Collaboration versus Competition: Design and Evaluation of Mechanics for Games with a Purpose

Kristin Siu, Alexander Zook, and Mark O. Riedl
School of Interactive Computing
Georgia Institute of Technology
{kasiu; a.zook; riedl}@gatech.edu

ABSTRACT

Games with a purpose (GWAPs) have proven to be effective solutions to solving difficult problems, labeling data, and collecting commonsense knowledge. Unlike traditional games, GWAPs must balance between acquiring accurate solutions or data and maintaining player engagement. However, when it comes to designing GWAPs, the effects of different game mechanics on accuracy and engagement are not well understood. We report on a study designed to investigate how scoring mechanisms based on principles of collaboration and competition impact the accuracy and engagement of players in commonsense knowledge collection tasks. Overall, we found that competition-based scoring mechanics generated data that was as accurate as more conventional collaborative scoring mechanics, but increased player engagement. Furthermore, when players were presented with both collaborative and competitive scoring options they performed worse due to the need to consider strategy. Our results suggest that different mechanics may be used with different players without loss of accuracy.

Categories and Subject Descriptors

Applied Computing [Computers in other domains]: Personal computers and PC applications – Computer games. Information Systems [World Wide Web]: Web Applications – Crowdsourcing.

General Terms

Design, Human Factors, Experimentation

Keywords

Games with a purpose, Game design, Collaboration, Competition

1. INTRODUCTION

Human Computation is a paradigm for leveraging human processing power to solve problems that computers cannot [16]. Human computation can take many forms, but one of the most intriguing forms is that of the computer game. *Games with a Purpose* (GWAPs) [15, 16] are games in which players generate useful data or solve problems as a by-product of play. For example, players may label images [16], discover the shapes of proteins [4], or categorize concepts to develop an ontology [13]. One important challenge in designing these systems is structuring the game to incentivize players to produce useful results. Another key challenge, especially for voluntary tasks, is giving players an engaging and enjoyable experience. Even for paid tasks, an engaging experience may make players more productive.

Unfortunately, making computer games is incredibly hard. Game design is still a black art: even with years of experience, designers have a difficult time predicting how their design decisions will impact player behavior. For GWAPs, understanding the impact of these design decisions on the player is an especially important goal. Particular game design decisions (e.g., how a game score is calculated) can have a profound effect on the accuracy of solutions and data produced by players, efficiency of workers toward the computational system's goals, and motivation of players to provide work.

As a starting point, von Ahn and Dabbish [15] enumerate a number of game design patterns that are believed to be especially effective. For example, the most commonly used pattern—called an *output-agreement game*—involves two players that collaborate to generate labels for things (e.g., images) and are rewarded when labels match. The other common designs rely on the principle of collaboration as well. But we don't yet understand *why* these design patterns work, nor whether alternative design patterns will be just as—or more—effective for certain types of tasks or players. In contrast to most GWAPS, for example, most entertainment-oriented games are designed around the principle of competition, wherein players compete to outperform each other.

The long-term goal of our work is to develop an understanding of how different *game mechanics*—rules that dictate how the game system behaves—impact player accuracy, efficiency, and motivation for a range of types of human computation tasks. In this paper, we investigate and compare how scoring mechanics based on principles of collaboration and competition impact the accuracy and engagement of players in commonsense knowledge collection tasks. Whereas other researchers have designed and implemented GWAPs that utilize co-operation for data verification purposes (e.g., [2, 4, 9, 11, 13, 16]), and some have introduced competition to encourage diversity of solutions (e.g., [7, 14] to date there has not been a principled study of the impacts of collaborative and competitive scoring mechanics on player behavior.

We report on a study in which we compare alternative versions of a game containing collaborative and/or competitive scoring mechanics. We measure player accuracy and engagement. We found that players of the collaborative and competitive versions of the game provided results of similar accuracy but found the competitive version more engaging. From this, we conclude that some GWAPs may benefit from competitive mechanics without loss of accuracy. Additionally, we examined the effects of giving players a choice to play either collaboratively or competitively, in a version of the game that utilized both mechanics. We found that due to the complexity of the game, players performed less accurately when confronted by a choice in strategy. These conclusions help us better understand how to choose game mechanics in which to build future GWAPs. This is a step toward increasing player engagement and, consequently, making players more productive towards a GWAP's goals.

In this paper, we first overview related work on GWAPS, with an emphasis on player collaboration and competition. We next describe our study methodology, including the game, *Cabbage Quest*, which we developed specifically for this study. Then, we present the results of our study, followed by a discussion of their implications. We finish with limitations, future work, and conclusions about where this work can lead with regard to the scientific study of human computation games.

