Crowdclass: Designing classification-based citizen science learning modules

Doris Jung-Lin Lee
University of Iliinois Urbana-Champaign
jlee782@illinois.edu

Joanne Lo, Moonhyok Kim, Eric Paulos
University of California, Berkeley
{jlo,moonhyok,paulos}@berkeley.edu

Abstract

In this paper, we introduce Crowdclass, a novel method that integrates the learning of advanced scientific concepts with the crowdsourcing microtask of image classification. In Crowdclass, we design questions to serve as both a learning experience and a scientific classification. This is different from conventional citizen science platforms which decompose high-level questions into a series of simple microtasks that require no scientific background knowledge to complete. We facilitate learning within the microtask by providing content that is appropriate for the participant's level of knowledge through scaffolding learning. We conduct a between-group study of 93 participants on Amazon Mechanical Turk comparing Crowdclass to the popular citizen science project Galaxy Zoo. We find that the scaffolding presentation of content enables learning of more challenging concepts. By quantifying the trade-off between user motivation, learning, and performance, we draw general design principles for learning-as-an-incentive interventions applicable to other crowdsourcing applications.

Introduction

Citizen science (CS) enables amateur scientists to make valuable contributions to tasks such as text transcription and image classifications. Over the past five years, online CS platform have grown in popularity, attracting citizen scientists of all ages and skills to help drive scientific discoveries. Even though CS is commonly regarded as a form of scientific outreach, recent studies have shown that learning in CS platforms often happens outside the context of the crowdsourcing task (Jennett, Kloetzer, and Schneider 2016; Iacovides et al. 2011). Existing CS project designs present very little scientific background to the users, in order to minimize the latency of getting the users started on the task. These designs are motivated by the concern that most participants have a notably short attention span (Eveleigh et al. 2014) and the initial barriers of learning both the game mechanics and project content can be discouraging for starters. As a result, most project tutorials contain minimal scientific content.

Moreover, popular online CS platforms, such as Eyewire[1] and FoldIt[2], often leverage domain-specific elements (e.g. coloring outlines of microscopy cross sections, protein folding visualizations) to gamify the crowdsourcing microtask. Since the learning content of these platforms are highly correlated to the game mechanics, the amount scientific content that could be learned from these framework is not very extensible. Additionally, the content generation step for these Games with a Purpose (GWAPs)(von Ahn and Dabbish 2008) is expensive and time-consuming for the scientist and focuses more on making the task fun to encourage returning contributions, rather than making learning the outcome.

To address these existing limitations, we applied Jennett et al.'s (2016) theoretical work on learning and motivation in CS projects to build Crowdclass, a CS framework that integrates content learning with the crowdsourcing microtask of image classification. In Crowdclass, users' responses serve as both a learning quiz and a classification task. We perform a between-group study to evaluate user performance, motivation, and learning in Crowdclass against a control study. This is the first CS project that proposes a generalized framework for combining learning and crowdwork at the microtask level. We propose two different user contribution models and the use of learning as an incentive for increasing participants learning and motivation. Our work not only opens the learning-within-microtask design space in future CS platforms, but we also draw general design principles for learning-within-microtask applications that can be extended to other crowdsourcing contexts.

Related Work

Galaxy Zoo

Galaxy Zoo (GZ) is a popular CS project where users help classify telescope images of galaxies according to their morphology. The volunteers answer a series of question about a particular image of a galaxy following a decision tree workflow. These classification results are useful for astronomers to conduct population studies of galaxies and characterize the relationship between galaxy morphology and their stellar compositions (Willett et al. 2013). The volunteer's scientific contribution extends beyond just labeling images. For example, in 2007 a Dutch schoolteacher pioneered a citizen-led discovery when she reported an anomalous image of a photo-ionized cloud while sorting through the GZ images, which generated considerable scientific interest in the theory

[1] eyewire.org

[2] fold.it

of black hole accretion at the center of these galaxies (Keel et al. 2012). Part of the inspiration for Crowdclass was recognizing the inherently meaningful connections between the crowdsourcing microtask and the scientific knowledge derived from these classes. As GZ is an example of a successful CS project, in this paper, we will be applying the GZ model as a particular use case of Crowdclass.

CS Learning & Motivations studies

In Jennett et al.'s (2016) extensive studies of volunteer motivation, learning and creativity based on a series of exploratory interviews with citizen scientists and researchers, they identified six types of learning commonly found in CS projects: task mechanics, pattern recognition, on-topic learning, scientific process, off-topic knowledge and skills, and personal development. Iacovides et al.(2011) distinguished between micro- and macro-levels of engagement in online CS projects, where micro learning happens when the user is directly engaged with the crowdsourcing microtask and macro learning originates from external sources, such as social forums and project documentations. Since most of the content learning happens informally at the macro level, they found that learning is correlated with the volunteer's level of engagement, rather than quantitative measures of the amount of time spent on the project or the number of microtasks completed.

Inspired by this body of work, we designed a system that optimizes participants' learning and motivation by enabling volunteers to learn something new as they complete each crowdsourcing microtasks. While community and social aspects of citizen science are highly valuable, in this paper, we focus on designing a system that enables formal, on-topic, content learning at the microtask level.

In our survey of existing online CS projects, we observed that many CS platform designs are based on research in the crowdsourcing communities, often performed on Amazon Mechanical Turk (MTurk)[3]. For example, Tinati et al. (2015) advocate designing tasks that make it hard to skip a question, in order to encourage participants to make their best guesses. Other examples include decomposing a high-level, scientific question to a series of simple tasks, as part of a decision tree and designing closed-ended, easy-to-answer responses that are solely based on the geometry, color, shape of the object in the image. While crowdsourcing on MTurk and CS share many similarities, applying these MTurk crowdsourcing approaches to the design of CS platforms is problematic since the motivations of citizen scientists is vastly different from motivation of MTurk crowd workers. Common motivations to participate in CS include a sense of contribution, learning, discovery, fun and interest in subject area (Iacovides et al. 2013; Raddick et al. 2013; Rotman et al. 2012), whereas MTurk workers are mostly driven by monetary rewards (Harris 2011; Marcus and Parameswaran 2015). Despite this difference in motivation, many CS platforms still employ MTurk crowdsourcing metrics for optimizing crowdworkers' speed, latency, and accuracy.

