and different approaches to creating games, we distinguish two different kinds of rewards that a game might offer, drawing on the notion of diegesis, a term from the study of film that refers to the notion of the "story world" vs. the "real world" (De Freitas & Oliver, 2006; A. R. Galloway, 2006; Stam, Burgoyne, & Flitterman-Lewis, 1992).

Diegetic rewards in games are those that have meaning within the game but no value outside of it. For example, a diegetic game reward might be an upgraded weapon given to the player by a game character upon finishing a quest. The weapon has meaning in the game: it is more powerful and can be used to slay more dangerous enemies. This reward is strongly tied to the story and the game world and has no use outside of it. In-game money and items are simple examples, but more abstract rewards also qualify as diegetic, including the immersive exploration of a beautiful game world, the enjoyment of a rich game story, the joy of playing with fun game mechanics, or the player's dialogue with game characters or other human players. Malone (1980) has noted how many of these can be motivating in the context of gamified experiences, specifically educational games.

In contrast, non-diegetic rewards are those that have only limited connection to the game world, but sometimes (not always) have meaning in the real life of the person playing the game. For example, "achievements" (a kind of merit badge) are a common non-diegetic reward used in entertainment games. Players can collect achievements by performing certain actions within the game (e.g., "jump from a great height," or "collect 1 million coins"). However, these achievements do not affect subsequent game play. Non-diegetic rewards like badges, points and scores are frequently used in citizen science games to acknowledge player accuracy, time spent, effort, or milestone accomplishments[1]. However, because non-diegetic rewards are only weakly tied to the game world and do not impact the game experience, players are likely to value them only to the extent that they value the actual accomplishments for which they are awarded.

For "science enthusiast" players who truly engage with the scientific elements of citizen science games, non-diegetic rewards might have great significance. However, it is possible that such players do not really need a game to motivate their contributions in the first place. For "non-enthusiast" players, non-diegetic rewards likely have limited appeal. If the real-world science activity itself is not highly valued, non-diegetic rewards for working on it will also not be valued.

Rather than badges or points, non-enthusiast players are most likely to find value in a game that can turn "boring science" into "play." Diegetic rewards can be crafted to be engaging and meaningful even to non-enthusiasts who are not inherently motivated by the task or related non-diegetic rewards. Diegetic rewards focus player attention upon the game story, game world, and game play instead of the real-world task, and can thus become a powerful form of feedback to keep non-enthusiasts immersed in a game that

[1] Examples include exergames like fold.it (http://fold.it), Phylo (http://phylo.cs.mcgill.ca), and Cropland Capture (http://www.geo-wiki.org/games/croplandcapture/), among others.

occasionally asks them to undertake a science task. There is promise in this approach, especially the possibility of attracting and engaging large crowds of non-enthusiast participants.

We have described the mismatch between non-diegetic rewards and motivation in the context of citizen science, but suspect that it applies more broadly. Indeed, many scholars and designers have become disenchanted with the typical connotation of the term “gamification,” finding it laden with inappropriate emphasis on performance metrics like badges and points (i.e., non-diegetic rewards). Many alternatives to the term “gamification” have been proposed: “games with a purpose,” “serious games,” “productivity games,” “persuasive games,” and “meaningful games” (Bogost, 2011; Deterding, Dixon, et al., 2011; Deterding, Sicart, et al., 2011; McGonigal, 2011; Salen & Zimmerman, 2004). These terms describe flexible approaches to gamification where diegetic rewards are common instead of rare, and game designers seek to craft meaning within the game world. In this present study, we adopt von Ahn’s (2006) term “games with a purpose” and its variant, “purposeful games,” to distinguish diegetic reward approaches from non-diegetic “gamification.” In our view, these terms strongly convey the task-oriented nature of citizen science but also emphasize our broad view of games as entertainment media that should focus on engagement, play, meaning, and fun.

2.2 The Challenges and Merits of Comparing Diegetic and Non-Diegetic Experiences

To date, there has been little formalized comparison of diegetic and non-diegetic rewards in gamified experiences, particularly as these relate to player performance and especially to data quality. Outside of a prior version of this work that featured a more limited data set and analysis (Prestopnik, Crowston, & Wang, 2014), to our knowledge there has been no formalized comparison made of citizen science data quality using the same science task as a basis for two very different modes of gameplay.

We posit that different reward structures and philosophies of gamification will impact player experience and subsequent performance independent of the task itself. For example, in most gamified citizen science activities, players are never allowed to stray very far from the tasks they are supposed to be doing. Players earn points and other rewards specifically for engaging with the science, and these data analysis activities comprise the majority of the game experience. Such games inherently place emphasis on the science, providing players with few opportunities or reasons to neglect the work.

On the other hand, our understanding of diegetic rewards suggests an alternative approach whereby players engage with an entertainment-oriented game world that only occasionally requires them to act as a “citizen scientist.” In this approach, the science task becomes just one mechanic among many, and not necessarily the most important or compelling aspect of the game. Though this approach to design could heighten the chances of attracting non-enthusiast players (Prestopnik & Tang, 2015), the concern is that these players will ignore, neglect, or otherwise undermine the quality of science data in lieu of playing other parts of the game. Even cheating – i.e., knowingly submitting bad data – could seem beneficial to players who are fixated on the entertainment experience and so motivated to skip over the science work.

2.3 Exploring of Data Quality: One Task, Two Games

To explore the relation between type of rewards, motivation and data quality, we designed our own citizen science platform. We chose to develop our own system because of the practical difficulties in introducing experimental interventions into an existing system. Most highly successful citizen science games are also specialized, unique, and focused primarily on *using* citizen science to further scientific goals within a specific context (e.g. *Fold.It*[2], *Phylo*[3], or *EyeWire*[4]). Only secondarily are these games viewed as tools to study citizen science per se, usually only after the games have achieved high prominence, a large user base, and operational procedures that can be challenging or risky to disrupt with an experiment.

We designed two very different games around the same purposeful activity in order to study the impact of different approaches to gamification on data[5]. One game, *Happy Match*, adopted a straightforward gamification approach, rewarding players for performance with non-diegetic score points and focusing primarily on the science task. The second, *Forgotten Island*, was entertainment-oriented, a point-and-click science fiction adventure where the science task was integrated alongside many other play mechanics (exploration, puzzle solving, item collection, virtual gardening) and designed as a means for advancing in the game. Rewards in this game were diegetic, and included in-game money as well as the ability to interact with various characters, progressively explore the game world, and advance the game story.

3. RESEARCH QUESTIONS

With our unique environment and overarching scholarly interests in mind, we developed a guiding set of research questions about data quality and participation. First, we wanted to know how our two games would differ in their ability to sustain participation and retain participants. Therefore, we address the question:

RQ1: *How does player retention differ between a gamified task and an entertainment-oriented purposeful game?*

Second, as discussed above, the different reward systems and play experiences offered by our two games raised the concern that data quality (i.e., accuracy) might vary between the two games, despite their nearly identical citizen-science task structure. If one gamification approach does indeed lead to measurably poorer data quality than another, that approach may be unsuitable for many kinds of citizen science tasks. We therefore address the question:

RQ2: *How does the quality of data produced by players differ between a gamified task and an entertainment-oriented purposeful game?*

Third, a common phenomenon in citizen science and many other forms of crowdsourcing is that a few “power” users provide the majority of the work, while a “long tail” of casual participants may provide only a small amount of labor each (Anderson, 2008; Franzoni & Sauermann, 2014). That is, many people may be curious enough to try a new system (the long tail of many participants, with few contributions each), but only a few will find it interesting enough to participate at a high level (the few

[2] http://www.fold.it
[3] http://phylo.cs.mcgill.ca/
[4] http://eyewire.org
[5] http://citizensort.org

power users who make many contributions each). As there are many players in the long tail, the combined number of classifications provided by less motivated individuals can be large. If it takes a long time or much effort for a player to learn the science task well enough to provide quality data, however, then the contributions from the long tail may be scientifically worthless. We therefore addressed a third question:

RQ3: *How is data quality affected by the number of classifications a participant provides?*

Lastly, further considering the differences in contribution between players, we were interested to explore how innate interests in science, nature, and games might impact data quality, play duration, and player retention in gamified tasks and entertainment-oriented purposeful games. For players with high interest in science and nature, a gamified task may have more positive outcomes for some or all of these dependent constructs. For players with a high interest in games or low interest in science or nature, an entertainment-oriented experience may more positively impact data quality, retention, or play duration. Accordingly, we examined a final question:

RQ4: *How are data quality, play duration, and retention related to a participant's initial interest in science, nature, and games?*

4. SYSTEM DEVELOPMENT

The two purposeful games that we designed to address these data and participation-centric questions were focused on a science activity, the taxonomic classification of plants, animals and insects. In sciences such as entomology, botany, and oceanography, experts and enthusiasts routinely collect photographs of living things. When captured with digital cameras or cell phones, photographs can be automatically tagged with time and location data. This information can help scientists to address important research questions, e.g., about wildlife populations or how urban sprawl impacts local ecosystems. Time and location tagged photos are only valuable, however, when the subject of the photograph (the plant, animal, or insect captured) is known and expressed in scientific terms, i.e., by scientific species name. This information is rarely recorded in an accessible fashion when the photograph is captured in the field by amateur enthusiasts and sometimes not even by professional scientists.