2. RELATED WORK

GWAPs have been used to solve a variety of computationally intractable problems, collect labeled data, or aggregate commonsense knowledge. Many games follow the model proposed by the *ESP Game* [16]. In the *ESP Game*, two anonymous players are presented with an image and are asked to independently describe (label) its contents. Here, consensus becomes verification for the solution. A number of GWAPs have been developed using analogous mechanics, including *Verbosity* [17], *Peekaboom* [18], and *Tag a Tune* [9]. These games highlight a class of tasks common to human computation, in which users are asked to classify or annotate data with additional information. Often, these tasks rely on commonsense knowledge that is difficult to acquire or represent using current algorithms. From their experiences designing GWAPs, von Ahn and Dabbish [15] derive three design patterns: *output-agreement* (e.g., the *ESP Game*), *inversion-problem*, and *input-agreement*. Under these design patterns, the mechanics of games are often solely structured around the human computation task and utilize collaboration-based scoring mechanics. This is in contrast to contemporary work in game design, which prioritizes and optimizes game mechanics for player engagement. However, due to the success of the *ESP Game* and other similar games, there has been little exploration of alternatives to collaboration-based scoring mechanics with respect to creating player engagement.

There are a number of notable exceptions to this trend. *OnToGalaxy* [3] is a GWAP designed to solve ontology-related problems, which is presented as a top-down 2D spaceship shooter as opposed to the more traditional GWAP interfaces for solving similar problems [13]. *HerdIt* [2], a music-tagging game targeted specifically at casual audiences, emphasizes an iterative development process and a user-centered design that prioritized user concerns ahead of the task. Our long-term goal builds off these efforts by attempting to formalize the relationship between seemingly arbitrary design decisions and the impact it has on players of GWAPs.

Most popular mainstream entertainment games often incorporate both collaborative and competitive mechanics into multiplayer modes. However, GWAPs typically focus solely on collaboration, since verification of solutions relies on a consensus. Even in single-player games such as *FoldIt* [4], good solutions are the result of multiple players coming together to solve the problem. One notable exception is *KissKissBan* [7], a variation of the *ESP Game* that introduces competitive elements on top of the existing collaborative scoring mechanics. In *KissKissBan* a third player attempts to impede the progress of the two other collaborators by banning obvious labels, resulting in a more diverse set of results. Another exception is *PhotoCity* [14], a game in which players take photographs of buildings in order to reconstruct 3D models of the environment. Their study involved a competition between two universities, in which players competed against the opposing school while collaborating with their peers at the same institution. While these projects showed that competition could improve or assist collaborative results, both game designs were not compared directly to a purely competitive design.

Studies of collaboration (co-operation) and competition have been conducted in a number of domains for a number of purposes. In studies involving single-player educational math games, both cooperative and competitive gameplay have been shown to have positive psychological effects on learning and motivation [8, 12], but students often preferred cooperative play. Other studies in multiplayer games have investigated the effects of co-operation [5] and challenge [10] on player engagement for the purpose assessing and improving game design. In a recent example, Emmerich and Masuch [6] examine the differences in the player experience between collaborative and competitive versions of their game *Loadstone*. However, their study differs from ours in that they do not focus on player performance. Additionally, *Loadstone* required players to be co-located, which is not the case for most GWAPs. Our study of competitive scoring mechanics is conducted in the context of more traditional, online GWAPS centered around acquiring commonsense knowledge. We further look at accuracy (performance) as well as engagement.

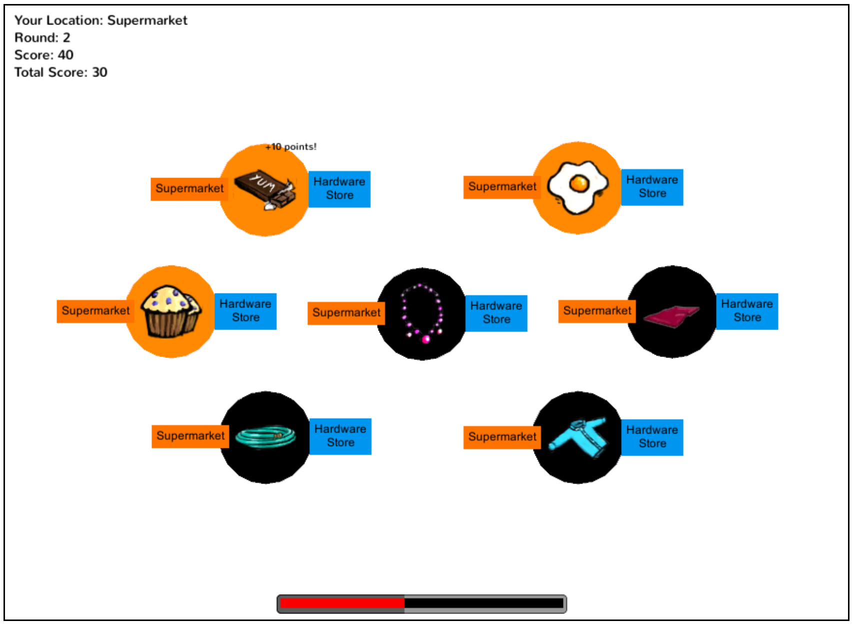
3. GAME DESIGN STUDY

In this paper, we report on a study to analyze the effects of collaborative and competitive mechanics on human player behavior in GWAPs. To achieve this, we designed *Cabbage Quest*, a simple GWAP for classifying everyday items with places that they can be purchased from. We first introduce the game, its design, and the mechanical variations we utilized in our study to compare collaborative and competitive mechanics. We then describe our study and discuss the results.

