Exploring Data Quality in Games With a Purpose

Nathan Prestopnik[1], Kevin Crowston[2], Jun Wang[2]
[1] Ithaca College, Ithaca, NY 14850
[2] Syracuse University, Syracuse, NY 13244

Abstract
A key problem for crowd-sourcing systems is 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 scientific interest raises concerns about the quality of the data provided, which is particularly important when the data are to be used for scientific research. To assess whether these concerns are justified, 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, the quality of data from short-time contributors was at a usable level of accuracy. These findings suggest that various approaches to gamification can be useful for motivating contributions to citizen science projects.
Keywords: Games with a purpose; data quality; diegetic rewards; citizen science; engagement

Acknowledgements: 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 so far: Jennifer Hammock, Nancy Lowe, John Pickering, Jian Tang, and Andrea Wiggins. This work was partially supported by the US National Science Foundation under grant SOCS 09–68470. Kevin Crowston is supported by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

A key problem for volunteer-based projects is motivating contributions from participants and ensuring the quality of these contributions. These concerns are interrelated, in that system designs intended to maximize the volume of contributions may do so at the cost of quality and vice versa.

In this paper, we examine the interplay between motivation and quality of participation in the context of online citizen science projects. 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, and collecting data from the physical environment 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).

Citizen science projects must address concerns about the quality of contributions, in this case, questions that arise from suitability of the generated data for the science goals of the projects. Data quality is a multi-dimensional construct (Orr, 1998; Pipino, Lee, & Wang, 2002; Wang & Strong, 1996), but the believability or accuracy of the data remains a particular concern for citizen science projects because many participants are not trained scientists. Their limited scientific knowledge may possibly affect the accuracy of the data they provide.

Cognizant of this concern, previous studies have examined 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. Goodchild and Li (2012) explored mechanisms for assuring the quality of citizen-provided geographic information. Delaney et al. (2008) checked data quality in a marine invasive species project, finding that participants were 95% accurate in their observations. 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 existing and enthusiastic communities of interest; 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 face limits on their available pools of participants, namely those who share a particular scientific interest. Less charismatic topics of inquiry that lack a large natural base of users could benefit from alternative mechanisms for motivating participants. Purposeful games have the potential to become one such 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.

Relying on games to motivate participation may have negative tradeoffs with data quality. Games are meant to be entertaining, and players may find themselves concentrating only on the fun elements of a game, ignoring, neglecting, or even cheating on embedded science tasks to get them over with quickly. Games that are designed to prevent such behaviors may improve data quality but not be fun for players and so fail to attract very many participants.

The interrelated issues of game-driven participant engagement and citizen science data quality are of interest to game designers, human-computation systems 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, and how different approaches to "gamification" influence player motivation and data quality. In this paper, we address these questions.

2 Theory: Gamification and Games With a Purpose

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 yet retain 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 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 real 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 and is strongly tied to the story and the game world. Conversely, non-diegetic rewards are those that have only limited connection to the game world, but sometimes (not always) can 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., "kill ten enemies in ten seconds," or "collect 1 million coins"). 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.

Because non-diegetic rewards are weakly tied to the game world (at best) and do not deeply 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 them in the first place. For "non-enthusiast" players, non-diegetic rewards likely have more 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. For these players, non-diegetic rewards are probably not an effective approach. Not surprisingly, 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.

Yet non-enthusiast players are those most likely to find value in a game layer that can turn "boring science" into "play." Diegetic rewards may be more engaging and more meaningful for non-enthusiasts, underscoring, as they can, the game story, game world, and game play instead of the real-world task.

Diegetic rewards can thus become a powerful form of feedback to keep non-enthusiasts immersed in a game that only occasionally asks them to undertake a science task. The benefits for those managing citizen science initiatives of employing diegetic rewards– i.e., an enhanced ability to attract and engage non-enthusiast participants – are also apparent.

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 itself. 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. Malone (1980) has noted how many of these can be motivating in the context of gamified experiences, specifically educational games. 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.

