Motivation and data quality in a citizen science game: A design science evaluation

Kevin Crowston & Nathan R. Prestopnik
Syracuse University School of Information Studies
crowston@syr.edu & napresto@syr.edu

Abstract

Citizen science is a form of social computation where members of the public are recruited to contribute to scientific investigations. Citizen-science projects often use web-based systems to support collaborative scientific activities. However, finding ways to attract participants and ensure the accuracy of the data they produce are key issues in making such systems successful. In this paper we describe the design and preliminary evaluation of a simple game that addresses these two concerns for the task of species identification.

1. Introduction

In citizen science projects, members of the public are recruited to contribute to scientific investigations [1, 2]. Notable successful citizen-science projects include GalaxyZoo, which asks participants to help classify astronomical photographs, eBird, which collects bird sightings or FoldIt, in which participants use spatial reasoning skills to fold protein strings. Such activities draw many individuals into a cooperative endeavor toward a common scientific goal.

While their scientific areas of inquiry vary widely, these projects adopt a common approach to data collection: large numbers of individuals are asked to submit individual observations or analyses via web- or mobile-based technologies. Projects often rely on some form of socio-computational system, as they feature a mix of tasks that can only be performed by people (e.g., making an observation or classifying an image) supported by computational scaffolding to organize these efforts. Prior research [3] has shown that projects apply a variety of technologies to support their science goals.

Exploring questions that lie at the intersection of the citizen science phenomenon and the systems designed to support it can be highly challenging. Studying ongoing citizen science projects and their associated technological infrastructure will often bear useful fruit, but the owners of these systems are typically hesitant to grant very much access lest experimental manipulation or other forms of inquiry disrupt their own successful data collection efforts. On the other hand, citizen science is a phenomenon where naturalistic methods of inquiry—studying real-world projects in realistic situations—is generally more useful than using simulations or other artificially controlled approaches. Our solution to this dilemma was to adopt a design science approach to inquiry.

Design science couples traditional research methodologies with the development of an IT artifact to address research questions along with design-related problems [4-6]. Design science is practiced (mostly without using the term) in many domains, particularly human-computer interaction (HCI) and computer science (CS) more generally. The term and its formal conceptualization come from the field of information systems (IS), where system design is often viewed as atheoretical and so not really research. In this setting, rigorous conceptualizations of design as a research tool are necessary to encourage its broader acceptance. However, even in fields where system design is generally embraced, the formal conceptualization of research as design science can be valuable, as too often the focus on designing useful artifacts results in inattention to larger research questions. For example, in [7] many HCI evaluation practices are criticized as “usability evaluations” instead of scientific “evaluations for research”, what [8] calls the “I did this and it’s cool” form of study.

Design science research has two equally important outcomes: 1) a functional IT artifact that helps address a specific, challenging and practical design problem within a given context and 2) meaningful scholarly contributions to a field of inquiry. Compared to typical social-science research approaches, the design science approach requires additional components, including interactions with subject-matter experts (SMEs), a situational focus on the context in which a design will be deployed, as well as system building and testing. Compared to typical systems research, the approach requires explicit use of theory to guide design decisions and—importantly—an ability to draw more general conclusions about these theories from the experience of building and evaluating the system. Problems suitable for a design science approach include both those that are unsolved and those that offer opportunities for newer or better solutions [4].

Our design science approach manifested as follows: partnering with naturalists and biologists, we set out to design and build a new citizen science initiative from scratch. This initiative, called *Citizen Sort*, addresses a challenging problem in the life sciences: the taxonomic classification of plant, animal, and insect species from photographs. Experts, enthusiasts, and curious members of the general public routinely collect and upload photographs of different living things. A photograph of an insect, plant, or animal, tagged with the date and location where it was taken, can provide valuable scientific data (e.g., on how urban sprawl impacts local ecosystems or evidence of local, regional, or global climactic shifts). To be useful though, it is necessary to know what the picture is of, expressed in scientific terms, i.e., the scientific name of the species depicted. *Citizen Sort* was developed to let average, ordinary members of the public view collections of pictures maintained by researchers and annotate them with data about the specimens they depict, with the goal of classifying the picture as a particular species.

To be meaningful to researchers outside of the specific problem space though, the IT artifacts developed for design science should be a vehicle for broader scientific inquiry. Three components—theory, design and evaluation—are thus interrelated in design science research, coherent pieces of a whole [9] and conducted iteratively [4, 5].

Theory: The word "theory" is used broadly here [10], encompassing the adoption of existing theory as a lens through which to approach design, as well as consultation with experts and review of non-theoretical, project-specific design literature. This stage may also result in the generation of new theory, produced either from literature or from data, and conceptualized either prior to design of the IT artifact, during its development, or after its evaluation. The theory stage may be seen as both a beginning and an end to design science research: theory adopted early will inform design and new theory will come from it.

Design: Design science research revolves around the design of an IT artifact, where theoretical and practical underpinnings shape a functional system. The designed artifact may ultimately produce new theory, so artifact design must take future evaluation into account. The design scientist must always keep in mind the research questions to be addressed through research evaluation of the artifact.