3.1 The Game

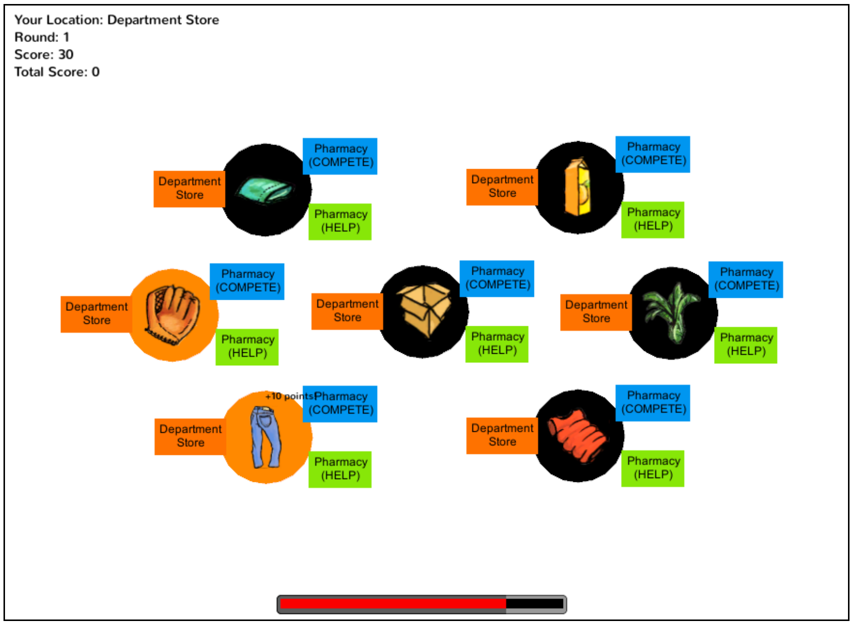
Cabbage Quest is a GWAP for classifying everyday items with potential purchasing locations. As with many GWAPs, *Cabbage Quest* is a two-player game, seeking player consensus of labels. The main interface for *Cabbage Quest* presents each player with a set of everyday objects, such as food or household items, and gives players two possible purchasing locations (see Figures 1 and 2). For example, players may be asked whether a cabbage can be purchased in a grocery store or a pharmacy. Players assign a purchasing location by clicking on corresponding location labels, or they may choose not to assign a location at all if the object cannot be bought at either location. Seven items are classified during a single round of play, and the duration of the round is enforced by a time limit. A score is computed based on the combined behavior of both players, as described below. We created three versions of the game, varying the scoring mechanics and observing how each version affected player behavior and subjective attitudes.

We selected the task of classifying objects with their purchasing locations because of its similarity to other classification and



Your Location: Supermarket
Round: 2
Score: 40
Total Score: 30

Figure 1: A screenshot of *Cabbage Quest*. Players are presented with a set of items that they must assign locations to by clicking on associated labels. The collaborative and the competitive versions of the game share this same interface.



Your Location: Department Store
Round: 1
Score: 30
Total Score: 0

Figure 2: A screenshot of the interface of the collaborative-competitive version of *Cabbage Quest*. This version introduces an additional label, allowing the player to choose to play competitively or collaboratively.

annotation-related problems in human computation. We note that the mechanics of our game are not explicitly tied to the context of purchasing items and believe these can generalize to similar problems. Additionally, we have also selected a task with a known solution. This enables us to objectively measure the accuracy of data generated by players and reflects the methodology used in other game studies (e.g., [11]) that use artificial problems with known solutions in order to study aspects of collaborative work.

Cabbage Quest was designed to evaluate the effects of both collaborative and competitive mechanics in GWAPs. We chose to vary the scoring mechanics because they provide objective feedback to the player about the consequences of his or her actions. Each time the player assigns a location to an item, he or she receives a change in score. This score is based on two criteria: correctness and agreement. Correctness is reflected in a base score, which remains consistent across all variations of the game. Agreement is based on the scoring paradigm for that particular variation of the game. In all three versions of the game, players are awarded a base score for assigning an object to a location correctly. Correctness was based on a gold standard we developed *a priori* for the study, but this could easily be replaced with the aggregate consensus for an object in the event that no such standard exists. The base score was used to encourage players to assign all objects correctly, regardless of the additional scoring mechanics. On top of this base score, each version of the game features a different scoring function, which reflects the presence of collaborative or competitive mechanics. The three game variations are *Collaborative*, *Competitive*, and *Collaborative-Competitive*, and are described in detail below.