[3] www.mturk.com

Learnsourcing applications

Recent works that combined crowdsourcing and learning have pursued three related directions: 1) crowdsourcing content generation for online education platforms, 2) on-task learning for crowdsourcing complex creative tasks, and 3) integrating learning into the design of a specific microtask. A common thread among these HCI applications is the notion of "learnsourcing" (Kim 2014), the idea that learners can act as the crowd that collaboratively improves the educational content and interface based on their interaction with the platform and thereby enriching their own learning experience. Existing application areas includes passive learnsourcing (where the user's clicks and interactions are monitored and used as feedback to improve the educational platform (Kim 2014)) and active learnsourcing (which focuses on integrating learning into the design of the microtask, such as summarizing or outlining key concepts in the lecture videos (Weir et al. 2015)).

From the requester's perspective, another motivation for creating platforms that integrate crowdsourcing and learning is to address the lack of a skilled MTurk workforce. In addition, it is difficult to evaluate the quality of complex, creative tasks and provide automated feedback. Existing work in this area include general frameworks for decomposing complex work into basic, independent microtasks (Kittur, Smus, and Kraut 2011; Little et al. 2009) and techniques for embedding learning components inside the crowdsourcing frameworks (von Ahn 2013; Dontcheva et al. 2014). For example, Dontcheva et al. (2014) developed LevelUp, an interactive plugin for Adobe Photoshop to train crowdworkers on the software mechanics and sharpen their photo-editing skills. Another example is Duolingo[4], a language-learning platform which make use of foreign vocabulary quizzes results to assist the crowdsourcing task of web translation. Both Duolingo and LevelUp are examples of how successful designs of learnsourcing frameworks can transform a complex crowdsourcing task (language translation, photo-editing) into a learning task for the crowdworkers. Inspired by these prior work, we sought to design a framework that 1) enable scientific concept learning within the crowdsourcing task of image classification, 2) increase user engagement and motivation, with 3) a generalizable content development workflow.

System Design

Decision Tree The program workflow is described by a decision tree that guides the users through a series of questions to collect information about a single image. At each level, the users are given a specific task with a set of possible responses. Once the users reach the bottom of the tree, they begin to classify a new image starting from the top of the tree. As shown in Fig.1, the Crowdclass decision tree is a modified version of the GZ4 DECaLS decision tree[5] and contains 9 tasks and 18 possible responses.

[4] duolingo.com

[5] data.galaxyzoo.org/gz_trees/gz_trees.html

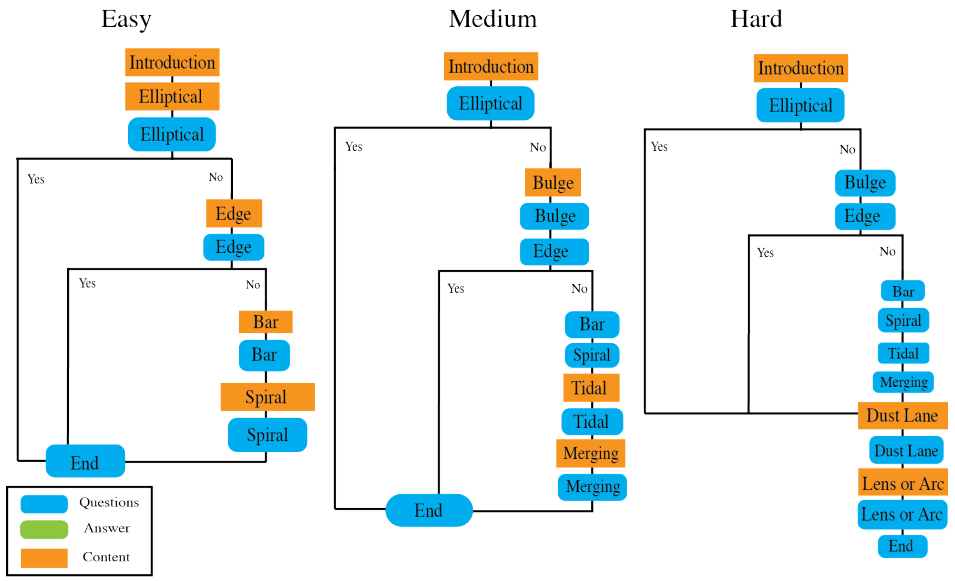

Figure 1: Decision Tree for Crowdclass.

Content We propose a simple, general workflow to minimize the amount of work for experts to generate the content and questions used in Crowdclass. First, the expert lists all the facts associated with a particular classification type, then she selects the widely-known facts that are unique to that class. Next, the facts are translated to a form of a True/False question where the answer to the quiz that indicates membership of a class is noted in the system. The facts are then used to generate hints and reading contents, while the questions are used for the classification microtasks. Future work includes automating this content-generation process by applying methods from natural language processing and information retrieval research. The educational contents used in this study have been reviewed by professional astronomers and educators to ensure the scientific correctness of the learning modules.

Tinati et al.'s (2015) study of Zooniverse's initial project design showed that a large number of users left within 90 seconds of first visiting the site. Even though videos are more interactive and richer in delivering content, the time commitment to watch the video may cause users to drop out of the project altogether. Such observations were noted during the design evaluation of the CS project Virtual Atom Smasher (Charalampidis, Segal, and Skands 2014). Therefore, we chose to use short text, rather than animations or short videos, for delivering the learning content.

Scaffolding Learning In Crowdclass, we decompose GZ decision tree workflow into three subtrees of different difficulty levels. This design was motivated by the problem that having to learn all the scientific content associated with a single classification can be overwhelming to novice users. Each subtree acts as a different learning module with a focused learning objective. The "easy" learning module introduces the scientific goals for studying galaxy morphology and the basic properties associated with elliptical and spiral galaxies. The "medium" level learning module covers galaxy dynamics (how galaxies interact in a collision and how that impacts their morphology and formation). The "hard" learning module focuses on the anomalous features in galaxies, such as gravitational lenses and dust lane galaxies. As the users progress through the program, they "unlock" the more advanced learning modules after reading all the contents available at that level.