Others have designed and studied entertainment games for citizen science. For example, *Fold.It*[1] and *Phylo*[2] are both citizen science projects in the form of entertaining games that have attracted substantial numbers of players and produced large amounts of scientific data. To date, however, there has been little formalized comparison of diegetic and non-diegetic rewards in gamified experiences, particular as these relate to player performance and data quality. In particular, to our knowledge there has been no formalized comparison made of these different approaches using the same citizen science task as a basis for two very different modes of gameplay. Yet it is plausible that different reward structures and philosophies of gamification could 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 of the game. Though this could heighten the chances of attracting non-enthusiast players, it may also be that these players will ignore or neglect the science in lieu of playing other parts of the game, potentially reducing data quality. Even cheating – i.e., knowingly submitting bad data – could be beneficial to players who are fixated on the entertainment experience and so motivated to skip over the science work.

To explore these issues we designed two very different games around the same purposeful activity in order to study the impact of different approaches to gamification on data. One game adopted a straightforward gamification approach, rewarding players for performance with non-diegetic score points and focusing primarily on the science task. The second was an entertainment-oriented purposeful game, 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 game money as well as the ability to interact with various characters, progressively explore the game world, and advance the game story.

Game designers and HCI researchers are also likely to find our work interesting. Very few purposeful games have been explicitly designed as story experiences featuring diegetic reward structures, and almost none have been built in a design science tradition with scholarly study as a key goal of the design and development process (Hevner, 2007; Hevner, March, Park, & Ram, 2004; Prestopnik, 2010). Our unique context (citizen science) and design-based approach to study can extend our understanding of purposeful game design, particularly with regard to the quality of data that different game design philosophies may achieve.

[1] http://www.fold.it
[2] http://phylo.cs.mcgill.ca/

3 Research Questions

We developed a guiding set of research questions with our overarching scholarly interests in mind. 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, the distinct 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. 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). 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 power users who make many contributions each). As there are many players in the long tail, the combined number of classifications provided (even 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, while voluminous, may be scientifically worthless. We therefore addressed one final question:

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

4 System Development

The two purposeful games that we designed to address these questions were centered 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., on 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 when the photograph is captured in the field by scientists or amateur enthusiasts.

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. However, taxonomic keys are usually written for expert users, and are often complex, highly variable, and difficult to translate into a form that will be suitable for use in a human computation systems (much less games).

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.

Citizen Sort features two purposeful games. The first, *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 a 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 earn high scores by correctly classifying the character-states of specimens in photographs for which the answers are known, by dragging each photograph to

the correct state, character by character (see Figure 1). The few known "happy" photos are mixed with other photos that are still to be classified, i.e., for which the character-state information needs to be collected.

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). The "happy" photos are revealed only at the end of the game, so players must strive to perform well on all photos to ensure a good score.

Figure 2. The *Happy Match* score interface.

Happy Match rewards players with points based upon their performance. However, we would argue that *Happy Match* differs from what Bogost (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, they are a meaningful performance metric and reward for those who do.

The second game, *Forgotten Island*, has players performing the identical classification task as *Happy Match*, one photo at a time, including the use of known photos as a way to check accuracy (Figure 3). However, 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). Rather than points, players are

rewarded with in-game money (a diegetic reward) for each classification. This money is spent to acquire further diegetic rewards: equipment and items that can advance the plot and open up new game spaces to explore. To motivate effort, incorrect answers for a known photo results in a warning and a slight deduction in game money.

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 3. The *Forgotten Island* classification interface.

Figure 4. 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, learning 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*. *Citizen Sort* and its two games are easy to find with online searches for citizen science activities and, in some communities, by word of mouth. This is to say that the participants for this study came to the project in a manner similar to any other current citizen science project. *Citizen Sort's* user base is therefore likely to be representative of many other citizen science initiatives.

The number of *Citizen Sort* users is growing. The data presented in this paper is drawn from approximately 900 user accounts, excluding developer accounts and approximately 100 temporary players (i.e., players who are 13 years of age or younger whose performance is not tracked between visits; under US law, users must be over 13 to create an account).

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. In this section, we present the results of this analysis.

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.* The 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 (87% of players for Happy Match and 74% for Forgotten Island) and a few "power" players played for many days. Therefore 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 ($p = 0.002$), with "power" players playing Happy Match for significantly more days.