3.1.1 Collaborative Version

Each player is awarded 10 additional points if they both agree on the same location for an object. This scoring function mimics the standard scoring mechanics of the *ESP Game* and other GWAPs.

3.1.2 Competitive Version

Each player is assigned a primary location: one of the labels belongs to him or her, while the other belongs to their opponent. As before, the player may choose to label an item with their primary location, their opponent's location, or neither. If a player assigns his or her opponent's location to an object and is the first to do so, then he or she gains 10 points while their opponent loses 10 points. Similarly, if the player assigns their primary location to an object faster than their opponent, they gain 10 points (but also prevent the opponent from causing them to lose points). Thus players must make decisions about *when* to assign their location and their opponent's to objects. A defensive strategy is to assign the player's location to corresponding objects first in an attempt to maintain their score. An offensive strategy is to attempt to assign the opponent's location to corresponding objects first to decrease the opponent's score. Hybrid strategies exist as well.

3.1.3 Collaborative-Competitive Version

The final variation of the scoring mechanics combines both the collaborative and competitive elements described above. As with the competitive version, players are assigned a primary location. However, players now have *two* possible ways to assign the other player's location to an object: collaboratively—indicated by the word "HELP" next to one location label—or competitively—indicated by the word "COMPETE" next to one location label as shown in Figure 2.

Assigning collaboratively follows the collaborative scoring mechanic: both players will receive 10 points if they agree, regardless of timing. Assigning competitively invokes the competitive mechanic. If a player assigns the other player's location first, he or she receives 10 points while deducting 10 points from the other player's score. Being the second to select this label yield no points to the player. Likewise, when assigning his or her primary location to an object, the player will gain 10 points if the other player uses the collaborative assignment, while potentially losing points if the other player uses the competitive assignment.

3.2 Methodology

For the study, we selected a set of common household items and a set of four possible purchasing locations. Before beginning the study, we built a gold standard mapping between items and locations by asking a panel of experts (volunteers unaffiliated with the project, but familiar with the items) to label items with possible purchasing locations. The gold standard contained fifty-three items, each of which was assigned one or more purchasing locations. We use the gold standard to award points for the base score as described above. We reference the standard later as a ground truth to verify the accuracy of our results.

Cabbage Quest was made available as a browser-based game using the Unity framework. Participants were recruited via email and social networks, and were directed to the website where they were assigned to play one of the three variations of the game. Upon launching the game, each participant was then given a short tutorial describing the scoring mechanics, followed by at least five rounds of the game. Each round lasted for fifteen seconds and contained a set of seven items. Following five rounds, participants were given the option to play additional rounds as they pleased. Once participants indicated they were finished playing, they were asked to complete a short survey. The survey included demographics questions about age, gender, and familiarity with games, plus three questions answered on a scale of 1 to 5: how challenging was the game, how engaging they found the game, and how likely they were to play the game again.

To simplify the online study, participants played the game asynchronously; each player was paired with a virtual player using a prior play trace. Since there was no guarantee that the current set of items and locations had been seen before, the virtual player chose object-location pairs randomly, but used the same timing as a prior trace. This is similar to the methods used by other GWAPs to compensate when there are not enough players online at any given time to match together [13, 15].

4. RESULTS

The study was conducted over the course of several weeks, during which the game was made available online to participants. 118 participants took part in the study. Of these, 44 played the collaborative version, 36 played the competitive version, and 38 played the collaborative-competitive version. Altogether, participants played a total of 796 rounds, resulting in 211 total unique object-location pairs. We discarded data from participants who played the game but did not complete the survey questions. Of the participants, 97 were male and 21 were female. All but 5 participants had prior gaming experience (although they were not necessarily familiar with GWAPs).

4.1 Accuracy

To assess how accurate a version of a game was, we first determined the accepted answers to the question of where each object could be purchased based on the input of all players. An accepted answer is the location selected the most times by all players. To compensate for the fact that an object could be purchased in multiple locations, we employed an additional rule: if another location had at least 80% of the total pairs given to majority location, it would also count as a valid assignment. In this way, an object with pairs split between two (or more) locations could be assigned to both. Each assignment was then compared to the gold standard for correctness. Thus for each version of the game, we measure its accuracy as the percentage of correct assignments over the total number of assignments.

As shown in Table 1, the collaborative version of the game yielded accuracy of 91.2%. The competitive version yielded an accuracy of 85.5%. The collaborative-competitive version yielded an accuracy of 77.4%. Both the collaborative and competitive versions resulted in more accurate labels than the collaborative-competitive version (shown as "both" in Table 1). However, only the pairwise difference between the collaborative-competitive and collaborative versions was significant ($p < 0.05$; Wilcoxon signed-rank test, a non-parametric version of the *t*-test that does not assume Gaussian distributions). Given the magnitude of the

Table 1. Per-version breakdown of the number of object assignments based correctness relative to the gold standard.