This scaffolding design is motivated by Wood et al.'s (1976) theory of scaffolding learning that advocates strategies for experts to facilitate learning of skills and knowledge that is initially beyond a student's zone of proximal development. While the theory was initially developed for training motor task skills in toddlers, scaffolding learning has been successfully applied to computer-supported collaborative learning contexts(Kahrimanis, Avouris, and Komis 2011). Here, we use the scaffolding design to enable formal learning of more advanced scientific concepts.

Assistive Tools The existing GZ interface provides cue words and example pages as assistive tools for guiding the users through the classification tasks (Fig.3). Cue words are short texts that remind the users about the category they are currently classifying. Example pages contain images with brief texts showing what an object that falls under a specific class looks like. These facilitate learning of project mechanics for novice users, but are not intended to convey any scientific content knowledge.

Other than facilitating better classification results, these assistive tools can provide additional pedagogical value when learning is incorporated in the microtask. For example in Duolinguo, when a user is prompted the translation task of "the apple" in French, they are shown an image of an apple and an audio clip of "la pomme" is played back. So even

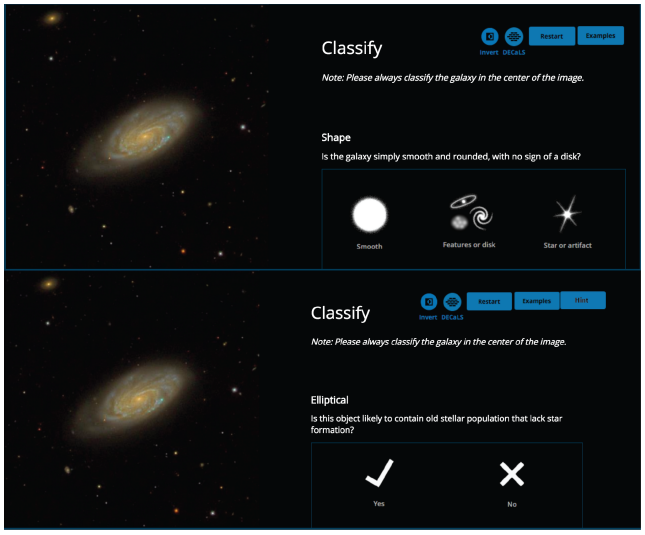

Figure 2: Example of the software interface for Group A (Top) and Group B(Bottom) for the same question and image.

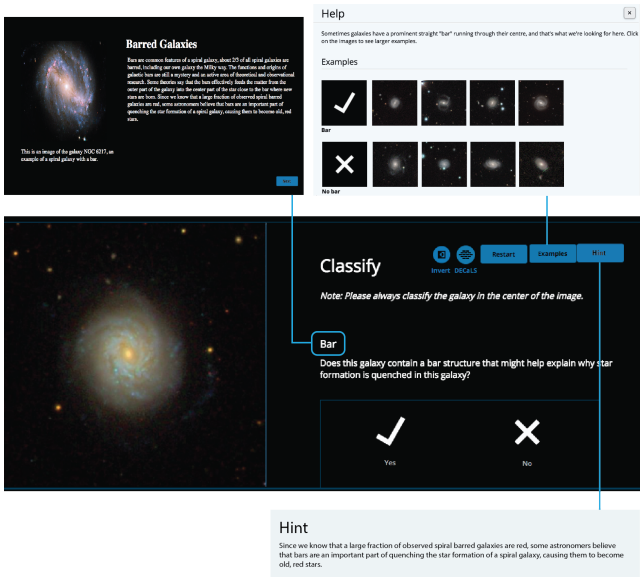

Figure 3: Examples of tools used for guiding users through the Crowdclass tasks. Here, "Bar" is the cue word that reminds the users about the content that they read.

if the users did not know the translation beforehand, they can still chose the right translation and in turn learn from the microtask. Likewise, cue words serve a similar role in Crowdclass to reinforce the learning of concepts associated with a particular class of objects. In Crowdclass, we also provide customized hints associated with each question to remind the users about the concept that they are trying to learn.

Technical Details The system is developed on Django[6] with CSS/HTML frontend hosted on a remote server. The software design is intended to be as generalizable as possible. Particular emphasis has been made on what are the minimum requirement required to build a working learning module. In the future, this will enable scientists to upload their data and create their own questions and learning modules with ease. Both the code for building the Crowdclass framework and for performing the statistical analysis described in this paper are open-source[7].

Method

Participants We recruited participants on MTurk over the month of May 2016 through a pre-screening HIT. To ensure work quality, we required that participants have a HIT approval rate of greater than 95% and have at least 50 HITs approved in their work history. To prevent language barrier as a confounding factor in comprehending the learning content, we required that the participant's primary language is English. The pre-screening HIT is separate from the main experiment and consists of a 3-minute survey where workers are paid $0.01. The user demographics is summarized in Fig.4.

Procedure We conduct a single-factor, two-level, between-group study that consists of five main steps in the following order: pre-screen survey, pre-test, main experiment, post-test, and post-study survey. The between-group study design prevents concepts and skills acquired from a previous trial from confounding the measurements of a subsequent study. We compensate the potential effect of learning variability by verifying that the measured and self-reported knowledge in astronomy in both populations are similarly distributed. As some astronomical images can be harder to classify than others, we prevent task difficulty as an additional confounding factor by displaying the same sequence of image tasks for both conditions.