Evaluation: The evaluation stage is about more than saying "yes, this worked," or, "no, this didn't work." It must address the project's broader research questions by validating the adopted theory or leading to the generation of new theory. Evaluation is not always an end point for research: evaluation will often suggest ways to improve the artifact (as a system to address the problem space or as a research tool) in its next design iteration.

We designed *Citizen Sort* to help us explore a variety of open questions about the citizen science phenomenon, keeping the theory, design, and evaluation stages well in mind. Prior research [3] evaluated a variety of current citizen science projects from a technological standpoint. The researchers noticed that games and game-like features are frequently cited as having great potential to motivate participation in citizen science, yet few projects actually include games. We therefore made purposeful gaming a central feature of *Citizen Sort*, developing *Forgotten Island*, a point and click adventure game in which taxonomic classification plays a central role, and *Happy Match*, a sorting and matching game that awards points and high scores for classification.

In this current paper, we limit our discussion to a preliminary evaluation of *Happy Match*, which was undertaken to address two research questions: 1) Can we design a classification game that achieves good data quality by using taxonomic keys prepared by experts? and 2) Will game-like features motivate users to contribute classifications?

2. Theory

Our two research questions oriented this study around two conceptual bases. First, the primary goal of many citizen science projects is to obtain scientifically valid data to support research. Therefore, a socio-computational system for citizen science must present tasks so that non-expert participants can accurately perform them, producing high quality data that can be meaningfully used by experts. This goal poses a serious design challenge, which we addressed by drawing upon theories and knowledge of classification from the natural sciences and data-quality frameworks.

Our second conceptual base arises from our interest in motivation and games. Given the voluntary nature of citizen science, ensuring adequate levels of participation and enjoyment are important design considerations. To address this design challenge, we drew upon theories of motivation and purposeful gaming from the information systems, collective intelligence, and educational gaming literature.

In the remainder of this section, we review the theoretical perspectives we adopted to explore our two research questions.

2.1. Data Quality

Data quality is a necessary precondition for the further scientific use of the data. We adopted the data quality framework suggested by [11], which is composed of four data quality attributes: 1) *intrinsic data quality*, the believability or accuracy of the data, 2) *contextual data quality*, how relevant, timely and complete the data is, 3) *representational data quality*, how interpretable and easy to use the data is and 4) *accessibility*, how easy the data is to access and use.

Intrinsic data quality is the key concern. As in many scientific problems, there is a "ground truth" (i.e. correct answers) within the taxonomic classification context. In other settings, the reactions of an individual to some item are of interest; e.g., the goal of a system to produce keywords for images is to identify words that people naturally use to describe the images. In contrast, for data to be scientific, valid and accepted, participants must produce the "right" answers, i.e., answers that are in agreement with experts. Participant opinions *per se* are not useful in this context.

Unfortunately, in many areas of science, specialized knowledge is required to provide data, while few citizen science participants are experts. Some participants will have the necessary knowledge (e.g., avid birders can generally identify the species they observe), but many potential participants will not. Therefore, finding methods to turn scientific tasks into things that non-scientists can do well, as well as finding techniques to confirm the validity of participant-provided data, are important research goals.

To identify specimens to the species level, biologists have developed tools in the form of taxonomic keys. These keys identify species from their particular combinations of characteristics, known as character-state combinations (i.e., attributes and values such as "colour: yellow")[1]. Specific characters and states vary by taxon, but are broadly similar in structure. Given sufficient characters and states, it is possible to identify a photographed specimen to a specific family, genus, species, or even sub-species. For example, characters useful for identifying the species of a moth include simple features such as its shape at rest or wing color as well as more subtle features such as the colour of its "discal" and "orbicular" spots, circular patches on the wing that are common in some families of moth and absent in others. The different colors of the spots and their borders help identify the particular species.

A challenging aspect of this problem is that researchers working within the same biological or ecological disciplines do not necessarily agree upon taxonomic keys. In fact, many researchers develop their own key variations to support their own specific research endeavors. Furthermore, keys are typically 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 socio-computational system, where expert understanding of characters, states and taxonomic identification cannot be assumed.

Even with an established key, some characters and states are beyond the ability of untrained members of the general public to identify (e.g., the previous "orbicular spot" example). Indeed, some require true expert knowledge (e.g., classifying species by their sex organs). An system designed to support classification will be unlikely to effectively support both extremely knowledgeable users and extremely novice users. Experts will require advanced tools with great flexibility, while novices may require simplified systems that have expert knowledge pre-built into them. Furthermore, some characters require specialized equipment (e.g., classifying species by their genetic makeup). A web-based classification system will only be able to support some kinds of characters and states, while others will be impossible.

Contextual and representational data quality, as well as data accessibility, are also concerns in the design of a socio-computational system for citizen science. To be successful, such systems must produce complete data organized in a manner that will be useful to expert scientists. In the classification context, omitting some characters or states because they are difficult for novice users could result in higher accuracy, but lower contextual data quality. Similarly, the data produced by such systems must be preserved in formats that can be returned to experts for relatively easy use.

2.2. Motivation

A second critical issue in socio-computational system design generally, and citizen science systems in particular, is attracting and retaining enough participants to make achievement of project goals possible. Systems with too little participation will be unlikely to generate meaningful quantities of scientific data, adversely impacting contextual and representational data quality (no matter what the data accuracy); these systems can also benefit participants themselves, for example through formal or informal learning benefits or a sense of personal achievement. As a result, the motivations of citizen science participants are important to understand in order to attract new participants and retain old ones.