Assignment	Collaborative	Competitive	Both
Incorrect	5	9	14
Correct	52	53	48
Total	57	62	62

Table 2. Per-version breakdown of mean delta times (in seconds) based on answer correctness.

Answer	Collaborative	Competitive	Both
Any answer	2.10	1.96	1.92
Correct answer	2.16	2.08	2.13
Incorrect answer	2.07	1.81	1.70

differences, we believe a larger sample size would find these differences to be significant.

We also looked at scores across the different versions of the game to see if score was an indicator of per-game accuracy (i.e. how accurate were labels in a game overall, as opposed to how accurate the labels were over all games). A well-designed scoring mechanism should theoretically incentivize players to perform accurately and also give feedback as to the accuracy of play. Scores were found to have a medium correlation with accuracy (r = 0.275, p < 0.001). This correlation was the largest for the competitive version of the game (r = 0.442, p < 0.001) and smallest for the collaborative-competitive version of the game (r = 0.138, p < 0.001). Figure 3 shows the correlations between score and accuracy.

We note that a potential design consideration for GWAPs might be to tailor the mechanics to optimize results given a particular demographic. Although not part of our original hypothesis, we found that gender also appeared to have a significant effect on accuracy. If we compare the mean percentage of correct assignments to total provided assignments (while accounting for score and accuracy using MANOVA), female participants' assignments were correct around 69% of the time, compared to 57% provided by male participants (p < 0.001). However, no significant differences in accuracy were detected within the three versions of the game.

Finally, we looked at the timing information across all three versions of the game as another potential design consideration that might be to select mechanics that increase the rate at which players provide answers. The distributions of the player timings of object assignments were similar across all versions. To get a better idea of the rate at which players were making decisions, we examined the mean time between player answer selections. Table 2 summarizes these results, controlled for correctness, game version, and gender using ANOVA. We found that participants in the collaborative version were around 7% slower compared to the competitive (p = 0.056) version and around 9% slower compared to the collaborative-competitive (p < 0.01) version. This suggests that the incorporation of competitive mechanics into the game induced people to work more quickly, likely due to the fact that competitions are won by answering faster than the other player.

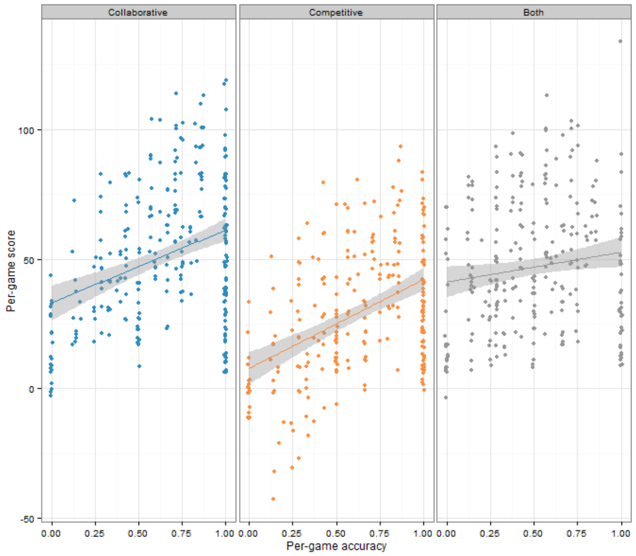


Figure 3: Linear correlation of accuracy to score across all three versions of the game.

4.2 Player Engagement

To measure player engagement, we looked at the number of rounds that participants played during the study, as well as their responses in the post-game survey. There was no significant difference in the number of rounds participants played across versions. The median number of rounds played in all three versions was 5; participants were asked to complete at least five rounds before being given the option to quit and fill out the survey.

In the post-game survey, participants rated the game on a 1-5 scale in three categories: the game's level of challenge, their enjoyment, and their likelihood of playing again (Figure 4). We equated higher values in all three categories to higher levels of player engagement.

Participants rated the games as similarly challenging (with medians of 2, 2.5, and 3 respectively, but with no significant differences). The collaborative-competitive version stands out as potentially the most challenging, as it was the only version to receive maximum challenge ratings. This is understandable considering that players must complete the task while also considering strategic implications of their actions.

Participants reported greater engagement in the competitive version; it had a smaller percentage of low ratings when compared to the other versions and a larger percentage of high ratings. Enjoyment ratings for the collaborative-competitive version were primarily in the low (1-3) range, indicating that players did not enjoy this version (perhaps due to its complexity). Conversely, the competitive version received higher ratings on the question of whether participants would play the game again.