The pre-test questions consist of 10 multiple choice questions displayed in random order. Each question contains four answer choices and an additional "I don't know" option. Six of the questions are fact-based questions, while the remaining four are "hybrid" question, which requires the participant to synthesize knowledge learned from multiple different learning modules that they have encountered. Correct responses to these questions demonstrates subject understanding beyond simple recall of facts. While participation in CS projects can often prompt users to do exploratory learning on their own (Masters et al. 2016; Kloetzer 2015), we chose to control the experiment by asking participants to answer the pre/post test questions without relying on external references (books, web ..etc.) to ensure that the test results only measures the learning that happens within the microtask.

After the pre-test, participants in Group A interact with a mockup of the GZ web interface and Group B participants interact with Crowdclass (Fig.2). During the experiment, we record the timestamps and choices performed in each mi-

[6] `djangoproject.com`
[7] `github.com/dorisjlee/crowdclass`

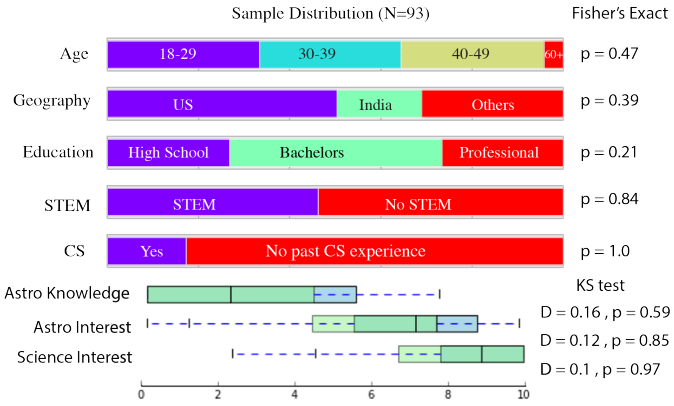

Figure 4: To ensure that the demographics of the two groups are similar, we conducted the Fisher's Exact test for the categorical variables and the Kolmogorov-Smirnov test for the 10-point, Likert-scale responses to show that the two groups are similar.

crotask as well as the number of times the assistive tools were used. At the end of the 30-minute study, the participants are prompted a set of post-test question (identical to the pre-test) and a post-survey. The participants are compensated an additional $1 for completing the main experiment and the post-survey. Since our pilot study showed that participants can experience fatigue and performance degradation after 40 minutes, the length of study was limited to 30 minutes including the pre-test, which takes on average 6.4 minutes to complete. We were initially concerned that the shorter experiment duration may not yield enough time for participants to reach the medium and hard difficulty levels to scaffold the content learning, but our experiment actually finds that 89% of the users reached the medium level and 80% reached the hard level.

User Study

We received responses from 506 workers; some workers were excluded from the analysis presented in this section for the reasons:

- 73 were excluded based on the language pre-screening requirement.
- 267 chose not to participate in the the main experiment.
- 15 in Group A and 12 in Group B dropped out of the main experiment.
- 17 in Group A and 29 in Group B were excluded based on our work quality filter. [8]

After these filters, there was 46 workers in Condition A and 47 workers in Condition B used for the analysis. Fig.4 shows that the demographics of the two groups are simmilar.

[8]Workers who classified over 95 or under 5 images in the 30 minute experiment were not included in the analysis, the cutoff was heuristically determined based on our pilot study results.

Usage of Assistive Tools While the usage of assistive tool varied largely depending on the individual, we use the number of times the examples and hint page is clicked as an indicator of how engaged the participants are within the program. The number of times the restart button is pressed additionally serve as a measure of the participant's perceived level of uncertainty regarding their classification results.

The hint page was accessed 5.59 times on average (SD = 8.21). The Mann-Whitney U test showed that Group B participants clicked on the example pages significantly more times than Group A participants and that Group B participants are significantly more confident in their responses than Group A participants (Table.1). Overall, we find that 26% of participants in group A and 72% in group B used at least one assistive tool once throughout the whole experiment.

RQ1: Can crowdworkers learn within microtask ?

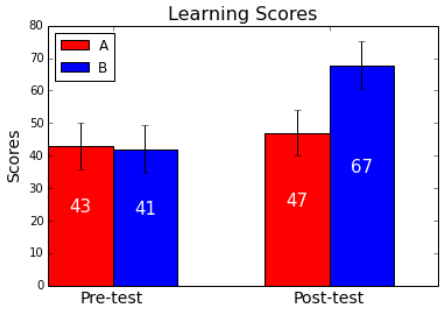

Figure 5: Bar chart comparison of pre/post test scores with error bars indicating 95% CI.

To quantify the amount of learning that happens within the microtask, we compute the overall score of the pretest and the post-test, excluding the responses where users selected "I don't know" option (`uncertain`). The Kolmogorov-Smirnov test to show that the initial pre-test scores ($D = 0.13$, $p = 0.83$) and the number of `uncertain` counts ($D = 0.23$, $p = 0.16$) came from the same distribution. As shown in Table. 1, the Mann-Whitney test showed a significant effect in the post-test scores with Group B outperforming Group A. The participants in Group B also self-reported higher levels of learning in astronomy than in Group A. There is no significant difference between learning of the scientific process in both conditions.

We also conduct an analysis on the `uncertain` counts as a measure of how confident the users are about the content that they learned in the main experiment. To calibrate for the large variation in guessing behavior among the participants, we compute the change in the the number of uncertainty choices for each individual user instead of comparing the aggregate of the whole group to compare how the individual `uncertain` counts change before and after the experiment. Table.1 shows that Group B participants are significantly more confident in their responses than Group A participants.

Table 1: Mann-Whitney U test results for self-reported learning and motivation measures.