Figure 5 shows the distribution of the number of scientific contributions (i.e., classifications) in the two games. The retention differences between *Happy Match* and *Forgotten Island* are also apparent in Table 1, comparing the percentage of retained players after just one classification decision, after 20 decisions, and after 50 decisions. Similar to many online systems, both games see a high initial attrition: when players try the game for the first time, most quickly lose interest and do not return. Attrition for *Happy Match* appears to continue at a steady rate, with only a small core set of "power" players continuing to contribute regularly.

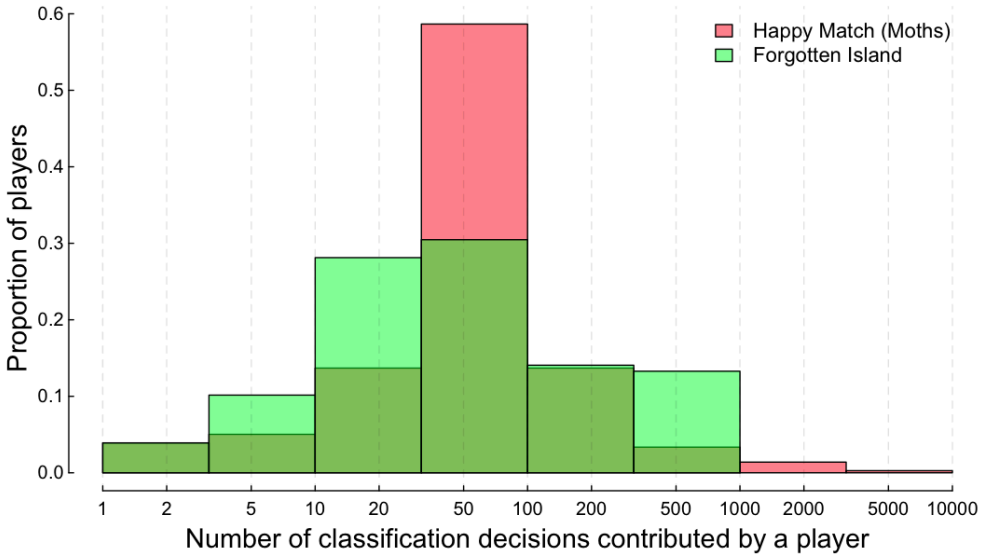

Figure 5. Distribution of number of decisions contributed by *Happy Match* and *Forgotten Island* players.

	Retained at 1 Decision	Retained at 20 Decisions	Retained at 50 Decisions
Forgotten Island	45%	32%	16%
Happy Moths	92%	79%	33%
Happy Rays	93%	76%	38%
Happy Sharks	89%	63%	21%

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

In contrast, for *Forgotten Island*, the rate of attrition seems to fall off after a larger initial loss. Note that players do not make their first classification decision until some way into *Forgotten Island*. Some players may decide that they are not interested in *Forgotten Island's* story and game world before making any classifications, which may explain the more rapid drop-off in retained players at the point of making the first classification decision. However, it seems that if a player does find their interest captured by *Forgotten Island* they are more likely to continue playing until the end of the game. Note also that unlike

Happy Match, *Forgotten Island* can be "won" and its story eventually concludes. While players can continue to play parts of the game, for the most part, *Forgotten Island* is finished at this point. It takes about 320 classification decisions to win *Forgotten Island*.

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*, both of which use the moth photo dataset.

We computed accuracy by comparing players' answers for pictures to the known correct answer. To increase the pool of classifications for the comparison, we ran the game 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). We compared the accuracies of players of the two games using a two-sample t-test. To our surprise, our results showed no significant difference in the accuracy of the data provided by *Happy Match* and *Forgotten Island* players.

	N (sample size)	Classification Accuracy
Happy Match (Moths)	289 players	0.806
Forgotten Island	81 players	0.802
		p-value=0.746

Table 2. Comparing classification accuracy.

Despite the overall similarity in accuracy, we did find some evidence that player classification behaviors and the accuracy of data produced by players interact and vary between the two games. Specifically, in *Forgotten Island* we observed a number of instances of "cheating" behavior, identified by checking the mean time spent by a player on a single classification and the overall accuracy of those classifications. Cheaters had a distinct signature: 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, and some players simply struggled with classification. However, fast classifications coupled with poor accuracy seemed to indicate the profile of a player more interested in game play than in classification.