In [12], three basic motivations for individuals are suggested: money, love, and glory. For citizen-science projects, regularly offering payment to participants is rarely an option as project resources are typically too low. Rather than expecting compensation for their efforts, participants indicate that inherent interest in the subject of scientific inquiry, the relevance of data collection efforts to particular interests or hobbies, the perception that a project will be fun and engaging, an interest in collaborate with experts, altruistic reasons, and hope for broader recognition as reasons for becom-

[1] See http://www.discoverlife.org/mp/20q for examples

ing involved in citizen-science projects [13-17]. These reasons match well with the notions of “love” and “glory” as motivators [12]. As a result, most citizen-science projects rely heavily on participants who have preexisting enthusiasm for the scientific topic of the project, be it astronomy, bird watching, or insects.

Unfortunately, while some scientific topics are highly “charismatic”, many others are not. For example, bird watching, astronomy, and conservation all have existing communities of interest and a certain appeal, even for non-enthusiasts. However, important work is also being conducted in areas that attract much less public interest, such as moth, mold, or lichen classification. While enthusiasts exist for virtually all areas of the natural sciences, socio-computational systems rely on attracting large numbers of participants. There has been less scholarly or practical attention paid to how citizen science systems might be designed to motivate participants who do not hold these predominantly intrinsic motivations.

In the broader collective computing domains, several models for attracting participation have been deployed. For example, the ESP game (an image tagging system) [18], Phetch (which produces accessible descriptions of images) [19] or TagATune (where users tag music clips) [20] are designed as games, capitalizing on “love” forms of motivation, and giving people enjoyable activities to undertake while producing meaningful work almost as a by-product.

A few citizen science projects like Fold.It[2] have used games as an effective motivator. Others, like Stardust@Home[3], encourage participants by providing individual scores and achievements similar to those found in games. Such work is consistent with a long line of scholarly inquiry that shows the potential for games and gaming as motivational tools in various contexts [e.g. 21, 22-24]. However, this approach seems to be used only rarely by citizen science projects. In the 27 websites reviewed in [3], only a few projects used games or game-like features. We therefore wanted to explore the possibility that games will be effective in motivating participation in citizen science endeavors, particularly among those without an inherent interest in the underlying topic; we were interested in potentially motivating features such as game scores, competition among players, collecting rewards or badges, fantasy elements, puzzle solving and interactive storytelling. We were also interested in what players of such games might take away with them, including the possibility of new knowledge, greater enthusiasm for scientific activities, or heightened awareness of different kinds of science.

Finally, we note that our two research questions may interact, as the effect of game-like interactions on data quality is unknown. For example, creating too strong an incentive to get a high score might lead to participants attempting to cheat or “game” the system, diminishing rather than increasing data quality. Contrariwise, a heavy emphasis on data quality might make the game less enjoyable to play and so decrease participation.

3. Design

In this section, we describe the development of *Happy Match*, a taxonomic classification game. By extension, we also address some design elements of the broader *Citizen Sort* project of which *Happy Match* is a component. As previously mentioned, purposeful games in the citizen science context are not entirely new. However, the recent advent of the online citizen science phenomenon itself and the relative novelty of purposeful games (in this or any other context) continue to make this a interesting area for research.

Citizen Sort and *Happy Match* (along with several other related IT artifacts) were designed and implemented by a development team of twenty-five professionals and students with varied technical and artistic expertise. Seventeen of the developers were hired on the project as either part- or full-time employees or volunteers. The remaining developers participated through their coursework (i.e., developing systems or components of systems for a class). In addition, we worked with several domain experts who provided expert knowledge on taxonomic classification in the biological sciences, as well as specific information about classification of species.

The *Happy Match* game has a variable title, based upon the photo collections that it is used to classify, e.g. *Happy Moths*, *Happy Sharks, Happy Plants*, etc. Characters and states are specialized for each version of the game. We are currently working with a moth-specific version of *Happy Match* (*Happy Moths*), which is built around characters and states established by professional lepidopterists, naturalists and biologists. The key for *Happy Moths* has four characters: 1) Shape at Rest, 2) Forewing Main Colour, 3) Forewing Distinctive Colour, and 4) Forewing Pattern. Each character has between 6 and 8 possible states. The collection of moth pictures was also provided by one of our domain experts.