	U	Z	p	r	median		mean rank	
					A	B	A	B
Example count	1318	3.13	0.0018	0.32	0	0	51.15	41.96
Restart count	848	-2.88	0.0041	0.44	0	0	41.93	51.96
post-test score	436.5	-3.99	6.77×10^{-5}	0.44	40	68.3	31.41	52.58
`uncertain` count	134	-5.42	4.09×10^{-9}	0.66	1	0	57.10	37.12
This software helped me learn something new about astronomy.	145.5	-2.06	0.039	0.31	8	9	19.39	27.44
This software showed me what the scientific process is.	297	1.64	0.10	0.25	8	7	25	18.53
I have a greater interest in astronomy after using the program.	264	0.83	0.41	0.13	9	7	23.76	20.5
This software increased my motivation to participate in other citizen science projects in the future.	255	0.63	0.54	0.09	9	8	23.44	21
I feel a sense of contribution to science from using the software.	247	0.43	0.68	0.06	8	7	23.15	21.47

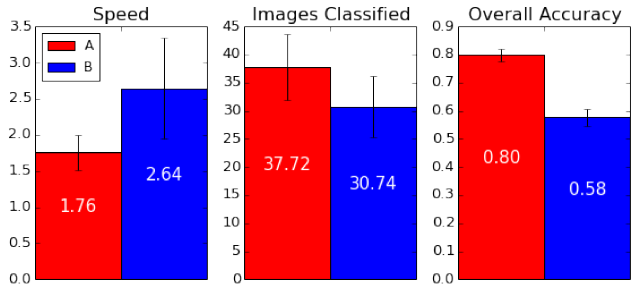

Figure 6: Bar chart comparison of performance measures with error bars indicating 95% CI.

Table 2: Mann-Whitney U test results for performance measures.

	U	Z	p	r	median		mean rank	
					A	B	A	B
`crowd`	2052	7.46	2.2×10^{-16}	0.77	0.81	0.61	68.11	26.34
`expert`	2088	7.74	1.04×10^{-14}	0.80	0.80	0.60	68.89	25.57
Bar	1452	4.85	1.26×10^{-6}	0.53	0.9	0.5	56.40	30.51
Bulge	1725	7.30	2.65×10^{-16}	0.78	1.0	0.68	61	26.43
Dust lane	1396	6.70	1.78×10^{-14}	0.75	0.85	0.36	53.84	18.89
Edge-on	389	-5.32	2.28×10^{-8}	0.55	0.54	0.83	31.95	61.73
Elliptical	1603	4.01	4.13×10^{-5}	0.42	0.72	0.6	58.34	35.91
Lens	1338	6.64	1.01×10^{-13}	0.76	1.0	0.33	52.59	18.84
Merging	1233	2.78	0.0052	0.30	0.96	0.75	50.30	35.68
Spiral	1350	3.94	5.42×10^{-5}	0.43	0.90	0.72	53.91	32.83
Tidal	1229	2.43	0.014	0.26	0.80	0.56	50.22	37.02
Speed	683	-1.89	0.059	0.20	1.71	2.01	38.35	48.49
Img classified	1328	1.90	0.058	0.20	35.5	26.0	52.36	41.75

RQ2: How does learning within microtask affect worker performance?

Accuracy Since there are no ground-truth labels for the dataset used for the image classification task, we used two different set of labels to measure user accuracy. The first set of data is the morphological classifications of main-sample spectroscopic galaxies from GZ2 (Willett et al. 2013)[9], which contains aggregate responses of 40∼50 citizen scientists for each image(`crowd`). We use the consistency-weighted fraction of votes for all the responses and threshold of 0.5 to determine membership of a class. Since the subclasses are mutually exclusive, in the cases where GZ contain multi-way selections, we sum the independent probabilities to binarize the results. The second set of data used for comparison is labeled by a professional astronomer directly in the Crowdclass schema(`expert`). The crowdsourced results generally converges with the expert classifications. Both comparisons and individual analysis of each microtask compared against the `crowd` results showed that Group A is significantly more accurate than Group B, as summarized in Table. 2. When examining time series plots of workers' accuracy, we found no significant learning effects of participants getting better at the classification task overtime for both groups.

Speed We use the timestamps recorded at the start of every microtask question to compute an average classification speed (number of images classified per minute). While our hypothesis was that Crowdclass questions may bear an additional cognitive load to the users, which would result in slower classification speed, Mann-Whitney U test showed no significant effect in the speed. We think this result might be partly attributed to the fact that binary selections in Crowdclass are easier to chose from than the multi-way selections in the GZ interface. Excluded from the classification speed is the overhead time spent on reading the content pages, which took on average 13.12 seconds per page (SD =10.64). Table. 2. shows that there is no significant difference between the number of images classified between the two groups.

[9] `data.galaxyzoo.org/data/gz2/zoo2MainSpecz.txt`

RQ3: How does user motivation change when learning is incorporated in the microtask ?

The 10-point Likert scale post-survey questionnaires ask the participants how the program changed their interest in astronomy, whether they feel a sense of contribution to science from using the software, and how the program affected their motivation to participate in other CS projects in the future. As shown in Table. 1, a series of Mann-Whitney's test showed no significant effect (p>0.05) on the self-reported, quantitative measures of motivation and learning.

Qualitative analysis

The qualitative post-survey questionnaires reveal important insights into participants' learning and motivation in the two experiments. The thematic analysis results are summarized in Fig. 7.

Most participants in Group A cited that being exposed to the subject area through the program workflow ("There

were so much different classification [sic] to learn and decide. I didn't know there were so many types of galaxies.") and pattern recognition ("I learned by repetition, and I gained a 'feel' for the dimensions of the bulge.") as their main source of learning. Many were also inspired by the images in the classification tasks and wanted to learn more. However, some participants expressed that "not knowing specific details and definitions of wording" for conducting the classification task is discouraging:

- "The program made me feel inadequate in the subject. I felt like I needed to do research in order to answer the classifications [...] I thought I was out of my league."
- "It was difficult enough to make me not want a science career. I am not even sure I accomplished it correctly. I felt defeated half way through."

While Group B participants also noted that the task was challenging and contributed to a source of confusion, they felt that the tasks and readings motivated them to continue on and learn more:

- "I was really motivated and thought really hard before answering each question. This motivation came from the extreme difficulty of the task at hand."
- "The task and the quiz were engaging and challenging, so they were very interesting, but they were easy enough that I never felt like I wasn't doing well or like I didn't want to continue the task."