Figure 6 plots *Forgotten Island* response time against performance for individual players. Red circles represent data for the first 20 photos and green circles represent data for all photos for a player. The two green circles in the lower left of the chart 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.

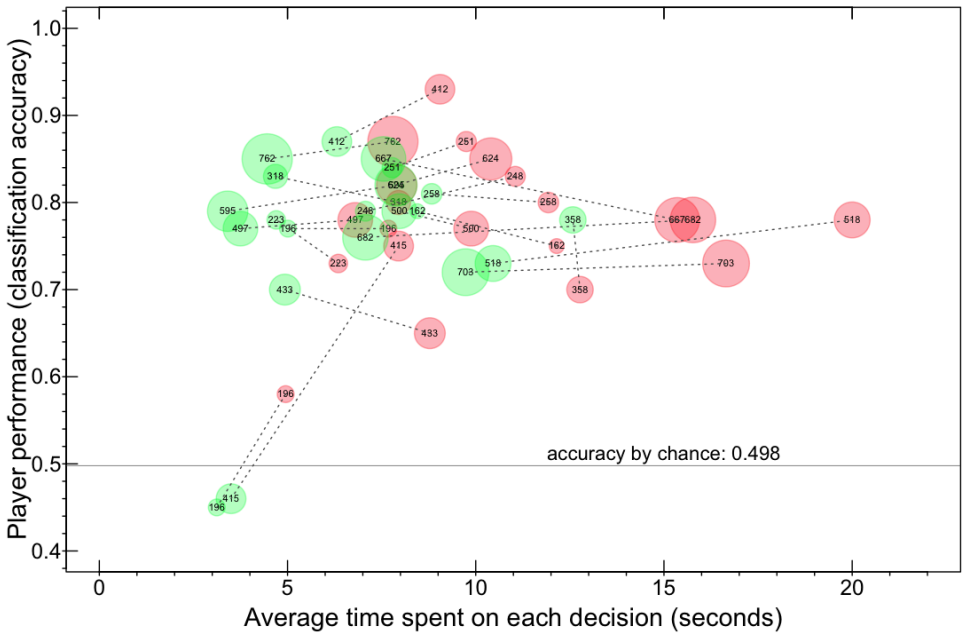

Figure 6. Performance vs. response time in *Forgotten Island*.

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." For example, in *Happy Moths,* just 4.4% of players contributed 50% of the decisions. 63% of players played only one game (including those who did not finish), 37% played at least two games, 19% played at least three games, 12% played at least four games, and only 8% played at least five games.

We expected that there would be a positive correlation between the number of classifications that players contribute and their performance, that is, that players would learn the game and the characters and states and so improve their performance. We used Spearman rank correlation to measure the relationship, and we restricted the sample, as before, to people who had contributed a minimum of 20 decisions. The results are shown in Table 3 and graphed in Figure 7. To our surprise, we did not find a significant correlation for any of the four games (*Happy Moths*, *Happy Rays*, *Happy Sharks*, or *Forgotten Island*), meaning that those who contributed longer were not more accurate.

	N (sample size)	Rho	p-value
Forgotten Island	81	–0.043	0.700
Happy Moths	289	–0.098	0.096
Happy Rays	208	0.070	0.317
Happy Sharks	107	–0.056	0.569

Table 3. Correlation between number of classifications and accuracy.