Happy Match is deigned to attract and retain participants to a citizen science project through the use of games, while still maintaining the quality of data provided. In *Happy Match*, each character of interest is presented in its own round, with up to ten characters per game and up to eight states per character. Players

[2] http://fold.it/

[3] http://stardustathome.ssl.berkeley.edu/

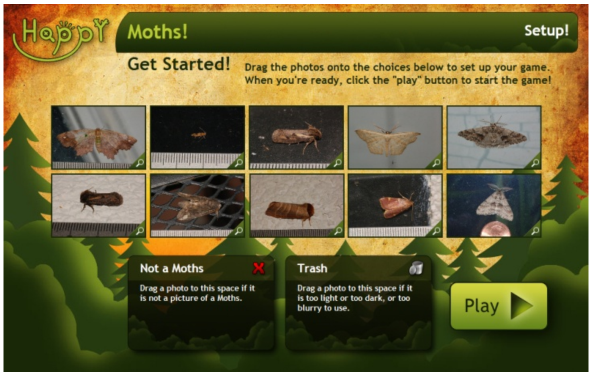
are presented with either five or ten photographs (depending on the number of rounds in the game) of some organism. They progress round-by-round, choosing the best state for each photo in each round. The system provides popup help to assist players in understanding the classification task and the game rules. A game might initially take as long as fifteen minutes to play, but an experienced player (one who knows the characters and states) might be able to finish a round in a minute or two. The classification aspect of the game is similar to the GalaxyZoo[4], in which participants classify for the shape of galaxies, though GalaxyZoo works galaxy-by-galaxy rather than character-by-character. Agreement among classifications performed by different users on the same photo can be used as an indicator of data validity for a particular specimen.

A key difference between *Happy Match* and GalaxyZoo is that *Happy Match* includes a scoring system. The game is seeded with photographs for which one of our domain experts has already determined the applicable states. In each round, at least two of the ten photographs given are already classified (these are referred to in the moth version of the game as the "happy moths"). *Happy Match* players are scored based on how well their classification decisions match those for the known photos (10 points for each correct decision plus 20 points for identifying the correct species). The scoring mechanism is intended to motivate players to do the classification carefully. Because players do not know which photos are the "happy moths" until the end of each game, they must try to do well on all photos to achieve a high score. As well, the accuracy of players on the known photos can be taken as evidence of their data quality.

4. Evaluation

In this section, we report on a preliminary evaluation of the performance of *Happy Moths* in order to answer our two research questions: can we design a system that achieves good data quality by using the character-state keys and will game-like features motivate users to contribute classifications?

To do the evaluation, we developed a descriptive and correlational study. We recruited subjects to play the game and collected data about the accuracy of their classifications and level of participation. For this preliminary evaluation, we used only pictures of moths for which we had known data so we could compute accuracy, that is, the players' level of agreement with the classified data provided by our domain experts. Participants also filled out a short survey. We then used descriptive statistics to compute accuracy and a correlational analysis to analyze predictors of motivation and accuracy. In the remainder of this section, we describe the elements of the research design in more detail.

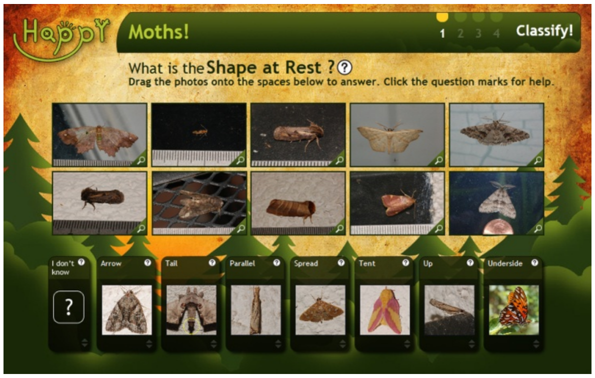


Drag the photos onto the choices below to set up your game.
When you're ready, click the "play" button to start the game!

Drag a photo to this space if it is not a picture of a Moths.

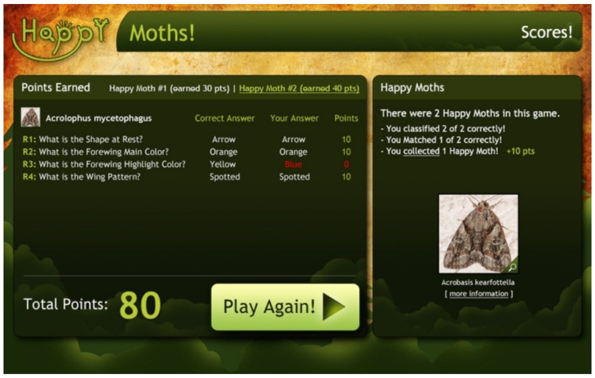
Drag a photo to this space if it is too light or too dark, or too blurry to use.

Figure 1. *Happy Moths* setup screen, where photos can be pre-sorted as bad images or not an example of the specimen of interest.



What is the Shape at Rest ?
Drag the photos onto the spaces below to answer. Click the question marks for help.

Figure 2. *Happy Moths* game round, where players are asked to answer a question (character) by dragging a photo to the appropriate answer (states).



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

Figure 3. *Happy Moths* scores page, where players are provided feedback on their performance, and rewarded for correctly classifying the hidden "Happy Moth."

[4] http://www.galaxyzoo.org/

Subjects. For this preliminarily evaluation of *Happy Moths*, we recruited players using Amazon Mechanical Turk (AMT), a “marketplace for work that requires human intelligence”[5]. The AMT system allowed us to dispatch a small task to a pool of workers who performed it in return for a small payment. We note that this subject pool is not really appropriate to test theories about motivation, as offering payment makes it difficult to assess the effects of other motivations (i.e., love and glory). However, in this preliminary evaluation our main interest was on our first research question, data quality, as well as the general usability/playability of *Happy Moths.* The rapid results offered by AMT seemed a good tradeoff for this stage of the project. As well, AMT users seemed to be representative of our target population of active Internet users.