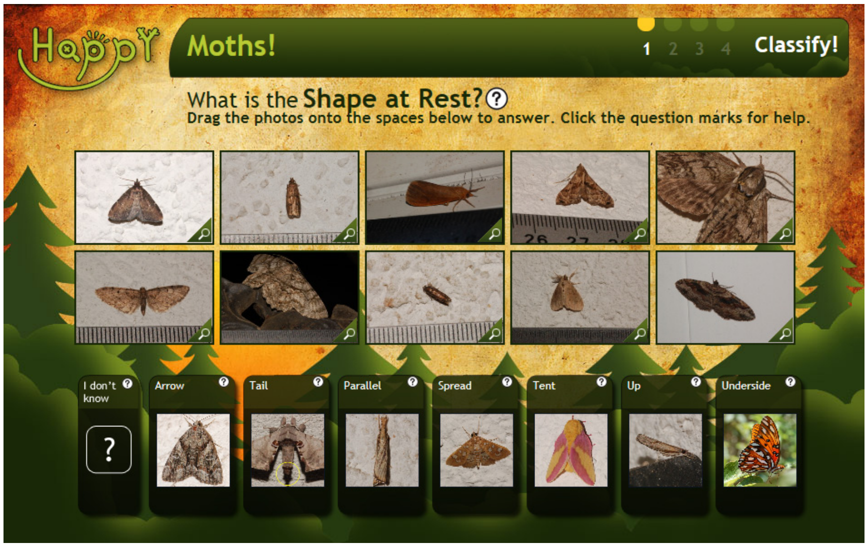
To classify specimens, biologists have developed taxonomic keys that guide the identification of species. These keys are organized around character-state combinations (i.e., attributes and values). For example, a character useful for identifying a moth is its "orbicular spot," with states including, "absent," "dark," "light," etc. Given sufficient characters and states assigned to a single specimen, it is possible to classify to family, genus, and even species.

Working within this area of the life sciences, we developed *Citizen Sort*, an ecosystem of purposeful games designed to let non-scientist members of the public apply taxonomic character and state information to large collections of time and location tagged photographs supplied by experts. We also conceptualized *Citizen Sort* to be a vehicle for HCI researchers to explore the intersecting issues of citizen science data quality and purposeful game design.

4.1 Gamified Task: Happy Match

Citizen Sort features two purposeful games that are the subject of this study. *Happy Match* is a score-based matching game that places the science activity in the foreground of the game, and seeks to attract "enthusiast" players who may already hold some interest in science, classification, or a particular plant, animal, or insect species. It may be considered a form of "gamified task," in that it is very much like an image sorting tool with a non-diegetic, points-based game layer added to it.

Happy Match can be played using photographs of moths, rays, or sharks (*Happy Moths*, *Happy Rays*, and *Happy Sharks* respectively). Players are tasked with earning a high score by answering character-state questions about the photos through a drag-and drop interface, character by character (see Figure 1). Most photos in the game are unclassified; the player is contributing data by answering the questions. Two photos in each round, however, are chosen from the set of photos with known gold standard answers generated by scientists. These two "happy" photos are used to score the game and verify player performance, as well as to calculate a bonus score when players do very well.

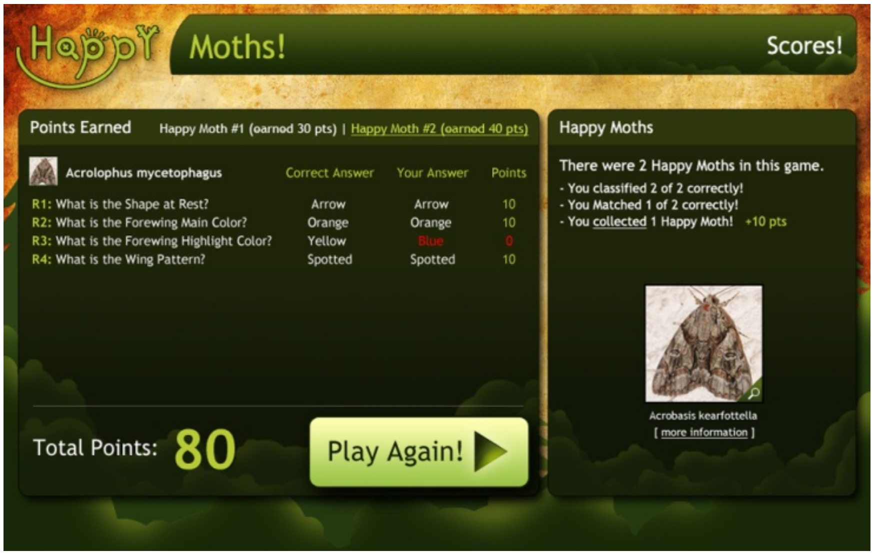


What is the Shape at Rest?

Drag the photos onto the spaces below to answer. Click the question marks for help.

Figure 1. The *Happy Match* classification interface.

At the end of the game, players receive feedback about the correctness of each of the character-state choices for the known “happy” photos and a score based on their performance (Figure 2). Which photos are the “happy” photos are only revealed at the end of the game, so players must strive to perform well on all photos to ensure a good score.



Points Earned Happy Moth #1 (earned 30 pts) | Happy Moth #2 (earned 40 pts)

Acrolophus mycetophagus	Correct Answer	Your Answer	Points
R1: What is the Shape at Rest?	Arrow	Arrow	10
R2: What is the Forewing Main Color?	Orange	Orange	10
R3: What is the Forewing Highlight Color?	Yellow	Blue	0
R4: What is the Wing Pattern?	Spotted	Spotted	10

Total Points: 80

Happy Moths

There were 2 Happy Moths in this game.

- You classified 2 of 2 correctly!
- You Matched 1 of 2 correctly!
- You collected 1 Happy Moth! +10 pts

Acrobasis kearfottella
[more information]

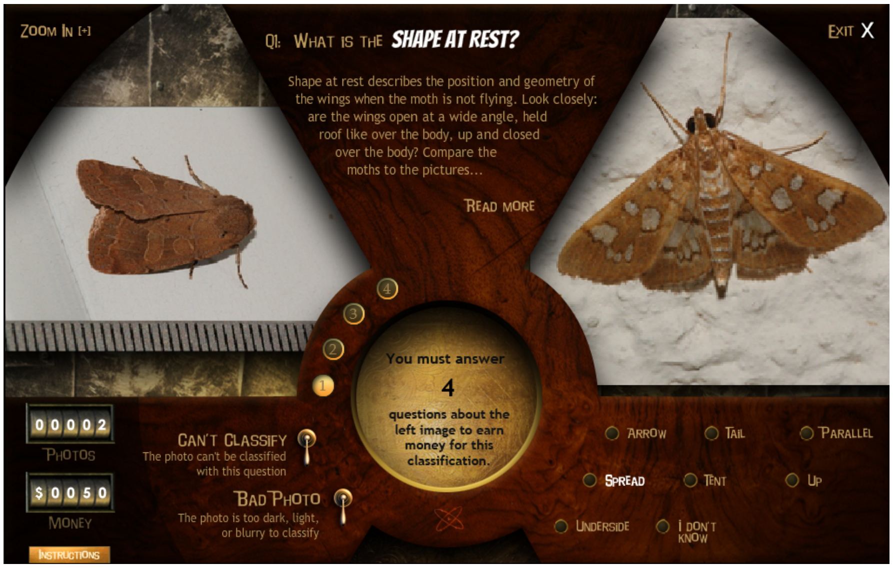
Figure 2. The *Happy Match* score interface.

The final score in *Happy Match* is a non-diegetic reward, a measure of performance and nothing more. It is not used as an input or modifier for future sessions of the game, nor is it connected in any way to story, since *Happy Match* is not a story-driven experience. Outside the game, the *Citizen Sort* website includes a leaderboard system that shows overall player performance, and rankings are based on the score.

Happy Match rewards players with points based on performance. This reward may not be meaningful to all players (Prestopnik & Tang, 2015), but we would argue that *Happy Match* differs from what Bogost (2011) calls “exploitationware” in that it is designed to be a meaningful experience for certain players: science enthusiasts who already have an interest in science, nature, living things, or classification. While *Happy Match’s* non-diegetic points have only limited meaning for players who do not care about these, as with many such games, they are a meaningful performance metric and reward for those who do.

4.2 Entertainment-Oriented Experience: Forgotten Island

The second game, *Forgotten Island*, has players performing the identical classification task found in *Happy Match*. Players classify the same data set as *Happy Match* using the same selection of character and state questions. The same help text, example photographs, and zoom features are available, just as in *Happy Match*. The major difference in classification between *Happy Match* and *Forgotten Island* is that *Forgotten Island* players classify just one photo at a time, rather than batches of ten images. This difference is because, in Forgotten Island, the classification task is connected diegetically to the game world through the use of endogenous fantasy (Figure 3).



Q1: What is the SHAPE AT REST?

Shape at rest describes the position and geometry of the wings when the moth is not flying. Look closely: are the wings open at a wide angle, held roof like over the body, up and closed over the body? Compare the moths to the pictures…

You must answer 4 questions about the left image to earn money for this classification.