When asked about how the difficulty of the task and content affected their performance, interest and motivation, they noted that the act of repeatedly applying what they have learned helped with knowledge retention:

- "The program was very helpful for learning. Because I was immediately applying the information that I learned in order to classify galaxies, I feel like I retained the information very well."
- "I liked being able to learn in an interactive manner and have little breaks from reading."
- "I liked that after every explanation, the program went back to questions about former subjects, so I didn't forget what I learned a minute ago."

While none of the participants explicitly discussed the scaffolding mechanism, some expressed how they enjoy "[going] from not knowing much to knowing some stuff" at each difficulty level and how the interactivity of the learning experience was "much better than using a dry textbook."

Discussion

Learning-within-microtask interventions

We apply Law et al.'s (2016) findings on curiosity intervention in crowdwork to suggest a possible explanation for participant motivation observed in the Crowdclass experiment. They find that curiosity interventions that reveal too much information satisfied the user's curiosity too early on, causing the intervention to be ineffective. However, revealing too little information discourages the effect of curiosity altogether. There is an ideal "tipping point" where the curiosity intervention is most effective for inducing the workers to complete the crowdwork.

While our study does not investigate user quitting behavior in detail, we find a similar trend in user motivation where the difficulty and amount of learning content characterize the information gap. When group A participants are shown simple geometry-based questions, they complain that the study is too long and boring. On the other hand, when they are given questions containing challenging technical terms without any context to understand these terminology, the participants are discouraged from doing the crowdwork. This latter case is especially seen as the group A participants find it difficult to classify gravitational lens, tidal debris, and dust lane galaxies which are less obvious classification categories compared to geometry-based categories such as elliptical or spiral. By using the scaffolding design in Crowdclass, we are able to show the participants just the right amount of information to sustain interest and motivation in learning more about the subject, while they are rewarded by reinforcing their new-gained knowledge for every microtask they complete.

Our study finds that while learning-within-microtask applications increases learning and motivation, participant's classification accuracy is lower than in conventional CS project designs. As Rotman et al. (2012) explain that motivation is "highly dependent on scientists' attitudes towards the volunteers – when scientists acknowledged the need to educate people and not 'treat them as dumb' ", we also find that participants regard the challenge associated with the learning as an incentive for motivating them to continue working on microtasks when they are provided sufficient context and background information. These multifaceted factors of motivations characterize the more global picture for the CS project in terms of the tradeoff between scientific productivity and learning, as described by Cox et al.(2015) and Masters et al(2016).

Design Space

After understanding the motivation factors and tradeoffs in learning-within-microtask designs, we consider how these findings can be applied to different contexts by proposing two potential use cases.

"Citizen Learners" for microvolunteering Crowdclass can be used as an educational software in a supervised classroom setting or for independent, curiosity-driven learners. Our learning-within-microtask design promotes a new view of citizen science platforms as reference modules, where the users' objective is to learn something quickly or to gain a general overview of a topic area.

These"citizen learners" differs from the typical GWAP model where gamification elements encourage contributions from returning users. Instead, these citizen learners can be seen as "dabblers", a term coined by Eveleigh et al. (2014) to describe a large percentage of contributors on existing CS platforms who are low-commitment, low engagement volunteers, often too busy to find time to commit to the project, but remain interested in the area. By offering learning as an incentive, Crowdclass make the crowdsourcing task more

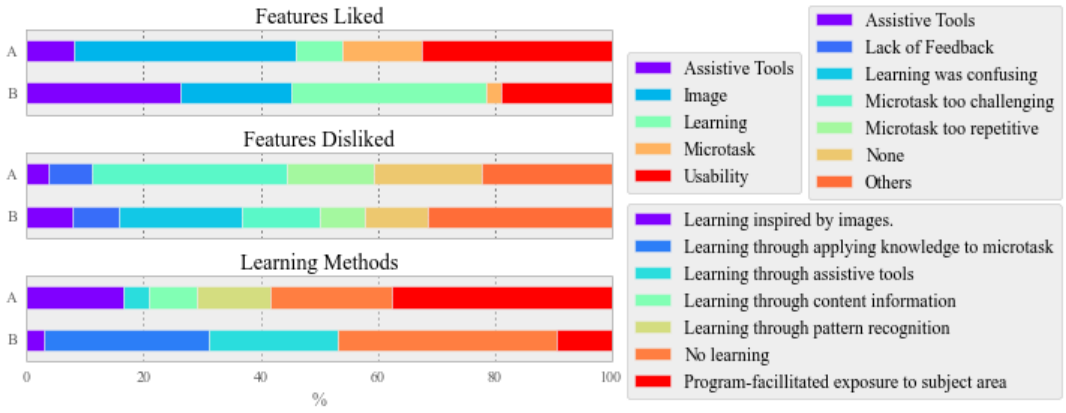

Figure 7: Theme frequencies from three survey questions.

worthwhile from the participant's perspective.

Dabbler participation through microvolunteering can be facilitated by embedding Crowdclass microtasks within traditional online reference learning sources. For example, the archival Zooniverse project Cell Slider[10] where volunteers were asked to label tissue microarray (TMA) images of cancer tumor cores. If each person who reads a Wikipedia page on cancer is prompted learning-within-microtask questions to apply the concepts that they have just read on cancer to score the TMA images, then the same amount of crowdwork completed in Cell Slider's four-year science campaign would have only taken a couple months to complete.

This boost in scientific productivity can not only accelerate the scientific discovery progress, but also has potential for attracting a more diverse population than in existing citizen science platforms. Recent studies have shown that most volunteers have already demonstrated an existing interest in science even before participating in the CS project (Rotman et al. 2012; Raddick et al. 2013). In addition, since CS projects are often advertised through scientific outreach outlets such as popular science magazines or social media channels, these platforms are missing participation from a large part of the "non-science" population, who would benefit the most from these CS experience. By prompting the users with learning-within-microtask crowdwork in a low-barrier, self-rewarding setting, we hope that the learning and sense of contribution from accomplishing the microtask would inspire further engagement in the CS project and sustained interest in science.