In setting up the AMT task, we offered to pay up to 100 users in each round of the study and ran two rounds in total, for a planned total of 200 participants. Because of the way AMT works, more than 100 people started each round. However, not all who started completed the task and of those who did, not all completed the survey that was required to be paid. For each round, we had the desired 100 responses within a day at a total cost of less than $100.00 per round.

Evaluation tasks. Those who accepted the AMT task were asked to accept an informed consent statement, to play *Happy Moths* at least once and to then fill out a survey. We offered to pay participants US$0.50 for completing a game and the survey. To motivate good performance on the game, we offered an additional US$0.50 for getting a score of 50/80 or higher. This score requires 5 of 8 correct classification decisions to be made on the two “happy moths” combined (4 decisions per each “happy moth”). We linked performance on the game to the survey results using a unique identifier, though a few players did not copy the identifier correctly, making their data unavailable for analysis.

Data collection. The *Happy Moths*/*Citizen Sort* system collected the number of games each player played and their score on each game. From the scores, we computed both the average score and high score. We also determined whether a player played additional games after obtaining the bonus (i.e., after a game on which they scored the 50 points required to obtain the bonus), and if so, how many games. Finally, the system recorded each classifications performed by the users (with some omissions in the first round, corrected in the second round). As noted above, for this evaluation, we had a record of professionally applied classifications for all the moths in the game, enabling us to check the agreement of every classification decision with the known data, not just decisions made on the “Happy Moths”, which determined the score From these data we computed each player’s overall accuracy (the fraction of their classification that agreed with the expert).

After playing, users filled out a 28-item survey administered through AMT. Survey questions were developed in discussion among members of the research group. The survey asked about: 1) Experience of game play (how well they knew how to play, how difficult they felt the game was, confidence in their decisions, which character was most difficult to classify and why, how much fun the game was to play, how often they used the pop-up help and any problems encountered); 2) Likelihood of playing in the future if not paid, with connections to social media, with the ability to play against friends and if they knew the work would help scientists; 3) How much they learned about moths, classifying insects and doing science, 4) Level of interest in gaming (including hours spent playing games) and in nature-related activities; and 5) Demographics (age, education, and gender). Most items were 7-point Likert scales. After running our first trial, we found problems interpreting the survey items in group 2. We changed these items for second trial, meaning we have usable data for these from only half the users.

Analysis. We first carried out exploratory data analysis, e.g., plotting histograms for each data item separately. Some variables that were found to be skewed were log transformed. We then developed tables for the descriptive data, e.g., of games played and accuracy. Finally, to answer research questions, we used stepwise regression to identify the factors that predicted accuracy or exhibited interest in game play.

5. Findings

In this section we present the results from our 2 AMT trials, deferring discussion of the findings to the following section. Table 1 presents the descriptive statistics for the data collected.

5.1 Subject demographics

A total of 323 people started the AMT task across the two trials. However 96 (30%) did not finish a game of *Happy Moths*, leaving 227 participants who actually played at least one game. Of these, 199 filled out the survey, of which we could link 185 to game play data (we paid one subject who encountered a system issue without filling out the survey). The subject pool included 67 women and 131 men (1 no response). Ages

[5] http://www.amazon.com/gp/help/customer/display.html ?nodeId=16465291

Variable	N	Mean	SD	Min	Max
How well knew how to play	195	5.46	1.43	1	7
How difficult to classify	192	3.94	1.32	1	7
How often used pop up help	194	3.81	1.99	1	7
How confident in answers	191	4.40	1.38	1	7
How much fun did you have	193	5.11	1.45	1	7
Play if:					
...not paid	97	4.18	1.76	1	7
…game connected to social media	97	3.88	1.89	1	7
…ability to compete with friends	99	4.46	2.00	1	7
…knew game helped scientists	97	5.46	1.38	1	7
Hours per week spent on:					
...single player games	191	4.62	5.20	0	40
...multiplayer games	191	5.01	13.6	0	160
Consider yourself a gamer	189	4.29	1.79	1	7
Interested in nature activities	191	4.88	1.61	1	7
How much did you learn about:					
...moths?	190	4.57	1.43	1	7
...insect classification?	190	4.60	1.45	1	7
...doing science?	193	4.12	1.60	1	7
Age	190	28.91	8.19	18	65
Games played	185	1.83	1.80	1	18
Played more than needed (1=yes)	185	0.35	0.48	0	1
Extra games played	185	0.64	1.62	0	17
Accuracy	185	0.73	0.13	0.33	1
Average score	185	53.76	15.0	10	80
High score	185	54.70	14.7	10	80

Table 1. Descriptive statistics for collected data.

ranged from 18 to 65, with an average of 29. 32 participants reported some level of high school education; 134, some level of post-secondary education; and 34, a graduate degree.

Across the two trials, participants played a total of 433 games, making a total of more than 10K classification decisions. Forty classification decisions are required to complete a single game of *Happy Moths*, but users had the option of selecting "don't know", which reduces the total classification decisions recorded. As expected, we found a skewed distribution of effort: most participants only played one game, but one person played eight games and one 18.