Can't Classify: The photo can't be classified with this question

Bad Photo: The photo is too dark, light, or blurry to classify

00002 Photos

$0050 Money

Figure 3. The *Forgotten Island* classification interface.

The classification activity in *Forgotten Island* is situated within an interactive point-and-click adventure story set in a vibrant, science fiction game world (Figure 4). Players explore this world by walking through various levels and locations, looking for equipment – one form of diegetic reward – that will help them to solve puzzles, advance the story, and open up new game spaces to explore. All the while, photos flutter from the sky, the result of a lab explosion, the inciting incident of the narrative. Players collect these photos and are rewarded with in-game money (another diegetic reward) for each classification they complete.

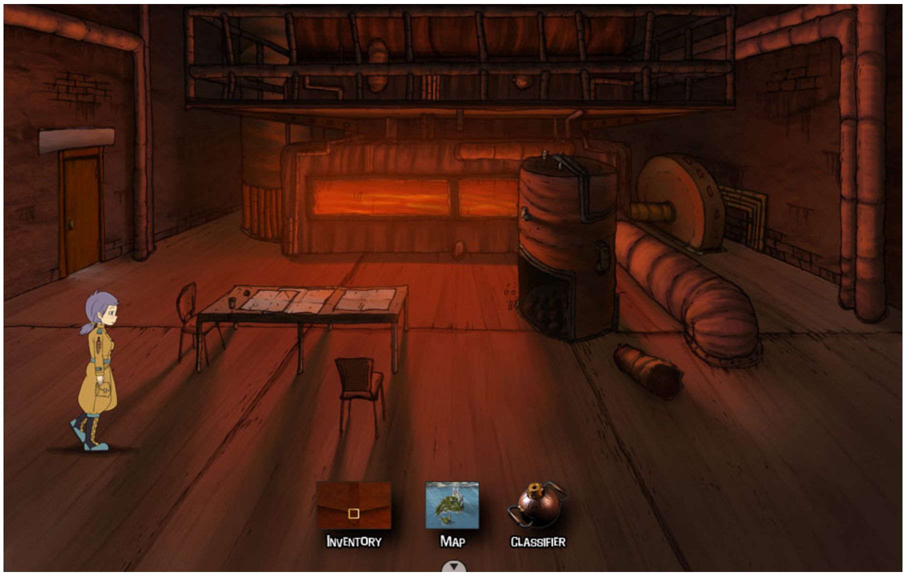


Figure 4. The "Abandoned Pump House" location from *Forgotten Island*.

Money can be spent in some levels to acquire additional equipment and items. To motivate effort, players are periodically given known photos to classify. Incorrect answers for a known photo result in a warnings and a slight penalty, a deduction of the player's in-game money. Both the warning and penalty are also diegetic, issued in bombastic fashion by the game's primary antagonist.

The *Forgotten Island* game experience – the game world and the story – is designed to be a form of continuous diegetic reward as it unfolds, as are (more concretely) the in-game money and equipment earned by players. All of these things have only limited meaning outside of the game, but can be important to players within the context of *Forgotten Island*.

Our intention in developing these two games was to explore some of the relative advantages and disadvantages of the two approaches. Scientists who envision purposeful games as an aspect of their crowdsourced scientific data collection or analysis activities need to understand how different game experiences lead to different player behaviors, as well as (potentially) different data outcomes.

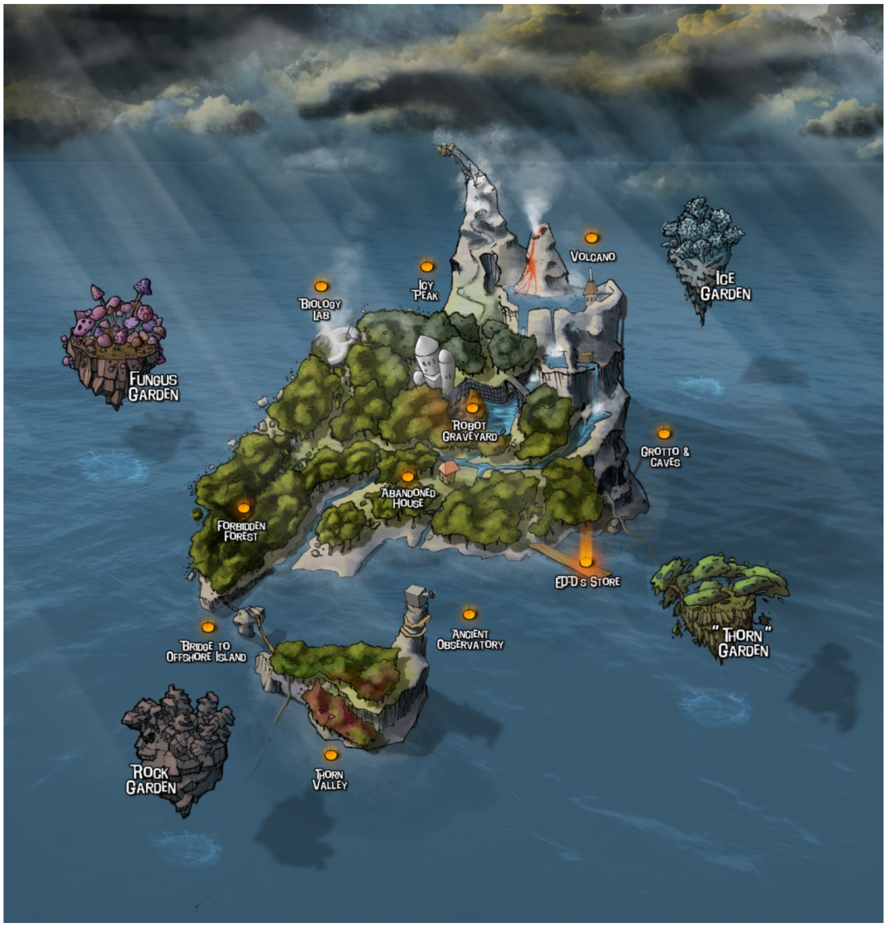


Figure 5. The *Forgotten Island* game world.

5. METHOD

To explore our research questions regarding motivation and data quality in the classification activity, we drew upon data generated by players of *Forgotten Island* and *Happy Match* who played using photos of moths (since this is the only dataset currently used in both games). For some additional analysis, we also drew upon data from other versions of *Happy Match* that used photos from different datasets (*Happy Rays* and *Happy Sharks*).

Participants were recruited naturalistically online beginning in October 2012. They learned about the project and the games from news posts, comments, and listings that appeared on various citizen science websites and in science publications such as *National Geographic* and *Scientific American*. The *Citizen Sort* home page gives equal emphasis to both *Happy Match* and *Forgotten* Island, and players were free to try either game (or both). Note, however, that most of our online outreach efforts targeted venues geared towards promoting citizen science, so our sample in this research may be biased toward individuals with some proclivity for participating in citizen science projects. This stands in contrast to a complementary study about this project (Prestopnik & Tang, 2015), that used a controlled experimental design and drew from a participant pool of computer science students who self-identified as "gamers" more than as citizen scientists.

The mean age of recruited participants was 32 (median = 29; mode = 23). Among younger players (below age17), 73% (349 out of 476) played *Forgotten Island* and 47% (223 out of 476) played *Happy Match*. Among players older than 17, 55% (1971 out of 3575) played *Forgotten Island* and 64% (2278 out of 3575) played *Happy Match*. Based on feedback left during the sign-up process, about 25% of our players were teachers or students. Approximately 8.3% of players verified their email but never started any game.

To date, *Citizen Sort* has 4554 registered users. The data presented in this paper is drawn from the 4174 user accounts of users who opened *Happy Match* or *Forgotten Island* at least once (e.g. users who did more with the system than simply create an account). This group of users excludes developer accounts and 879 temporary player accounts (anonymous, one-time use accounts created for players who are 13 years of age or younger whom we did not track for research purposes).

Relying on data from naturalistic participation has advantages and disadvantages for our study. The main disadvantage is the lack of control: we cannot say if the differences we observe between the two games are due to differences in the features of the games or to difference in the participants who choose to play the games or (most likely) some combination. However, this confounding of game and players is simultaneously a feature of our study: practitioners attempting to deploy such systems would also be constrained by the characteristics of the audiences attracted. Put alternately, we conceptualize the comparison we are drawing as between socio-technical systems that comprise both the games themselves and the specific kinds of players they attract.

6. FINDINGS

We ran a variety of tests on *Citizen Sort's* classification and player data. We used data from the "moths" version of *Happy Match* for all direct comparisons between the games, but include additional supporting data from *Happy Rays* and *Happy Sharks* where appropriate. In this section, we present the results of this analysis, organized by the research question addressed.

6.1 Player Retention

RQ1: *How does player retention differ between a gamified task (Happy Match) and an entertainment-oriented purposeful game (Forgotten Island)?*

To address this question, we compared the retention of players for *Happy Match* and *Forgotten Island.* Retention was measured as how many days a player visited a game and made contributions. The distribution of player visiting days was highly skewed: most players only played the game for one day, 88.6% (n=1571) for *Happy Matc*h and 72.2% (n=840) for *Forgotten Island*, but a few "power" players played for many days. Because the data are not normally distributed, we used the non-parametric Wilcoxon rank sum test to compare retention between the two games. We found a significant difference between the two games (p = 0.0001).