4.3 Multiple Simultaneous Mechanics

The collaborative-competitive version yields additional data not present in the other versions. In particular, players must make a choice of which scoring mechanic they will use. In this version of the game, players have four assignment options for each item: they can use their location ("self"), they can use the other player's location either collaboratively ("help") or competitively ("compete"), or they can decide not to assign the object at all. Notably, the distribution of use of the first three options across all

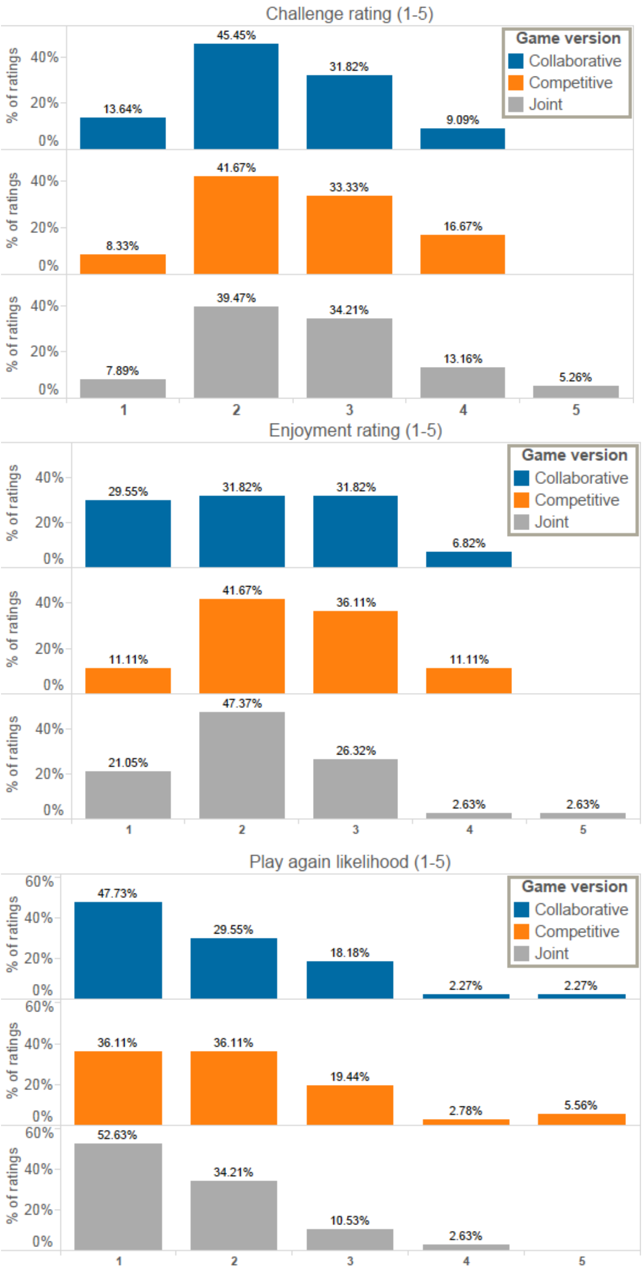


Figure 4: Results of the post-game survey for each version of the game, broken down by the percentage of answers from players of a given game version.

of the participants was not uniform. That is, participants did not have a strategy of using these labels evenly.

Participants demonstrated a preference towards using certain options as opposed to others. Roughly 34% of participants chose to use the other player’s location options (either "help" or “compete) near-exclusively. We consider an option to be used “near exclusively” if the participant used it more than 60% of the time. That is, they ignored their own location in favor of the other player’s. More strikingly, when participants used the other player’s location options, they tended to chose one option and use it consistently. Roughly 32% of participants used the "help" option near-exclusively while 42% of participants used the "compete" option. In short, nearly 3 out of 4 players chose to either exclusively compete or exclusively collaborate.

Table 3. Per-option breakdown of the number of object assignments in the collaborative-competitive version based on correctness relative to the gold standard.

Assignments	**"help"**	**"compete"**	**"self"**
Incorrect	33	31	19
Correct	33	44	48
Total	66	75	67

Table 4. Per-label breakdown of mean decision times (in seconds) in the collaborative-competitive version based on answer correctness.

Answer	**“help”**	**“compete”**	**“self”**
Any answer	1.80	1.75	2.10
Correct answer	2.05	2.02	2.26
Incorrect answer	1.65	1.50	1.93

The choice of strategy appears to have some influence on player accuracy. When using the "self" option (the player's primary location), accuracy was highest at 71.6% when compared with either the “help” option or “compete” option, which had lower accuracies of 50% (p < 0.05) and 58.7% (p = 0.106), respectively (Table 3).

Likewise, we also investigated the timing associated with using each of the options. Table 4 details the mean decision times across the different options. These times represent the average number of seconds players spent between using the different options. When using the "self" option, participants took significantly longer than when using either the "help" or "compete" option (both $p < 0.001$). Additionally, participants took significantly more time to provide correct answers than incorrect answers (all $p < 0.05$).