5.3. Data Quality

Our first research question was if users would be able to successfully classify moths following the character/state design we implemented. We obtained mixed results. The overall accuracy for all 227 players was 73%, which seemed reasonable, though not high: we would need a large number of raters per moth to be reasonably confident of the results. Looking character by character (as shown in Table 2), we see that overall accuracy was reduced by low accuracy, only 51%, for the character "forewing pattern". This character was also rated as most difficult by about 50% of users.

We used stepwise regression to identify factors that predicted an individual participant's accuracy (only for the participants who completed a survey). Though this variable is a percentage, exploratory data analysis suggested that it was more-or-less normally distributed, so for simplicity we used ordinary multiple regression. The only significant predictors for Accuracy were the variables "How well knew how to play" and "Use of popup help". Because of the exploratory nature of this research and space limitations, we will not report these results in detail.

Character/State	% Correct	# Correct	Total
Shape at Rest	*73%*	*2066*	*2831*
Arrow	79%	586	741
Parallel	36%	114	317
Spread	95%	1006	1055
Tail	50%	96	192
Tent	85%	211	249
Underside	5%	5	108
Up	28%	48	169
Forewing Main Color	*85%*	*2276*	*2678*
Black	14%	26	182
Brown	96%	916	953
Gray	93%	1029	1104
Green	57%	45	79
Orange	64%	95	149
White	78%	165	211
Forewing Distinctive Color	*80%*	*1978*	*2462*
Blue	5%	6	118
Green	38%	42	110
None	94%	1609	1714
Orange	68%	98	144
Red	70%	149	214
Yellow	46%	74	162
Forewing Pattern	*51%*	*1188*	*2346*
Banded	90%	566	629
Checkered	3%	6	193
Complex	71%	229	324
None	4%	16	360
Speckled	19%	45	240
Spot	82%	79	96
Stripe	55%	207	376
Veins	31%	40	128
Grand Total	**73%**	**7508**	**10317**

Table 2. Accuracy by character and state.

5.4. Motivation

As noted above, the fact that we paid participants to play means that we cannot truly answer our research question about motivation, as the presence of an external reward makes it hard to know whether the intrinsic reward of the game play would motivate play. Nevertheless, we do have data about the participants' actual playing behaviours, as shown in Table 3. For this analysis, we consider all 323 participants who started the AMT task, not just those who completed the survey.

First, we found that 96 (or 30%) of those who started the AMT task did not actually play even one game, while others played more than the minimum required for payment. Recall that players had to play at least once to be paid and received a bonus for a score of 50 or above. Table 3 shows that about 42% of those who played, played more than 1 game and 35% of those who played continued playing even after they received a score sufficient for the bonus, for a total of 151 extra games, about 35% of the total games played.

We used Poisson regression to attempt to predict the number of games played and logit regression to predict whether person played extra games (i.e., games past the bonus level). We found a positive relationship between the number of games played and "learned about moths", though it seems likely that the causality works in the other direction. We did not find significant predictors of playing extra games.

In addition to recording actual playing behaviour, we asked respondents on the second round if they would be interested in playing the game even if they were not paid. The average response to this question was 4 out of 7 (i.e., neutral): 47/97 answered positively, 17/97 neutral and 33/97 negatively. We used stepwise regression to identify variables that predicted a positive attitude towards playing, which identified "How much fun", "Interest in nature activities" and "Learned about moths" as positive factors and "Understood how to play" and "High score" as negative.

To identify which players might be motivated by the game, we used stepwise regression to identify factors related to participants reporting that the game was fun. We found that reported learning about classification, understanding how to play the game, confidence in answers and age were significantly related to finding the game fun.

Breakdown of Participants		
Total who started AMT task	*323*	
- Didn't play	96	30%
- Played	227	70%
Of those who played	*227*	
- Played only one game	132	58%
- Played two or more games	95	42%
Of those who played	*227*	
- Played only to get bonus	149	66%
- Played past bonus	78	34%
Breakdown of Games Played		
Total played	*433*	
- First games	227	52%
- Second game or more	206	48%
Games played	*433*	
- Only to get bonus	282	65%
- Played after bonus	151	35%

Table 3. Breakdown of players and games by times played.

6. Discussion

Our preliminary trial of *Happy Match* suggests that with some work, a "gamification" approach could produce usable data about the pictures and could be a useful tool for scientists in the biological and life sciences. Accuracy on three of the characters was high enough to be usable for creating research data.

Accuracy for the "pattern" character was reduced because several states for this character were chosen often but nearly always incorrectly (e.g. "checkered" and "none"). Accuracy would likely be improved by providing additional examples or training on these categories, e.g., by providing better exemplars and explanation of the character and state. For example, the "complex" state is found frequently within the gold standard data set, but the current example image for this state is not very representative of the kinds of moths that should normally be assigned as "complex."

It may be useful in future releases of *Happy Match* to let participants know which states are more common and which are rarer, although this could have the effect of erroneously biasing players toward some choices and away from others.