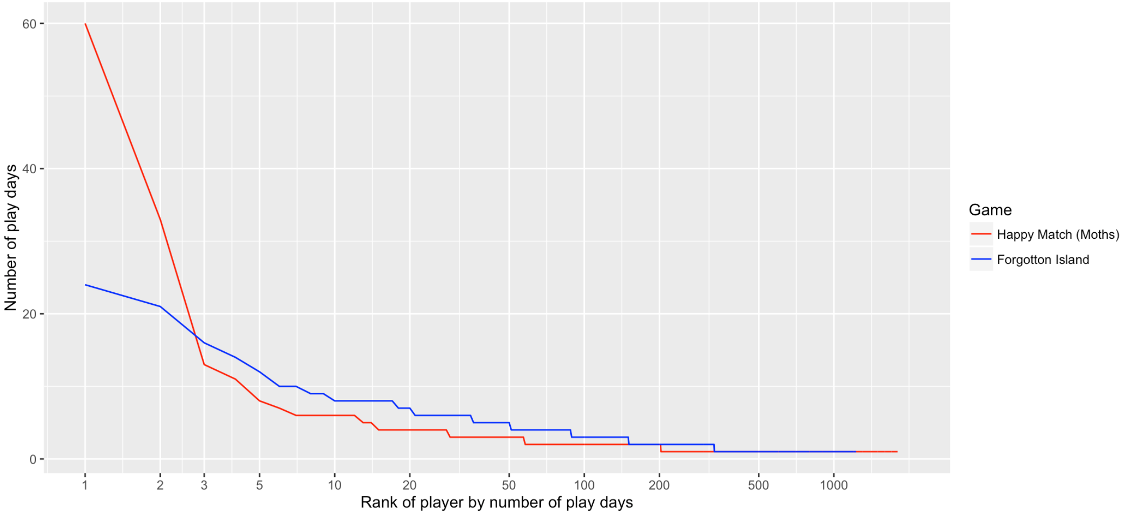


Figure 6. For both *Happy Match* and *Forgotten Island*, players were ranked by the number of days they played each game (rank #1 is the player who played for the most days). The graph above shows the ranked players in order with the number of days played. A small number of *Happy Match* power users played for significantly more days (~60) than *Forgotten Island* power users (~24), though the distribution of days played is otherwise similar between the two games. Note that *Forgotten Island* is a story-driven game, so players naturally would stop playing once the story has been completed. The amount of days from start to finish would depend entirely on the player and their actions within the game.

Figure 7 shows the distribution of the number of scientific contributions (i.e., classification decisions) in the two games. Note that the games show similar distributions, with most players producing between 20 and 500 decisions.

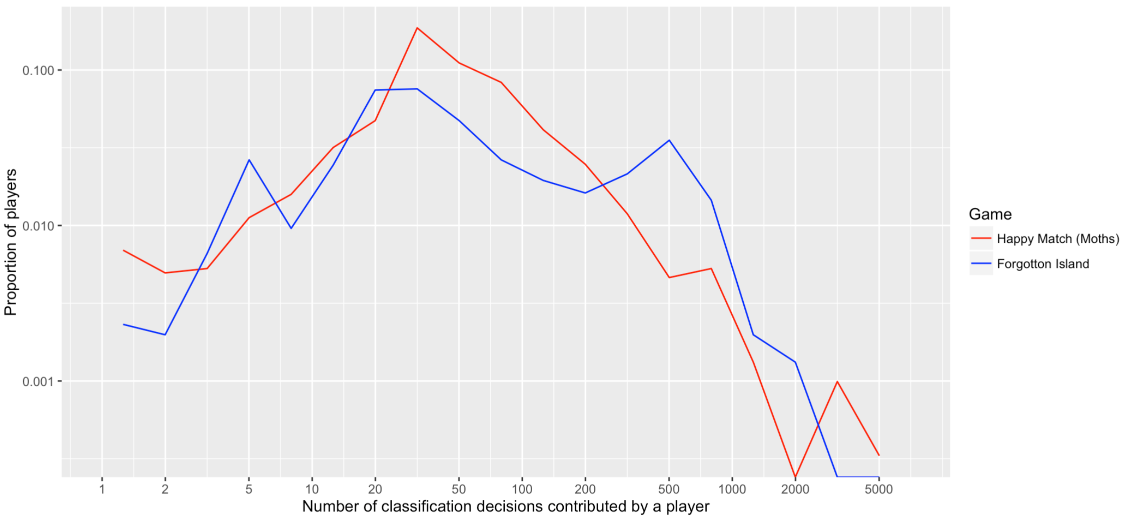


Figure 7. Distribution of decisions contributed by *Happy Match* and *Forgotten Island* players.

On the other hand, the retention differences between *Happy Moths* and *Forgotten Island* (as well as *Happy Rays* and *Happy Sharks*) are apparent in Table 1, which compares the percentage of retained players after just one classification decision, after 20 decisions, and after 50 decisions. Similar to many online systems, the games see a high initial attrition: when players try a game for the first time, most quickly lose interest and do not return. Attrition for *Happy Moths, Happy Rays,* and *Happy Sharks* appears to continue at a steady rate, with only a small core set of "power" players continuing to contribute regularly.

	# of Players	Retained at 1 Decision	Retained at 20 Decisions	Retained at 50 Decisions
Forgotten Island	2407	48% (n=1155)	34% (n=818)	18% (n=433)
Happy Moths	1912	92% (n=1759)	78% (n=1491)	29% (n=554)
Happy Rays	635	95% (n=603)	85% (n=540)	46% (n=292)
Happy Sharks	937	91% (n=853)	70% (n=656)	32% (n=300)

Table 1. Percent of players retained by number of decision made.

In *Forgotten Island*, the rate of attrition seems to fall off after a large initial loss, and overall the game retains fewer of its initial users. Note that players do not make their first classification decision until some way into *Forgotten Island*, and a large percentage (~50%, n=1200) of players leave the game before that point. It may be that *Forgotten Island's* story and world are uninteresting to the players in our online sample, or they may be leaving for other reasons.

This outcome is in contrast to qualitative feedback collected about *Happy Match* (moths) and *Forgotten Island* in a controlled study (Prestopnik & Tang, 2015), in which participants (n=29) greatly preferred *Forgotten Island* because of its emphasis on story, fantasy, and diegetic (story-motivated) rewards. Participants in this controlled study self-identified as "gamers," while the participants in our current research heard about *Citizen Sort* primarily from citizen science venues, suggesting that different styles of game will indeed be appropriate for different target participants.

Note also that unlike *Happy Match*, *Forgotten Island* can be "won" and its story eventually concludes. It takes about 320 classification decisions to win *Forgotten Island*, whereas players can play *Happy Match* indefinitely. Rephrased, *Forgotten Island* has a built in retention "threshold," beyond which players are likely to abandon the system because they have finished its play-oriented aspects.

6.2 Data Quality and Activity Type

RQ2: *How does the quality of data produced by players differ between a gamified task and an entertainment-oriented purposeful game?*

We expected that *Happy Match* players would show better data quality than *Forgotten Island* players because *Happy Match* was designed to be classification task-focused and *Forgotten Island* was entertainment and adventure-focused, with the science task as a side element of the game. To test the difference between the two games, we compared classification accuracy for players of *Forgotten Island* to *Happy Moths*, again only using versions of the games that drew from our moth photo dataset.

We computed accuracy by comparing players' answers for pictures that had a known correct answer. To increase the pool of classifications for the comparison, we ran the game for some time using only pictures for which we already knew the species of moth represented. However, there is not a one-to-one mapping from species to state (e.g., individuals of a particular species can be different colors). We counted as correct any of the possible answers, which inflated the computed accuracy.

We restricted the sample to people who had done a minimum of 20 classification decisions on moths (equivalent to 5 photos, since classifying each photo requires 4 decisions). Our comparison was conducted using a two-sample t-test. Comparing all of these players, there was a slight difference in accuracy (about 2 percentage points in favor of Happy Match). Though not of much practical importance, the difference is statistically significant due to the large sample size.

	Sample size	Accuracy	Sample size for accuracy at least 0.6	Accuracy
Happy Match (Moths)	1500 players	0.792	1464	0.799
Forgotten Island	842 players	0.775	769	0.805
		p-value=0.0003		p-value =0.051

Table 2. Comparing classification accuracy.

On further examination, the slightly lower accuracy of data provided in Forgotten Island seemed to be due to a somewhat larger fraction of users in *Forgotten Island* providing data of low accuracy. Dropping players with low accuracy (< 0.6), there was no significant difference in the accuracy of the data provided by *Happy Match* and *Forgotten Island* players, as shown in the right of Table 2; and it is *Forgotten Island* that improves.

Specifically, in *Forgotten Island* we were able to identify a number of instances of "cheating" behavior. These were recognized when we compared the mean time spent by an individual player on single instances of the classification task (making a decision about character and state) and the overall accuracy of his or her classifications.