5. DISCUSSION

The goal of the study was to investigate the effects of different mechanics on GWAPs. Analyzing our results on accuracy and engagement led to some interesting observations about the collaborative version and the competitive version of the game. Additionally, our results from the joint collaborative-competitive version of the game provide information about how layering multiple mechanics might affect player strategy and accuracy.

5.1 Collaboration versus Competition

Previous GWAPs have emphasized collaborative (agreement) mechanics; these appear to compliment problems in human computation that require consensus as a means of verification. What effect do competitive mechanics have on both the accuracy of the solution and player engagement? Here, we discuss similarities and differences between the collaborative and the competitive versions of the game.

Based on our data, we did not find a significant difference in accuracy between the collaborative version and the competitive version. Therefore, we can plausibly conclude that a GWAP of the type studied in this paper could be designed to use either collaborative or competitive mechanics and expect similarly accurate results. GWAPs such *KissKissBan* have already demonstrated the potential for competitive elements to be added to traditionally collaborative games, but it has been unclear just what the effects of these elements have been on the accuracy of information gathered. Our results show that these elements do not necessarily compromise accuracy or quality. This opens up new ways to potentially design GWAPS. For example, if the problem domain permitted it, players who prefer to play exclusively collaboratively or exclusively competitively can be presented with a personalized or preferred option.

Likewise, timing information brings up other design considerations. If we care about the rate at which we acquire results, then competitive scoring mechanics seem to encourage players to provide information faster in general. This is likely affected by the fact that our competitive scoring mechanics are based on time, as players who perform faster than their opponents receive a higher score (while penalizing their opponent). Since per-round time constraints are a common mechanic in other GWAPs, we believe that our approach to competitive scoring could be applied to similar games.

When it comes to player engagement, the competitive version of the game stands out as more engaging compared to the collaborative version of the game. Participants tended to rate the collaborative version lower with respect to enjoyment and likelihood of playing again. We found players are more engaged with competitive game mechanics. This conclusion is reinforced by data from the collaborate-competitive version, where more people exclusively used a competitive strategy. A possible concern when designing GWAPs might be that players would be engaged and distracted by competitive elements, therefore providing less accurate information. Our results demonstrate otherwise. There seems to be a potential benefit to considering competitive mechanics for GWAPs.

5.2 Multiple Mechanics

Our initial results indicate that competitive mechanics might be an equivalent, if not more engaging alternative to the traditional collaborative design of GWAPs. However, many games consist of a number of complex, interacting systems requiring strategic decision-making. Not to mention, multiplayer games often incorporate both collaborative and competitive elements side-by-side, offering players choices about how to interact with others in the game. Our existing mechanics provided the opportunity to test collaborative and competitive mechanics side-by-side. What would be the result of introducing this choice in a GWAP? The collaborative-competitive version of the game had the lowest accuracy and, while it appeared to be the potentially the most challenging, participants enjoyed it the least. It is likely that incorporating multiple scoring mechanics ultimately made the game too complex (cognitive load-inducing) for players expecting short play experiences.

Our results show that players had a preference for particular options, which resolve to distinctly collaborative or distinctly competitive play styles. This observation was echoed in additional feedback from participants. Informal conversations with some participants reiterated that they tended to "stick to one option" over the other. In short: when given the choice, players tended to play either collaboratively or competitively, but rarely use both strategies.

Given that players choose one particular strategy over another, the next question is whether choosing particular options has an effect on the accuracy of the results. When assigning their own location ("self"), players had a much higher accuracy rate. We hypothesize this is due in part to the complexity provided by the choice of having two options "help" and "compete" for the other player's location. Players performed consistently better at assigning a location (their own) when only one option was present. Thus, in exchange for giving players the freedom to pick their style of play, the accuracy of the results was clearly compromised and players ultimately found the game more frustrating.

We do not conclude that adding multiple mechanics to a GWAP is necessarily detrimental, however our results illustrate some of the complications and complexity involved in doing so. It is possible that complex scoring mechanics are not a good fit for short round categorization games where there are no long-term incentives to learn the strategy that will work most effectively with the other player. Giving players the choice to strategize during the game appears not only to be a distraction, but in the case of *Cabbage Quest*, ultimately did not matter because players generally picked a consistent strategy regardless of any other incentives (in this case, score).

5.3 Limitations and Future Work

Our study has several limitations. First, it is unclear how our results would be affected by a problem with a larger number of objects to classify or a larger audience of players. Our game closely resembles other output-agreement style games, but our results may not generalize depending on the specific nature of the problem. We also demonstrate only one way to make scoring mechanics competitive: through use of time. There may be other ways to implement competitive mechanics that also preserve verification via player agreement. Additionally, although we took steps to disguise the asynchronous play, many players may have realized they were not playing against real, synchronous humans. This may have led players to prefer competition. However, we note that the study by Emmerich and Masuch [6] with co-located subjects also found a preference for competition. Finally, the demographics of our study pool may not be representative of audiences likely to play GWAPs.