This finding has repercussions for other variants of the *Happy Match* game (e.g. *Happy Sharks* or *Happy Plants*). In all instantiations of the *Happy Match* game, care must be taken to ensure that examples and help text are clear and representative of typical good quality answers. *Citizen Sort* includes an administrative interface that enables the development team to make rapid edits to existing games. Data from a study such as the one presented in this paper can therefore have an immediate and beneficial impact on the design of the characters and states in a game, so data quality might eventually be "tuned" to desirable levels.

In the regression analysis, we found that accuracy was related to the average score and use of pop up help. It may be that participants based their assessment of how well they knew how to play on their average score, which is correlated with accuracy, though not perfectly (r=0.38). The effect of use of popup help suggests that the system help functions were useful.

Our evaluative efforts also helped us to improve our contextual and representational data quality, as well as the accessibility of our data. Round one of the AMT test revealed a variety of technical improvements that could be made in the *Citizen Sort* project database and game code to ensure a more complete and understandable data set that could be more easily queried for a variety of results. These improvements were implemented for the second round of the AMT test, and will have positive repercussions for our ability to provide data to our domain experts and achieve our own research goals. In addition, these improvements enhance the flexibility of the *Happy Match* game design, making it easier for us to improve character and state examples or add additional guides and help information.

With regards to our second research question, our analysis of motivation, while preliminary and colored by the use of a paid subject pool, suggests that some proportion of participants did find the game motivating. We do not expect that *Happy Match* (or any other video game) could ever appeal all users, so the question is whether it attracts enough to be useful.

Of the participants who played, about 60% played only once, suggesting that money was the primary or only motivation for many players. 96 of the people who started the task didn't finish even a single game, suggesting that for this group even the small payment was not sufficient motivation to play. However, 95 participants played two or more times, for a total of 206 extra plays of the game beyond the minimum required to be paid.

We did offer a bonus for achieving a certain score, and 149 people played only minimum number of games needed to obtain the bonus (i.e., they stopped when they scored at the bonus level), again suggesting the primacy of the monetary incentive. However, 78 participants (about 1/3 of the total) continued to play even after they earned the bonus; indeed, the individual who played 18 times earned the bonus on the first game. An extra 151 games, or about 1/3 of the total games played were thus essentially "volunteered." The number of volunteered games even in the presence of a monetary reward, coupled with some positive responses for how fun the game was perceived to be and likelihood to play without pay, suggests that *Happy Match* may be able to find a sufficient audience.

In our survey, participants were asked what the impact would be on their likelihood to play if they could compete with friends, with an average score of 4.46 out of 7. This suggests that a competitive element may, in fact, work as an additional motivator for *Happy Match*. The *Citizen Sort* system includes affordances for registered users to create groups of friends and compare scores. In our future evaluations, it will be interesting to more thoroughly explore how competition or cooperation can enhance the motivational impacts of purposeful games.

The stepwise regression identified the factors "How much fun", "Interest in nature activities" and "Learned about moths" as positive factors predicting a positive attitude towards the game and "Understood how to play" and "High score" as negative predictors. The first three factors have a certain face validity. Unfortunately, they do not seem to be factors that we might be able to manipulate through the design of the game, but rather factors for identifying participants to recruit, e.g., nature enthusiasts who have some intrinsic motivation to participate.

That "Understood how to play" and "High score" were negatively related to interest suggest that if the game is felt to be too easy, it will not be motivating. However, the appropriate level of challenge is a complex issue that bears further study. Games that are too difficult could be as de-motivating as games that are not challenging enough; striking an appropriate balance is a key goal. Furthermore, what is found to be challenging likely depends heavily on the particular player, posing a design challenge to reach a wide group of players. For example, many citizen science projects include elements of science outreach or education that is targeted specifically at children or novices.

Happy Match, because it can support classification on a wide variety of species with a wide range of characters and states, may provide a flexible level of challenge. Based on participant responses, *Happy Moths* was perceived by our players to be of moderate difficulty. Other games such as *Happy Sharks* or *Happy Plants* could be designed around either simpler or more difficult questions. Ideally, these various games would attract different kinds of users with different levels of expertise and preferences for level of challenge. It may also be possible to vary the level of challenge within a single game by selection of different pictures.

Finally, an additional future direction for this research is also an important overall goal of the *Citizen Sort* project: comparison between different types of games and tools. *Citizen Sort* includes two systems in addition to *Happy Match*: a classification tool called *Hunt & Gather* and a point-and-click adventure game called *Forgotten Island*. Together, these three systems represent a continuum from "tools" to "games," with *Happy Match* hypothesized to fall somewhere in the middle as a "tool-like game." We are very interested in evaluating both *Hunt & Gather* and *Forgotten Island* to explore how different kinds of users might be motivated by these different systems; *Hunt & Gather* is hypothesized to be of interest primarily to professional scientists and their assistants, while *Forgotten Island* seeks to capture the attention of "casual gamers" who may be entirely uninterested in scientific participation *per se*. We consider casual gamers to be a significant and as-yet untapped pool of potential citizen science participants, and we are anticipating interesting finding from the evaluation of *Forgotten Island* and its comparison to *Happy Match* and *Hunt & Gather*.