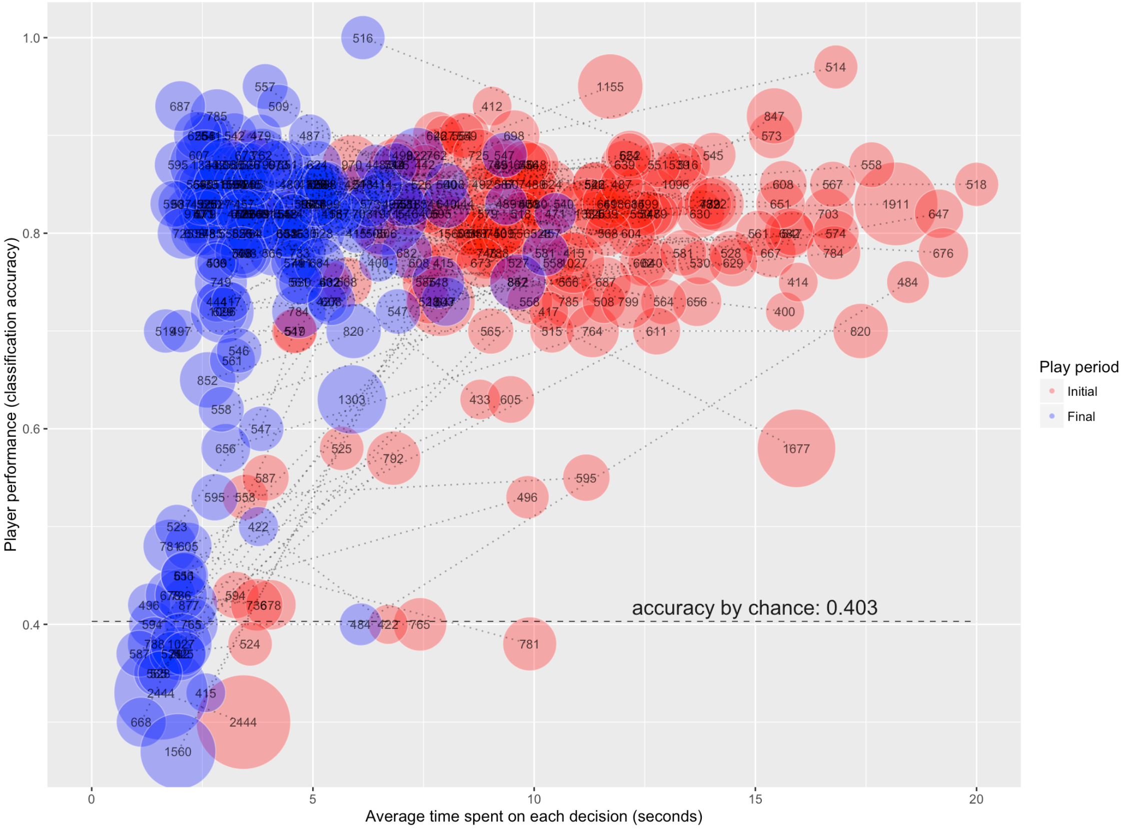


Figure 8. Plot of classification accuracy vs. response time in *Forgotten Island*.

Figure 8 plots the average response time against average accuracy for 159 individual *Forgotten Island* players who contributed at least 400 decisions (100 photos) each. Red circles represent the time and accuracy for a player's first 80 decisions (i.e., for 20 photos). Blue circles represent the data for all photos for a player.

Cheaters leave a distinct signature in Figure 8: very rapid decision making with low accuracy (at the level of chance). Neither low accuracy nor rapid decision making were, by themselves, indicators of cheating. "Power" players who were deeply invested in either *Forgotten Island* or *Happy Match* often became proficient enough to rapidly make accurate classification decisions (upper left of Figure 8), while other players simply struggled with the classification task and did poorly. However, fast classifications coupled with poor accuracy seemed to indicate the profile of a player uninterested in doing well at classification.

Specifically, the blue circles in the lower left of Figure 8 represent players whose performance decreased to the level of chance as their response time per question also decreased, which we interpret as evidence of cheating. Figure 9 plots response time against performance only for those individuals whose performance so decreased, providing a clearer view of this data signature. Figure 10, showing the distribution of answers selected, shows that cheaters habitually select only the first of seven answer choices, while other players show the expected mix of answer choices.

In summary, somewhat to our surprise, the accuracy of data provided in the two games was nearly the same, with differences explained by a small number of *Forgotten Island* players who appear to stop attempting to provide correct answers.

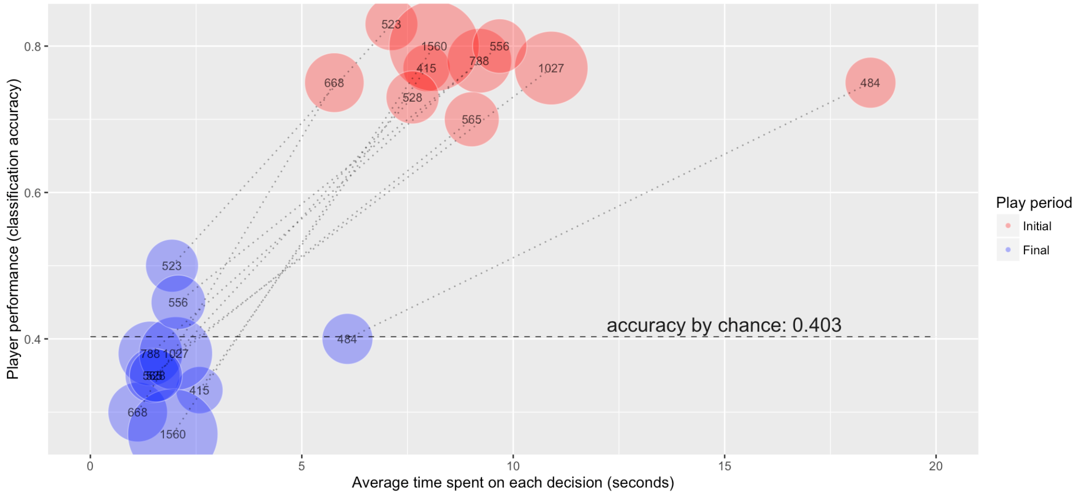


Figure 9. Plot of classification accuracy vs. response time in *Forgotten Island* for participants identified as "cheaters" due to their decreasing performance over time.

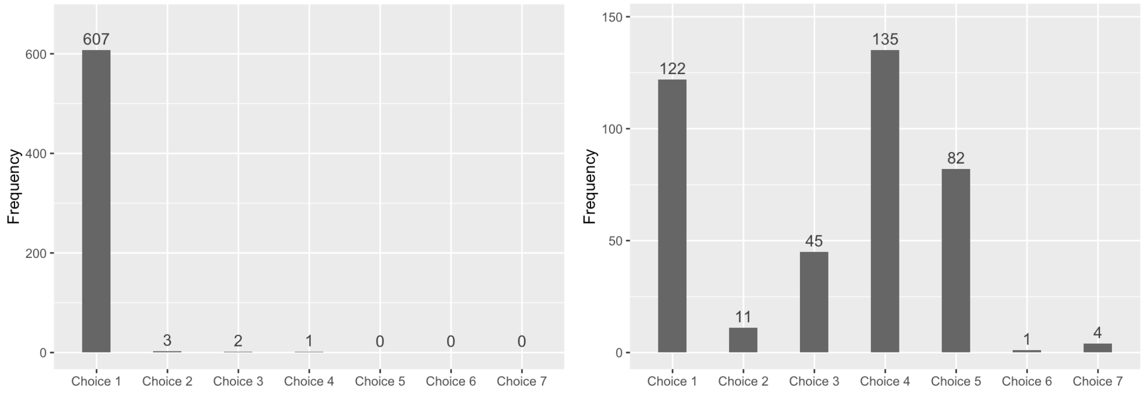


Figure 10. "Cheater" answer choices (left) vs. "normal" distribution of answer choices (right).

6.3 Data Quality and Contribution Amount

RQ3: *How is data quality affected by the number of classifications a player provides?*

As expected, the number of classifications made by players of *Happy Match* and *Forgotten Island* exhibits a highly skewed "long tail." In *Happy Moths,* just 10% of players (n=191) contributed half of the decisions. On the other hand, 67% of players (n=1281) played only one game (including those who did not finish), 33% (n=631) played at least two games, 17% (n=325) played at least three games, 11% (n=210) played at least four games, and only 8% (n=153) played at least five games. We expected that players would need some time to learn the game and the characters and states and so the accuracy of the data they contributed would improve over time. However, if it takes some time to learn to contribute accurately, then the data contributed in the early games (which is the majority of contributions) might not be usable.

We examined this learning effect by comparing the accuracy of a player's first game to their accuracy in later games, as shown in Table 3 for the *Happy Match* games. We used a matched-sample t-test to determine statistical significance of these comparisons. The results show that a player's accuracy stays nearly the same for all three games. In one, *Happy Sharks*, the increase (about 2.5%) is statistically significant, though the difference is of limited practical importance.