In *Cabbage Quest*, we manipulated and investigated the effects of collaborative and competitive scoring mechanics. A possible next step would be to investigate other types of mechanics at this scale in order to determine their effects. We believe our methodology provides a useful framework for testing a wide variety of GWAP mechanics, such as the effects of timing or aesthetical elements.

More broadly, we wish to understand how different mechanics affect different kinds of tasks. Do some classes of human computation problems map better to certain mechanics than others? We would like to see how our results apply to other problems in this space. To do so will help us better predict the effects of game mechanics on human workers in GWAPS.

6. CONCLUSIONS

In our study, we have taken a first step toward more completely understanding the effects of collaborative and competitive scoring mechanics on the accuracy and engagement of players of GWAPs. While most GWAPs are collaborative in nature, our results suggest that (a) a purely competitive version of output-agreement can be implemented that still uses player agreement for label verification, and (b) competition can lead to results that are just as accurate. Our data further suggest that there is a general preference for competitive games. This makes sense given that many entertainment-based games are competitive. Additionally, a side observation is that competitive and collaborative versions of the same game may be deployed to different audiences based on their personal preferences. Complicated scoring mechanics that require players to think strategically, however, appear to be ill-suited for the type of simple GWAP used in the study.

The methodology we used to compare game mechanics in a GWAP is similar in nature to the A/B testing conducted on online educational games (cf., [1]). We used it for the first side-by-side study of collaboration and competition in GWAPs. The methodology is also general in that it can be applied to many other mechanics besides scoring.

Though the study reported is limited to just collaborative and competitive scoring mechanics, it offers insights into how the game mechanics affect the behavior and enjoyment of players. By looking at other, alternative mechanic pairs, we aim to develop a better understanding of how game design decisions will impact the motivation of players to provide work, efficiency of workers toward the computational system's goals, and accuracy of solutions and data produced by players. To that end, we hope to facilitate the development of GWAPs that more closely resemble commercial games of the sort found on social media sites, phones, and tablets. By making GWAPs more engaging, we hope to provide players more compelling experiences that can produce more useful data for solving difficult problems.

7. REFERENCES

[1] Andersen, E., Liu, Y.-E., Snider, R., Szeto, R., Cooper, S., and Popović, Z. 2011. On the harmfulness of secondary game objectives. In *Proceedings of the 6th International Conference on Foundations of Digital Games*. ACM, New York, NY, USA, 30-37.

[2] Barrington, L., O'Malley, D., Turnbull, D., and Lanckriet, G. 2009. User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA, 7-10.

[3] Carranza, J. and Krause, M. 2012. Evaluation of Game Designs for Human Computation. In *AAAI Workshop on Human Computation in Digital Entertainment and Artificial Intelligence for Serious Games*, 9-15.

[4] Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. In *Nature 466*, 756-760.

[5] El-Nasr, M. S., Aghabeigi, B., Milam, D., Erfani, M., Lameman, B., Maygoli, H., and Mah, S. 2010. Understanding and evaluating cooperative games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 253-262.

[6] Emmerich, K. and Masuch, M. 2013. Helping Friends or Fighting Foes: The Influence of Collaboration and Competition on Player Experience. In *Proceedings of the Eighth International Conference on the Foundations of Digital Games*. ACM, New York, NY, USA, 150-157.

[7] Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J.Y., and Chen, K.-T. 2009. KissKissBan: a competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA, 11-14.

[8] Ke, F. and Grabowski, B. (2007), Gameplaying for maths learning: cooperative or not?. British Journal of Educational Technology, 38: 249–259.

[9] Law, E., von Ahn, L., Dannenberg, R. B., and Crawford, M. 2003. Tagatune: A game for music and sound annotation. In *International Conference on Music Information Retrieval*, 361-364.

[10] Lomas, D., Patel, K., Forlizzi, J. L., and Koedinger, K. R. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 89-98.

[11] Mason, W. and Watts, D. J. 2011. Collective Problem Solving in Networks. Retrieved November 28, 2013 from: http://dx.doi.org/10.2139/ssrn.1795224

[12] Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., and Perlin, K. 2013. The Impact of Individual, Competitive, and Collaborative Mathematics Game Play on Learning, Performance, and Motivation. *Journal of Educational Psychology*, 105, 1050-1066.

[13] Siorpaes, K. and Hepp. M. 2008. Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems*, vol. 23, no. 3, 50-60.

[14] Tuite, K., Snavely, N., Hsiao, D., Tabing, N., and Popović, Z. 2011. PhotoCity: training experts at large-scale image acquisition through a competitive game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1382-1392.

[15] von Ahn, L. and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51, 8 (August 2008), 58-67.

[16] von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 319-326.

[17] von Ahn, L., Kedia M., and Blum, M. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 75-78.

[18] von Ahn, L., Liu, R., and Blum, M. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 55-64.