7. Conclusion

The next step in our research plan is to launch *Citizen Sort*, *Happy Match* and our other tools and games publically and in a natural setting, allowing them to run as a public citizen science project akin to many others currently in use. This deployment will provide a rich new stream of data with fewer limitations with regard to participant motivation. We are currently designing a marketing plan to promote *Citizen Sort* and its associated games, specifically seeking to reach expert scientists, enthusiasts and casual gamers. Deploying a marketing campaign will necessarily impact our research on participant motivation, since by their nature marketing initiatives are intended to create enthusiasm for a product or service. On the other hand, no citizen science project can be successful if it is unknown to the general public, so we consider marketing efforts to be part of our research agenda. We expect our findings on *Citizen Sort* to be of value to scientists who seek to deploy socio-computational systems and games for citizen science, and are seeking practical guidelines on this exciting and complex phenomenon.

8. Acknowledgements

This research was partially supported by NSF Grant 09-68470. We acknowledge the contributions of Susan Furest, Dania Souid, Trupti Rane, Gongying Pu, Shu Zhang, Nathan Brown, Chris Duarte, Yang Liu, Nitin Mule, Sheila Sicilia, Jimit Shah, Jessica Smith, Peiyuan Sun, Supriya Mane, Xueqing Xuan and Zhiruo Zhao, of Syracuse University and Onondaga Community College.

9. References

1. Cohn, J.P., *Citizen Science: Can Volunteers Do Real Research?* BioScience, 2008. **58**(3): p. 192-107.
2. Wiggins, A. and K. Crowston. *From Conservation to Crowdsourcing: A Typology of Citizen Science*. in *44th Hawaii International Conference on System Sciences*. 2011. Kauai, Hawaii.
3. Prestopnik, N. and K. Crowston. *Citizen Science System Assemblages: Toward Greater Understanding of Technologies to Support Crowdsourced Science*. in *iConference 2012*. 2012. Toronto, ON.
4. Hevner, A.R., et al., *Design Science in Information Systems Research.* MIS Quarterly, 2004. **28**(1): p. 75-105.
5. March, S.T. and G.F. Smith, *Design and natural science research on information technology.* Decision Support Systems, 1995. **15**(4): p. 251-266.
6. Peffers, K., et al., *A Design Science Research Methodology for Information Systems Research.* Journal of Management Information Systems, 2007. **24**(3): p. 45-77.
7. Dix, A., *Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail.* Interacting with Computers, 2010. **22**: p. 13-27.
8. Ellis, G. and A. Dix, *An explorative analysis of user evaluation studies in information visualisation*, in *Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization*. 2006, ACM: Venice, Italy.
9. Prestopnik, N., *Theory, Design and Evaluation – (Don't Just) Pick Any Two.* AIS Transactions on Human-Computer Interaction, 2010. **2**(3): p. 167-177.
10. Gregor, S., *The Nature of Theory in Information Systems.* MIS Quarterly, 2006. **30**(3).
11. Wang, R.Y. and D.M. Strong, *Beyond accuracy: what data quality means to data consumers.* J. Manage. Inf. Syst., 1996. **12**(4): p. 5-33.
12. Malone, T.W., R. Laubacher, and C.N. Dellarocas, *Harnessing Crowds: Mapping the Genome of Collective Intelligence*, in *MIT Sloan Research Paper No. 4732-09*. 2009.
13. Bradford, B.M. and G.D. Israel, *Evaluating Volunteer Motivation for Sea Turtle Conservation in Florida*. 2004, University of Florida, Agriculture Education and Communication Department, Institute of Agriculture and Food Sciences: Gainesville, FL. p. 372.
14. King, K. and C.V. Lynch, *The Motivation of Volunteers in the nature Conservancy - Ohio Chapter, a Non-Profit Environmental Organization.* Journal of Volunteer Administration, 1998. **16**(5).
15. Raddick, M.J., et al., *Citizen science: status and research directions for the coming decade*, in *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey*. 2009. p. 46P.
16. Raddick, M.J., et al., *Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers.* Astronomy Education Review, 2010. **9**(1): p. 010103-18.
17. Wiggins, A. and K. Crowston, *Developing a conceptual model of virtual organizations for citizen science.* International Journal of Organizational Design and Engineering, 2010. **1**(1/2): p. 148-162.
18. von Ahn, L., *Human computation*, in *Proceedings of the 4th international conference on Knowledge capture*. 2007, ACM: Whistler, BC, Canada.
19. von Ahn, L., et al., *Improving accessibility of the web with a computer game*, in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2006, ACM: Montreal, Quebec, Canada.
20. Law, E. and L. von Ahn, *Input-agreement: a new mechanism for collecting data using human computation games*, in *Proceedings of the 27th international conference on Human factors in computing systems*. 2009, ACM: Boston, MA, USA.
21. Garris, R., R. Ahlers, and J. Driskell, *Games, Motivation, and Learning: A Research and Practice Model.* Simulation & Gaming, 2002. **33**(4): p. 441-467.
22. Malone, T.W., *What makes things fun to learn? heuristics for designing instructional computer games*, in *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems*. 1980, ACM: Palo Alto, California, United States.
23. von Ahn, L., *Games with a purpose.* Computer, 2006. **39**(6): p. 92-94.
24. von Ahn, L. and L. Dabbish, *Designing games with a purpose.* Commun. ACM, 2008. **51**(8): p. 58-67.