	N (sample size)	Accuracy of first game (s.d.)	Accuracy of later games (s.d.)	t (p value)
Happy Moths	347	0.802 (0.09)	0.796 (0.07)	–1.505 (0.133)
Happy Rays	231	0.793 (0.10)	0.806 (0.07)	1.829 (0.069)
Happy Sharks	180	0.629 (0.13)	0.654 (0.12)	3.072 (0.002)

Table 3. Accuracy in first game vs. accuracy in later games for those who have played multiple games. Happy Sharks shows reduced accuracy compared to the other games. We hypothesize that this may be because of the nature of the photos for this game, which sometimes show only portions of the animal in question, making some questions more difficult to answer.

We also examined the possibility of population effects, that is, that people who play only one game (a large fraction of contributors) are less accurate than those who continue playing. Table 4 shows the accuracy in the first game for those who continue vs. those who do not. We used an independent-sample t-test to determine statistical significance. Again, the average accuracy of players in the first game is nearly the same, regardless of whether they continue or not. In one, *Happy Moths*, the difference (about 1%) is statistically significant, though again of limited practical importance.

	N	Accuracy of first game (s.d.)	N	Accuracy of only game (s.d.)	t (p value)
Happy Moths	347	0.802 (0.09)	907	0.790 (0.10)	2.099 (0.036)
Happy Rays	231	0.793 (0.10)	211	0.780 (0.10)	1.394 (0.164)
Happy Sharks	180	0.629 (0.13)	333	0.635 (0.14)	–0.480 (0.632)

Table 4. Accuracy in first game vs. accuracy in later games for those who have played multiple games.

Figures 11 and 12 show the learning effect in *Forgotten Island* and *Happy Match*, plotting player accuracy against total number of decisions. For each game, there is little change from players who contributed less to players who contributed a great deal. (NB. the lines of points apparent on the left side of the graphs are an artifact of the fact that only a few accuracy levels are numerically possible for participants who have made only a few contributions, e.g., 15 correct out of 20, 16 out of 21, etc.).

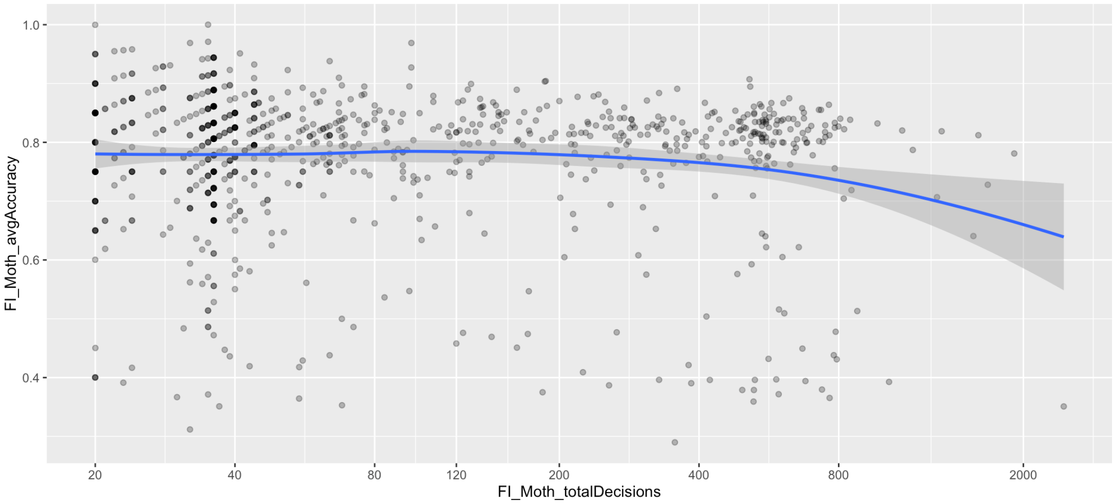


Figure 11. Average accuracy plotted against number of decisions made in *Forgotten Island*.

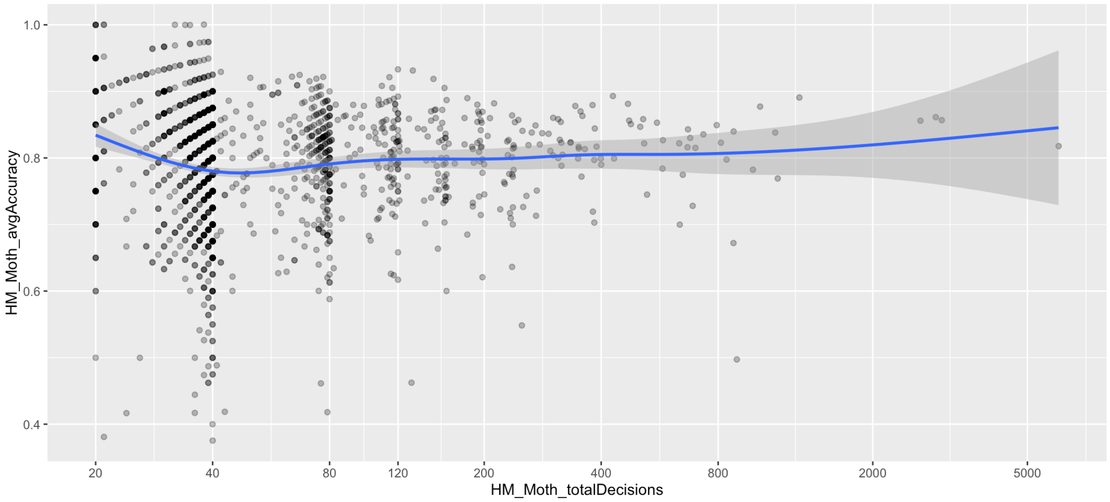


Figure 12. Average accuracy plotted against number of decisions made in *Happy Moths*.

6.4 Participant Initial Interest

RQ4: *How are data quality, play duration, and retention affected by a participant's initial interest in science, nature, and games?*

Finally, RQ4 asks about the relationships between initial interest and various gameplay and task outcomes. To address this question, we drew upon data from a brief questionnaire issued to all new users of *Citizen Sort*. When participants first joined the system, we asked about their current interest in science, in nature, and in games. The questionnaire used Likert-style answers and contained one question for each of these topics. Not all participants answered all of the initial questions.

To assess the relationship between interests and outcomes, we examined correlations. Because some of the data (e.g., play duration) were highly skewed, we used a non-parametric measure, Spearman correlation, ρ. As we have many correlations, we applied a Bonferroni correction to the critical value and so declared a relationship significant only if the p value is 0.001 or less. As in the rest of the paper, we focus on the comparison between *Forgotten Island* and *Happy Moths*, reporting on the other two matching games in passing.

	N (sample size)*	ρ (p value) Science	ρ (p value) Nature	ρ (p value) Games
Forgotten Island	655–663	0.09 (0.018)	0.13 (0.001)	0.02 (0.685)
Happy Moths	1267–1275	0.19 (0.000)	0.10 (0.000)	–0.04 (0.155)
Happy Rays	468–471	0.10 (0.027)	0.05 (0.327)	0.05 (0.282)
Happy Sharks	551–556	0.32 (0.000)	0.11 (0.007)	0.05 (0.241)

Table 5. Correlation between self-reported initial interest in science, nature or games and data accuracy.

* Note: number of responses varies by question

For *Happy Moth* and *Happy Sharks*, we found small but significant correlations between reported interest in science and data quality (as shown in Table 4). The correlations for *Forgotten Island* and *Happy Rays* are smaller and with the Bonferroni correction fail to be significant. The correlations between reported interest in nature and data quality are also small, but significant except for *Happy Rays*. On the other hand, as might be expected, we did not find a significant correlation between interest in games and data quality. These data suggest that an interest in science and especially nature do motivate better accuracy (perhaps by encouraging closer attention to the science task), even in *Forgotten Island*, though the effect is small.

	N (sample size)*	ρ (p value) Science	ρ (p value) Nature	ρ (p value) Games
Forgotten Island	655–663	–0.03 (0.468)	0.07 (0.071)	–0.01 (0.759)
Happy Moths	1267–1275	–0.05 (0.097)	0.03 (0.309)	0.02 (0.442)
Happy Rays	468–471	–0.09 (0.050)	–0.05 (0.270)	–0.00 (0.964)
Happy Sharks	551–556	–0.15 (0.000)	0.01 (0.785)	–0.06 (0.159)

Table 6. Correlation between self-reported initial interest in science, nature or games and decisions contributed.

* Note: number of responses varies by question

	N (sample size)*	ρ (p value) Science	ρ (p value) Nature	ρ (p value) Games
Forgotten Island	655–663	–0.03 (0.464)	0.12 (0.002)	0.02 (0.699)
Happy Moths	1267–1275	0.02 (0.484)	0.04 (0.138)	–0.02 (0.485)
Happy Rays	468–471	–0.07 (0.134)	–0.03 (0.52)	0.01 (0.831)
Happy Sharks	551–556	0.03 (0.462)	0.10 (0.018)	0.00 (0.994)

Table 7. Correlation between self-reported initial interest in science, nature or games and days played.

* Note: number of responses varies by question

Surprisingly there were almost no significant correlations after the Bonferroni correction between any of the interest variables and number of contributions (Table 5) or days played (Table 6). Indeed, the correlation between interest in science and number of contributions was negative across all projects, though significantly so only in *Happy Sharks*. This outcome is surprising as we expected participation in *Happy Match* to be driven by intrinsic interest in the topic. It could be that the task on offer in the system simply did not have enough science content to hold interest for those expecting it.