Feedback-facilitated active learning In our qualitative analysis, participants expressed that the lack of feedback on whether a classification is done correctly or not can often be a source of frustration, which can lead to dropout behaviors. While most existing CS platforms do not provide active feedback for the users, we feel that this is a crucial part of learning-within-microtask applications that needs to be addressed in future work. As many existing studies have shown that feedback can help sharpen crowdworkers' skills and yield better work quality (Dow et al. 2012; Glassman et al. 2016), feedback mechanisms for identifying common fallacies in learners' understanding or mistakes in their classifications are especially important when learning is incorporated in the microtask.

One exciting area of future work is to apply active learning to scaffold both image classification skills and content learning within the microtask(Costa et al. 2011; Branson et al. 2010). Current state-of-the-art machine learning algorithms are capable of classifying galaxy morphology at above 90% accuracy, but their results reflect the inherent errors due to human classifiers in the GZ training set(Ferrari, de Carvalho, and Trevisan 2015). Some machine learning algorithms such as Support Vector Machines and Linear Discriminant analysis can additionally generate an uncertainty value associated with the labels, which can be used as difficulty estimates of how easy it is to classify a particular image. These uncertainty values will enable us to display the easy-to-classify images to novices and use these high-accuracy, machine-generated labels to provide a automated feedback mechanism to guide novices.

Feedback are not only important for novices, but essential for training expert classifiers. For long-time, committed contributors, known as "super-volunteers" in (Eveleigh et al. 2014), their objective is to make valuable contributions to a CS project. Using the Crowdclass scaffolding mechanisms, as these super-volunteers progress up the difficulty levels, they will be given harder tasks and learning content over time. In doing so, the users are kept motivated as they learn about the more challenging concepts suited at their levels and become skillful enough to distinguish the more ambiguous images that are often misclassified by machine learning algorithms.

Different CS projects can benefit from the two different modes of engagement. For example, GWAPs such as FoldIt and EteRNA[11] encourage the supervolunteer model, since supervolunteers often become experts at the project mechanics and submits high-quality crowdwork. This contribution model aligns with the science goal of the project: finding the functionally-best protein folding, rather than collecting many

[10] cellslider.net

[11] eternagame.org

possible different protein-folding configurations. On the other hand, in many Zooniverse projects where the microtasks are largely independent and easy-to-learn, the science goal (e.g. classifying 90,000 galaxies) can be achieved, whether the platform attracts a large number of dabblers, a few super-volunteers, or a hybrid of both.

Limitations

As discussed in an earlier section, while CS and MTurk share similarities, there are notable differences in the participant motivations. Even though we recognize that recruiting our participants on MTurk may threaten the external validity of our experiment, it was our best option for gathering large numbers of participants to test our primary hypothesis of whether learning occurs under these interventions. While we tried to compensate these effects by filtering out users with malicious intents, this population difference may contribute to why we did not see a difference in motivation between the two groups from the Likert scale responses. We suspect that since the participants may not have fully understand the connection between citizen science, they did not feel a very strong sense of contribution to science from both interfaces. Our hypothesis is that if this experiment was deployed on a CS platform such as Zooniverse, we would expect to see a more notable differences in participant's interest and motivation.

One limitation to our existing implementation is that all the questions in Crowdclass are True/False questions. We have not yet been able to incorporate learning within microtasks involving multiple choices with numerical or categorical outputs. Our future work includes developing suitable methods for extending microtask learning beyond binary classification.

Conclusion

In this paper, we present Crowdclass, a novel, generalizable system for combining learning and crowdwork at the microtask level. Crowdclass prompts the users with scientific knowledge about the types of object that they are classifying and quizzes the users on these content to obtain the classification results. Using a scaffolding mechanism, we are able to show them the appropriate learning modules suited to their knowledge level with focused learning outcomes. We conducted a between-group user study with 93 participants where we compared user's performance, learning, and motivation in Crowdclass to the existing GZ interface. While we find that Crowdclass enables users to learn significantly more about the science involved in the project than in GZ, the classification accuracy of Crowdclass users is also significantly lower than the GZ users. This is potentially due to user's confusion resulting from the cognitive overhead of trying to learn and classify at the same time. The issue can potentially be alleviated if feedback is provided to the users to correct learners' misunderstandings.

We did not by find any significant differences in worker's efficiency (speed and number of microtasks completed) between the two conditions. The qualitative results showed that the challenge of the task and reading motivated the Crowdclass users to continue working on the microtask. Crowdclass users also find that the act of applying their newly-gained knowledge to classify the galaxies helped them retain the information better.

Crowdclass offers the HCI community a novel approach to citizen science, where content learning is presented as an incentive for user motivation and closely integrated with the design of the microtask. From our study results, we draw general design principles for learning-within-microtask applications that are applicable to other crowdsourcing contexts. Future directions of our work include leveraging task difficulty generated by machine learning systems to provide a mechanism for guided feedback and scaffolding learning.

Acknowledgments

The authors would like to thank Charlene Jennett, Laure Kloetzer, Fabrico Ferrari, Bjoern Hartman, Kristin Stephens-Martinez, Cesar Torres, and the annonymous reviewers for many useful inputs and helpful discussions. The work is supported by NSF grant IIS-1451465.

References

[Branson et al. 2010] Branson, S.; Wah, C.; Schroff, F.; Babenko, B.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6314 LNCS(PART 4):438–451.

[Charalampidis, Segal, and Skands 2014] Charalampidis, I.; Segal, B.; and Skands, P. 2014. Initial prototype of "Virtual Atom Smasher" game. `https://sciencenode.org/feature/virtual-atom-smasher-lhchome.php`.

[Costa et al. 2011] Costa, J.; Silva, C.; Antunes, M.; and Ribeiro, B. 2011. On using crowdsourcing and active learning to improve classification performance. *International Conference on Intelligent Systems Design and Applications, ISDA* 469–474.

[Cox et al. 2015] Cox, J.; Oh, E. Y.; Simmons, B. D.; Lintott, C. J.; Masters, K. L.; Greenhill, A.; Graham, G.; and Holmes, K. 2015. Defining and Measuring Success in Online Citizen Science:. *Computing in Science & Engineering* 17(4):28–41.