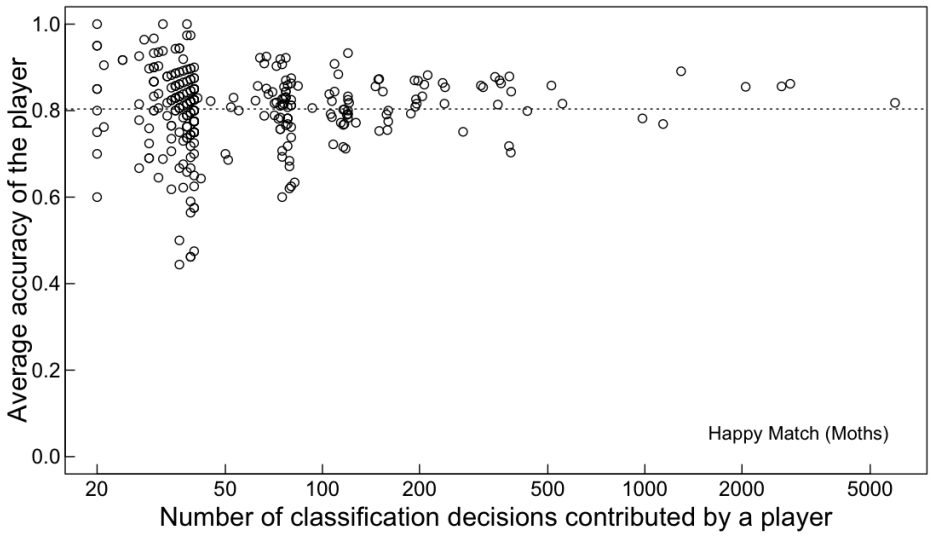

Figure 7. Player performance vs. contributions.

7 Discussion

The most interesting findings from the comparison above were the overall similarity between the two games, with the exception of cheating, and the lack of a learning effect.

7.1 Cheating Behavior

Cheating behavior was apparent only in *Forgotten Island*, underscoring how non-diegetic and diegetic reward systems can have different impacts on player behaviors and data quality. There is little reason for *Happy Match* players to cheat: for power players, achieving a score-based reward without also achieving some meaningful experience would be pointless, while for long-tail players, neither the points nor the game experience are worth the effort of cheating. Power players will attempt to do well because they are personally interested in doing so, while long-tail players who are uninterested in the science activity simply stop playing *Happy Match*.

Forgotten Island, on the other hand, has built-in incentives that make cheating more likely and potentially beneficial to certain players. The diegetic reward system 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. Players who are interested in the science activity may still not want to do well on it. However, for players who enjoy the game but not the science task, cheating will make the overall game faster and easier, allowing players to focus on the diegetic rewards – game money, the game world, and the story – rather than the work required to progressively experience them. As a result, cheating may be an attractive proposition for players who realize that they can still make enough money to play through the game even when doing poorly in the classification task.

Forgotten Island is currently configured so that cheating is a feasible strategy, a decision driven by the overall game experience and not just the need for accurate classifications. It would be 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, 2012) 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 could easily render the game too difficult to play, as non-cheating players would be regularly punished for well-intentioned but incorrect answers. Accordingly, the game was configured to be easy so that players can continue to make progress even though this design choice means that cheating is viable for players who realize it.

Overall though, there was no significant difference in performance between *Happy Match* and *Forgotten Island*, comparing players who made a minimum of 20 classification decisions (i.e., the level of cheating was not high enough to significantly affect the overall results). 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.

7.2 Lack of Learning Effects and the Value of the Long Tail

We found no evidence for learning effects in either *Happy Match* or *Forgotten Island* when looking at players who had classified at least 5 photographs. This is an interesting, unexpected and useful finding. 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 lack of a learning effect coupled with the acceptable accuracy found in *Happy Match* and *Forgotten Island* suggests that long-tail classifications are not a waste. 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 *look* 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. These games may be intentionally designed as short-term experiences, or may simply be games that fail to live up to their promise. Either way, if they look interesting and are tried by enough players, they may very well 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 (Prestopnik & Crowston, 2012) 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.

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 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 *Happy Match* and *Forgotten Island* demonstrate, 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.

A limitation of the current study is the approach taken to computing accuracy based on the species classification. To address this limitation, we are exploring more precise 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.

10 References

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

Bogost, I. (2011). Persuasive Games: Exploitationware. *Gamasutra: The Art and Business of Making Games.* 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.

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

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*(0), 110-120. doi: http://dx.doi.org/10.1016/j.spasta.2012.03.002

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems, 19*(2).

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, 28*(1), 75-105.

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. (2010). Theory, Design and Evaluation – (Don't Just) Pick Any Two. *AIS Transactions on Human-Computer Interaction, 2*(3), 167-177.

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.

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.