Even more surprisingly, for *Forgotten Island*, an interest in nature (but not science) had a small but nearly significant correlation with days played (0.12, p=0.002). This correlation suggests that even with the more developed game aspects in this version, an interest in nature was a part of the attraction, even if only a small part. Finally, a reported interest in games was not correlated with either contributions or days played for any of the projects, *Forgotten Island* in particular. The lack of a correlation suggests that none of the versions was more appealing to gamers, which is surprising for *Forgotten Island* especially.

7. DISCUSSION

The most interesting findings from the comparison above were the contrasts we noted between our earlier research on play experience and this current research on data quality, the overall similarity in player performance between *Forgotten Island* and *Happy Match* (with the notable exception of cheating), and the lack of a learning effect.

7.1 Contrasting two *Citizen Sort* Studies

Earlier work focused upon self-identified "gamers" found a strong participant preference for *Forgotten Island* (Prestopnik & Tang, 2015). This work was conducted under controlled, laboratory conditions, and the key reasons cited for this preference included *Forgotten Island's* story and world, story-based incentives, sensory stimulation through art and sound, and feelings of enhanced control and agency compared to *Happy Moths*. In short, the gamer participants in that study appreciated the overall play experience of *Forgotten Island,* and considered it to be, as one player stated, "an entire virtual world that I become invested in as I directed my avatar which way to go and what tasks to accomplish."

In our current research on data quality, we noted that a) preference for *Forgotten Island* does not seem to extend to our online participants in the same way that it did for our experimental participants, and b) whatever preference for *Forgotten Island* there might be does not impact play duration or data quality. In real-world use, *Forgotten Island* actually retains a smaller proportion of participants than *Happy Match*. In addition, even players who expressed an initial interest in gaming when filling out the *Citizen Sort* sign-up questionnaire did not play *Forgotten Island* for any longer or contribute any more data than those who did not.

One possible reason for this discrepancy could be differences in the participant demographics between the two studies. The controlled study drew upon undergraduate computer science students who, as a group, expressed a strong interest in playing video games. The online user base for the current study is larger and mixed, but was recruited primarily from online citizen science venues such as SciStarter[6], Scientific American[7], and the like. This recruitment process has likely resulted in an online participant base that is more interested in science or nature than in games, per se. Our sign-up questionnaire data seems to bear this out. For example, approximately 5% of our users self-identified as teachers, and another 19% identified as students.

An additional important difference between *Forgotten Island* and *Happy Match* is the type of play that these games encourage and, subsequently, the kinds of players they attract and retain. Bartle (1996) proposed a taxonomy of player types, including “killers,” “socializers,” “achievers,” and “explorers.” None of the *Citizen Sort* games were designed for “killers” or “socializers,” but *Forgotten Island* was designed specifically with “explorer” players in mind, and *Happy Match* was designed to attract “achievers.” The play experience in *Forgotten Island* is organized around unlocking new areas, finding interesting locations, solving puzzles, and acquiring information, all aspects of exploration-driven play. *Happy Match* is bound to a leaderboard, located at the *Citizen Sort* website, where players can compete for primacy amongst fellow players based on their score in the game.

Browser games tend to be small, casual, and quick, focused on simple experiences, and so the web browser may not be an ideal environment to capture explorer players for a game like *Forgotten Island*, which is somewhat lengthy (~5 hours to complete), has a complex narrative, and may require several sessions of play to finish. That is, *Forgotten Island* may succeed at attracting “gamer” players, but be poorly positioned to attract the right kind of gamers: explorers. Publishing *Forgotten Island* in a different, gamer-oriented venue (e.g. the Steam marketplace) might result in very different aggregate play patterns.

Given the noted differences in preference, we consider there to be more work to do in this area, especially building and studying citizen science (or other purposeful) games targeted explicitly to gamers, especially gamers of different types. The strong preference for *Forgotten Island* shown in one study suggests that it might be possible to recruit large numbers of non-science enthusiasts to many different kinds of projects with the right balance of play, story, aesthetics, and tasks. On the other hand, this gamer-intensive approach could be less interesting to science enthusiasts, as seems to be the case with *Forgotten Island* from clues in our online data. That is, a strong emphasis on stories

[6] http://scistarter.com/project/689-Citizen Sort
[7] https://www.scientificamerican.com/citizen-science/citizen-sort/

and play might deter science enthusiasts, but a strong emphasis on science, could have a similar effect on gamers. More work remains to tease out the possibilities and limitations of this unique approach to facilitating scientific work.

7.2 Cheating Behavior

A corollary to differences in preference for our games was the difference in player behaviors, especially in *Forgotten Island*. In our exploration of RQ2, we noted the presence of cheating behavior in *Forgotten Island*, a finding that underscores how non-diegetic and diegetic reward systems can have different impacts on player behaviors and data quality.

Why should players cheat in a science game like *Forgotten Island*, and why didn't we spot cheating in *Happy Match*? We hypothesize that the different approaches to play in these games attract different kinds of players with different reasons for playing and different incentives for making progress in the games. *Happy Match* foregrounds the science activity, turning the task itself into a form of play. Additionally, cheating in *Happy Match* will result only in a low score, with no additional offsetting benefit. Players with an inherent interest in the science task should be uninterested in achieving score-based rewards without also achieving some meaningful science experience. To cheat would be pointless for these players, since cheating would be in direct conflict with their science-oriented reasons for playing. For players with less inherent interest in the task, neither the points nor the game experience are worth the effort of cheating. These players will simply stop playing *Happy Match* rather than finding ways to cheat.

Contrariwise, *Forgotten Island* has built-in incentives that make cheating potentially beneficial to some players, and the possibility of these perverse incentives is a real risk of designing gamer-oriented games with a purpose. The diegetic reward system in *Forgotten Island* connects classification activity to in-game rewards like game money, new areas to explore, new puzzles to solve, and new story elements to engage with. For players who enjoy the game but not the science, cheating has real advantages: it speeds up the game, makes it less cognitively demanding, and allows players to focus on the diegetic rewards – game money, the game world, and the story – rather than the sometimes tedious science work required to progress the experience. Players risk a small in-game penalty if they are caught cheating, but in the current system, the risk and the penalty are both low. As a result, cheating can be an attractive proposition for players who realize that they can still make enough money to play through the game even when intentionally doing poorly in the classification task.

It is possible to adjust the classifier in *Forgotten Island* to discourage cheating more strongly. Players who answer incorrectly on known photos could be punished to the point that making money would be impossible without carefully attending to the classification task. However, feedback collected during play tests and other evaluation exercises for both *Happy Match* and *Forgotten Island* (Prestopnik & Crowston, 2011, 2012; Prestopnik & Tang, 2015) suggest that species classification is inherently difficult to do well, and that many honest players struggle to do a good job. Configuring *Forgotten Island* to make cheating impossible would also punish those who honestly err on the task, and could easily render the game too difficult or unpleasant to play.

Overall, there was little practical difference in performance between *Happy Match* and *Forgotten Island* when comparing players who made a minimum of 20 classification decisions. That is, the level of cheating was not high enough to significantly affect the overall results. Furthermore, cheaters leave a clear signal, meaning that their data are easy to eliminate. This finding suggests that both diegetic and non-diegetic reward systems can be viable for citizen science human computation tasks. However, precautions should be taken to identify and exclude data from cheaters or outliers who may be more interested in the game's entertainment experience than its science, e.g., by including a few known items to detect poorly performing players and omitting their data from analysis. It might also be possible to target punishment more precisely if cheating is detected. The need for such measures may be especially high or games or projects that specifically target "gamer" demographics as a user base, since these players are likely to be far more interested in play than in science (or whatever other) tasks.

7.3 Learning Effects and the Long Tail

We found limited evidence for learning effects in either *Happy Match* or *Forgotten Island*, and what learning there was, was minimal. This finding is interesting, unexpected and useful. Many citizen science initiatives heavily rely on power players to provide the majority of data. In our exploration of player behaviors, we noticed this division of labor as well, with 4.4% of players contributing 50% of the classification decisions in the *Citizen Sort* system. These "power players" provide the bulk of scientific data and so are critical to the success of the project. In other settings though, value can come also from the lower volume mass. For example, Anderson (2008) espoused the value of the "long tail" in the context of online marketplaces. Though most items in a market may sell only a few units each, the cumulative sales of the tail can be comparable to the fewer best-selling items that seem at first to be more lucrative. Similarly, 50% of classification decisions in *Citizen Sort* came from what we dub long-tail players. However, verifying that long-tail classifications are as accurate as power-player classifications is important, because if they were not, their 50% of the data would be useless. The acceptable accuracy found in *Happy Match* and *Forgotten Island* suggests that long-tail classifications are not a waste for this task. The overall accuracy of classifications generated by players is relatively consistent over time and at high enough level that new players, even those who leave shortly after trying a game, can provide data that is usable and comparable in quality, if not quantity, to long-term power players.

The usefulness of long-tail classifications raises another interesting issue regarding the design of purposeful games and gamified tools: the distinction between games that are genuinely engrossing to play and games that merely *seem* engrossing to play. Game designers aspire to the former, hoping to produce great experiences for players that will keep them entertained for hours, days, months, and even years. In the context of purposeful activities, however, there can still be value in producing games that fail to achieve this standard but still attract a critical mass of short term, "long tail" players. Indeed, our data about *Forgotten Island* suggests that this could be one such game – appealing at first, but only engaging for select users over the long term. Even so, if games look interesting and are tried by enough players, they may still produce data that is useful.