[Dontcheva et al. 2014] Dontcheva, M.; Morris, R. R.; Brandt, J. R.; and Gerber, E. M. 2014. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems (CHI '14)*, 3379–3388. New York, New York, USA: ACM Press.

[Dow et al. 2012] Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*, 1013. New York, New York, USA: ACM Press.

[Eveleigh et al. 2014] Eveleigh, A.; Jennett, C.; Blandford, A.; Brohan, P.; and Cox, A. L. 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems (CHI '14)*, 2985–2994. New York, New York, USA: ACM Press.

[Ferrari, de Carvalho, and Trevisan 2015] Ferrari, F.; de Carvalho, R. R.; and Trevisan, M. 2015. MorfometrykaâĂŤa New Way of Establishing Morphological Classification of Galaxies. *The Astrophysical Journal* 814(1):55.

[Glassman et al. 2016] Glassman, E. L.; Lin, A.; Cai, C. J.; and Miller, R. C. 2016. Learnersourcing Personalized Hints. *Proceedings of the 2016 conference on Computer supported cooperative work (CSCW '16)*.

[Harris 2011] Harris, C. G. 2011. You're Hired ! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)* 15–18.

[Iacovides et al. 2011] Iacovides, I.; Aczel, J.; Scanlon, E.; Taylor, J.; and Woods, W. 2011. Motivation, Engagement and Learning through Digital Games. *International Journal of Virtual and Personal Learning Environments* 2:1–16.

[Iacovides et al. 2013] Iacovides, I.; Jennett, C.; Cornish-Trestrail, C.; and Cox, A. L. 2013. Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. *Proceedings of the 2013 annual conference on Human Factors in Computing Systems - CHI EA '13* 1101.

[Jennett, Kloetzer, and Schneider 2016] Jennett, C.; Kloetzer, L.; and Schneider, D. 2016. Motivations, Learning and Creativity in Online Citizen Science. *Journal of Science Communication.* 15(03):1–23.

[Kahrimanis, Avouris, and Komis 2011] Kahrimanis, G.; Avouris, N.; and Komis, V. 2011. *Interaction Analysis as a Tool for Supporting Collaboration: An Overview*, volume 350.

[Keel et al. 2012] Keel, W. C.; Chojnowski, S. D.; Bennert, V. N.; Schawinski, K.; Lintott, C. J.; Lynn, S.; Pancoast, A.; Harris, C.; Nierenberg, A. M.; Sonnenfeld, A.; and Proctor, R. 2012. The Galaxy Zoo survey for giant AGN-ionized clouds: Past and present black hole accretion events. *Monthly Notices of the Royal Astronomical Society* 420(1):878–900.

[Kim 2014] Kim, J. 2014. Learnersourcing : Improving video learning with collective learner activity by Juho Kim.

[Kittur, Smus, and Kraut 2011] Kittur, A.; Smus, B.; and Kraut, R. 2011. CrowdForge Crowdsourcing Complex Work. *Proceedings of the 2011 annual conference on Human Factors in Computing Systems - CHI EA '11* 1801.

[Kloetzer 2015] Kloetzer, Laure; Schneider, D. d. C. J. . A.-A. O. J. C. 2015. Technology Enhanced Creative Learning in the field of Citizen Cyberscience.

[Law et al. 2016] Law, E.; Yin, M.; Goh, J.; Chen, K.; Terry, M.; and Gajos, K. Z. 2016. Curiosity Killed the Cat , but Makes Crowdwork Better. *Proceedings of the ACM conference on Human factors in computing systems (CHI '16)*.

[Little et al. 2009] Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2009. TurKit: Tools for iterative tasks on mechanical turk. *2009 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC 2009* (Figure 1):252–253.

[Marcus and Parameswaran 2015] Marcus, A., and Parameswaran, A. 2015. Crowdsourced Data Management Industry and Academic Perspectives. 6(1):1–161.

[Masters et al. 2016] Masters, K.; Oh, E. Y.; Cox, J.; Simmons, B.; Lintott, C.; Graham, G.; Greenhill, A.; and Holmes, K. 2016. Science Learning via Participation in Online Citizen Science. 15(03):32.

[Raddick et al. 2013] Raddick, M. J.; Bracey, G.; Gay, P. L.; Lintott, C. J.; Cardamone, C.; Murray, P.; and Vandenberg, J. 2013. Galaxy Zoo: Motivations of Citizen Scientists M. Jordan Raddick. 1–41.

[Rotman et al. 2012] Rotman, D.; Preece, J.; Hammock, J.; Procita, K.; Hansen, D.; Parr, C.; Lewis, D.; and Jacobs, D. 2012. Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*, 217–226. New York, USA: ACM Press.

[Tinati et al. 2015] Tinati, R.; Luczak-roesch, M.; Kleek, M. V.; Simpson, R.; Simperl, E.; and Shadbolt, N. 2015. Designing for Citizen Data Analysis : A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)* 4069–4078.

[von Ahn and Dabbish 2008] von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):57.

[von Ahn 2013] von Ahn, L. 2013. Augmented intelligence: the Web and human intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1987):20120383.

[Weir et al. 2015] Weir, S.; Kim, J.; Gajos, K. Z.; and Miller, R. C. 2015. Learnersourcing Subgoal Labels for How-to Videos. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* 405–416.

[Willett et al. 2013] Willett, K. W.; Lintott, C. J.; Bamford, S. P.; Masters, K. L.; Simmons, B. D.; Casteels, K. R. V.; Edmondson, E. M.; Fortson, L. F.; Kaviraj, S.; Keel, W. C.; Melvin, T.; Nichol, R. C.; Raddick, M. J.; Schawinski, K.; Simpson, R. J.; Skibba, R. A.; Smith, A. M.; and Thomas, D. 2013. Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey.

[Wood, Bruner, and Ross 1976] Wood, D.; Bruner, J. S.; and Ross, G. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17(2):89–100.