Are such games a form of Bogost's (2011) so-called "exploitationware?" If the intention is to attract players with false promises about the game experience, the answer must be "yes." However, if the intention is simply to create a good, short-term experience for players, the answer may be "no." Furthermore, while game designers never aspire to create bad games, for a variety of reasons, bad and mediocre games are far more common than great ones (Schell, 2008). Given the resources required to create an entertainment-oriented purposeful game, it is reassuring to know that even modestly engaging games can still produce meaningful data if they are tried by enough short-term players. Though not ideal, this effect mitigates at least some of the risks involved in producing purposeful games. It may also give scientists leeway to contemplate the design of game experiences that aspire to more than task-focused gamification.

8. FUTURE DIRECTIONS

While most citizen science games favor non-diegetic rewards and task-centric game play, *Forgotten Island* shows how diegetic rewards and a game world that is not tightly bound to the science activity can still produce data of value to scientists. This "game taskification" approach raises interesting possibilities, among them the potential to create scientific research tools that are also commercial entertainment products. Two possibilities seem especially interesting: 1) develop and release games like *Forgotten Island* for profit, supporting scientific research and game development with sales of the game, or 2) partner with existing game studios to integrate science tasks into commercial titles. Each approach has advantages and disadvantages.

For purpose-built citizen science games, the primary advantage is that the game can be exactly tailored to the science task, while the primary disadvantages are the time and resources required to plan, design, implement, release, and support the game as well as the difficulty of marketing and attracting players. This difficulty may be behind our finding that an interest in games was not correlated with more time spent on Forgotten Island. This outcome was surprising, as we had developed that game specifically to attract gamers and in other research about this project, we did indeed see that *Forgotten Island* was strongly favored by gamers. However, it may be that the system, as developed, does not have enough game elements, desirable mechanics, or appropriate balance to be attractive to this audience when deployed online as opposed to in a lab

For entertainment games that have science activities grafted onto them, the advantages and disadvantages are roughly reversed. Science activities may suffer in service to the entertainment game experience, even if development resources become less of an issue. Yet a for-profit game title that included a real world science activity, perhaps as a diegetically motivated mini-game, could have a potential marketing advantage over its competitors.

It is easy to envision how "grinding" tasks found in many current game titles, i.e. repetitive activities that allow players to accumulate diegetic resources like in-game money, experience points, or building materials, could be turned into real-world, purposeful activities. In many cases, this could be done without compromising the integrity of either the game experience or the science; for example, a space adventure game could easily integrate real-world astronomy activities, just as a plant biology activity might become part of an alchemy exercise in a medieval fantasy. As *Forgotten Island* demonstrates, data quality need not suffer unduly in entertainment-oriented games, as long as player activities are adequately measured so that bad data and unwanted player behaviors do not adversely impact the final data set.

9. LIMITATIONS AND CONCLUSION

In this study we explored a variety of differences between two purposeful video games for citizen science. Specifically, we studied how the diegetic and non-diegetic reward systems of purposeful games and "gamified" tools shape play experiences, impact player activities, and, most significantly, affect data quality.

We found that different reward systems and gamification approaches can certainly impact player recruitment and retention, as well as the ways that players experience purposeful games, but that these modalities need not adversely impact data quality. We also found that while most data in purposeful games for citizen science will be contributed by a few power players, the many players who make just a few contributions still provide quality data. The quality of contributions made by these long tail players does not appear to be adversely impacted by the specific reward structures or gamification approach that is used. However, a limitation of our study is that we examined only one task. The accuracy of newcomers on more difficult tasks might be lower, enough so that their contributions are not useful. We also drew upon a participant pool recruited largely because of their interest in citizen science. Previous work (Prestopnik & Tang, 2015) has suggested that different types of users might prefer different types of games, and this preference could also have an impact on data quality. This remains an open question worth exploring in future work.

A further limitation of the current study is the approach taken to computing accuracy based on the species classification, which overstates the actual accuracy. However, the problem affects all conditions equally, meaning that it does not affect the conclusions of this study. Still, it would be preferable to have more precise estimates of accuracy. To address this limitation, we are exploring other ways to compute accuracy. For example, with enough players, we could measure individual agreement with the consensus rating for a picture.

In future, we hope to explore how game design, commercial game design in particular, and purposeful game design might intersect to reach greater numbers of players in service to the creation of meaningful play experiences, the economics of the game industry, and the data requirements of scientists.

ACKNOWLEDGMENTS

The authors would like to thank the development team for their efforts on this project: Nathan Brown, Chris Duarte, Susan Furest, Yang Liu, Supriya Mane, Nitin Mule, Gongying Pu, Trupti Rane, Jimit Shah, Sheila Sicilia, Jessica Smith, Dania Souid, Peiyuan Sun, Xueqing Xuan, Shu Zhang, and Zhiruo Zhao. The authors would also like to thank the following for their partnership and assistance in Citizen Sort's design and evaluation efforts: Anne Bowser, Jennifer Hammock, Nancy Lowe, John Pickering, Jennifer Preece, Dana Rotman, Jian Tang, and Andrea Wiggins. This work was partially supported by the US National Science Foundation under grant SOCS 09–68470.

REFERENCES

Anderson, C. (2008). *The Long Tail: Why the Future of Business is Selling Less of More*. New York, NY: Hyperion.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research, 1*(1), 19.

Bogost, I. (2011). Persuasive Games: Exploitationware. *Gamasutra: The Art and Business of Making Games.* Retrieved from http://goo.gl/jK1VR

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience, 58*(3), 192-197.

De Freitas, S., & Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education, 46*(3), 249-264. doi:http://dx.doi.org/10.1016/j.compedu.2005.11.007

Delaney, D., Sperling, C., Adams, C., & Leung, B. (2008). Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions, 10*(1), 117-128. doi:10.1007/s10530-007-9114-0

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). *From game design elements to gamefulness: defining "gamification"*. Paper presented at the Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). *Gamification. using game-design elements in non-gaming contexts*. Paper presented at the CHI 2011, Extended Abstracts on Human Factors in Computing Systems, Vancouver, BC, Canada.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy, 43*(1), 1-20.

Galloway, A. R. (2006). *Gaming: Essays On Algorithmic Culture (Electronic Mediations)*: University of Minnesota Press.

Galloway, A. W. E., Tudor, M. T., & Vander Haegen, W. M. (2006). The Reliability of Citizen Science: A Case Study of Oregon White Oak Stand Surveys. *Wildlife Society Bulletin, 34*(5), 1425-1429. doi:10.2193/0091-7648(2006)34[1425:trocsa]2.0.co;2

Law, E., & von Ahn, L. (2009). *Input-agreement: a new mechanism for collecting data using human computation games*. Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems, Boston, MA, USA.

Malone, T. W. (1980). *What makes things fun to learn? heuristics for designing instructional computer games*. Paper presented at the Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems, Palo Alto, California, United States.

McGonigal, J. (2007). Why I Love Bees: A Case Study in Collective Intelligence Gaming. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, -, 199-227. doi:10.1162/dmal.9780262693646.199

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world.* New York: Penguin Press.

Orr, K. (1998). Data quality and systems theory. *Commun. ACM, 41*(2), 66-71. doi:10.1145/269012.269023

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Commun. ACM, 45*(4), 211-218. doi:10.1145/505248.506010

Prestopnik, N., & Crowston, K. (2011). *Gaming for (Citizen) Science: Exploring Motivation and Data Quality in the Context of Crowdsourced Science Through the Design and Evaluation of a Social-Computational System.* Paper presented at the 7th IEEE International Conference on e-Science, Stockholm, Sweden.

Prestopnik, N., & Crowston, K. (2012). *Purposeful Gaming & Socio-Computational Systems: A Citizen Science Design Case*. Paper presented at the ACM Group: International Conference on Supporting Group Work, Sanibel Is., FL.

Prestopnik, N., Crowston, K., & Wang, J. (2014, 4-7 March, 2014). *Exploring Data Quality in Games With a Purpose.* Paper presented at the iConference, Berlin, Germany.

Prestopnik, N., & Tang, J. (2015). Points, stories, worlds, and diegesis: Comparing player experiences in two citizen science games. *Computers in Human Behavior, 52*, 492-506.

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. Cambridge, MA: The MIT Press.

Schell, J. (2008). *The Art of Game Design: A Book of Lenses*. Burlington, MA: Elsevier, Inc.

Stam, R., Burgoyne, R., & Flitterman-Lewis, S. (1992). *New vocabularies in film semiotics*. London: Routledge.

von Ahn, L. (2006). Games with a purpose. *Computer, 39*(6), 92-94.

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM, 51*(8), 58-67. doi:http://doi.acm.org/10.1145/1378704.1378719

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst., 12*(4), 5-33.

Wiggins, A., & Crowston, K. (2011, January 04-January 07, 2011). *From Conservation to Crowdsourcing: A Typology of Citizen Science.* Paper presented at the 44th Hawaii International Conference on System Sciences, Kauai, Hawaii.