Reward Systems in Human Computation Games

Kristin Siu and Mark O. Riedl
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
{kasiu, riedl}@gatech.edu

ABSTRACT

Human computation games (HCGs) are games in which player interaction is used to solve problems intractable for computers. Most HCGs use simple reward mechanisms such as points or leaderboards, but in contrast, many mainstream games use more complex, and often multiple, reward mechanisms. In this paper, we investigate whether multiple reward systems and ability to choose the type of reward affects human task performance and player experience in HCGs. We conducted a study using a cooking-themed HCG, Cafe Flour Sack, which implements four reward systems, and had two experimental versions: one which randomly assigns rewards and the other which offers players the choice of reward. Players were recruited from both Amazon Mechanical Turk and university students. We report the results across these different game versions and player audiences. Our results suggest that offering players a choice of reward can yield better *task completion* metrics and similarly-engaged *player experiences*, and may improve these metrics and experiences for audiences that are not experts in crowdsourcing. We discuss these and other results in the broader context of exploring different rewards systems and other aspects of reward mechanics in HCGs.

ACM Classification Keywords

H.5.3. Information Interfaces and Presentation (e.g. HCI): Computer-supported cooperative work; K.8.0 Personal Computing: Games

Author Keywords

human computation games; games with a purpose; rewards; game design

INTRODUCTION

Human computation games (HCGs) are games in which player interaction is used to generate data or solve problems traditionally too difficult or intractable for computers to model. These games, also called *Games with a Purpose* (GWAPs), have been effectively deployed in domains such as data classification (e.g, image labeling [24]), scientific discovery (e.g., protein folding [3]), and data collection (e.g., photo acquisition [23]). However, despite an increased public appetite and a growing societal benefit for games, HCGs have not seen widespread adoption. Some of this can be attributed to the fact that game design and development is still a difficult and time-consuming process.

Developing human computation games still remains challenging because like other serious games, HCGs have two often-orthogonal design goals. On the one hand, the human computation task must be solved effectively and on the other, the game should provide an entertaining player experience. Balancing the two is still a formidable task, even for experienced game designers. To complicate this, we know very little about how to design these games. Conventional game design theories often do not accommodate the additional requirements imposed by solving the task. Existing design knowledge in HCGs is limited to templates and anecdotal examples that do not easily generalize to new tasks and changing audiences. Growing this design space would enable scientists, researchers, and amateur developers to create HCGs more effectively, allowing for more games to solve many interesting problems that might otherwise be computationally intractable.

In this paper, we focus on the *reward systems* in human computation games. Without players, the underlying human computation tasks in HCGs may never be completed, and reward systems — the sets of gameplay mechanics responsible for providing positive feedback — allow us to compensate players directly for contributing their time and effort to solving these problems. This makes rewards some of the most important gameplay elements to investigate in HCGs because of their role in motivating and engaging players.

Currently, most HCGs tend to adopt simple reward systems such as point systems and leaderboards, mirroring collaborative elements of puzzle games combined with social (and sometimes competitive) mechanics. However rewards in mainstream digital games are often far more complex, and take on a wide variety of forms not seen in current HCGs. One longstanding question in HCG design is how to adopt the mechanics of modern digital games in a way that respects both the *task completion* —player performance at the task — and the *player experience* —player interaction and engagement with the game. Rewards are no exception to this, but unfortunately, we know very little about how different rewards systems behave in human computation games, let alone which ones are the most effective. Mainstream digital games often incorporate multiple, different reward systems in order to appeal

CHI PLAY 2016, October 16–19, 2016, Austin, Texas, USA.

ACM ISBN 978-1-4503-4456-2/16/10 ...$15.00.
http://dx.doi.org/10.1145/2967934.2968083

to a wide variety of player motivations and allow for diverse player experiences. However, we know little about how these more-complicated systems might behave in HCGs.

This paper is a first step towards untangling the effects of using different rewards systems in human computation games. We wish to investigate the question of whether or not randomly distributing rewards to players as opposed to offering players a choice between different reward systems has any effect on the task completion and the player experience. Beyond looking at reward systems, we are interested in understanding how different reward systems affect different audiences of players.

To facilitate our investigation, we instrumented a human computation game with four kinds of reward systems. We then ran a study comparing two versions of the game: one which randomly assigns rewards to players and one which offers players the choice of different rewards. We conducted the study using two different player audiences: workers from an online crowdsourcing platform and university students. We evaluated the results as they relate to both the *task completion* and the *player experience*. Our results show significant differences between the *random* and *choice* conditions of the game, as well as differences between player audiences and interactions between these variables. For example, players in the *choice* condition completed tasks more accurately and more quickly than players in the *random* condition. We discuss these results, highlighting design recommendations around reward systems in HCGs, along with future directions for studies in this area. Ultimately, we believe this is a step to better understanding how reward systems work in HCGs, which would open possibilities for new, effective, and entertaining games.

BACKGROUND AND RELATED WORK

Rewards in HCGs

Traditionally, most human computation games have adopted simple reward systems with mechanics that focus on the collaborative nature of human computation tasks and the social features of crowdsourcing. Point and scoring systems are generally the most common form of feedback to the players. In addition to being easy to implement, they provide both a form of direct feedback to players and a way for task providers to monitor and evaluate performance at the task. Most recent survey work in HCGs [7, 16] explores rewards as different forms of incentives available to the player, although these are again in the context of scoring systems.

Design knowledge in HCGs has focused primarily on what kind of behavior players should be rewarded for, specifically as it pertains to collaboration and competition. Early design work in von Ahn and Dabbish's game templates [25] for classification tasks outlines that players should be rewarded for collaborative agreement — which maps to the divide-and-aggregate approach to solving the underlying human computation task. The respective games described [9, 24, 26] all implement collaborative scoring systems, which reward players for agreeing on task results, while leaderboards provide a social interface for players to interact, share scores, and compete. Design of scoring systems and leaderboards is further explored in *Foldit* [4], where the authors describe the design of their scoring function and the evolution of their leaderboards to better enable collaborative sharing (of protein solutions) while still providing an interface for players to compete. Competitive play in HCGs has been explored in games such as *PhotoCity* [23], which utilized an explicit competition between students at two universities and *KissKissBan* [6], which implemented a three-person competitive variant of the original *ESP Game* [24]. Finally, a study comparing collaborative and competitive scoring systems [20] suggests that collaborative scoring systems may yield better task completion results while competitive scoring systems may provide a more engaging player experience. However, none of these design investigations and games explore different or alternative kinds of reward systems beyond point systems and leaderboards; we investigate alternative systems in this paper. This relates to a longstanding question [8, 7, 22] of how to incorporate gameplay elements of modern, commercial games into HCGs in ways that do not compromise the quality of either the *task completion* or the *player experience*.

Rewards and Motivation in Games

Outside the domain of human computation games, reward systems have been widely explored. In game design for digital games, common approaches towards understanding and designing effective rewards in games are driven by theories on motivations and incentives. Early approaches in game design sought to understand how player motivations mapped to mechanics and reward in games, often for game genres with diverse player bases such as multi-user dungeon games (MUDs) [1], tabletop roleplaying games [10], MMOs [28], and online games [2]. Models for player motivation and engagement incorporate psychological theories, such as self-determination theory [17]. A comprehensive overview of motivational theory as it applies to gamification and serious games can be found in the work of Richter et al. [18]. The authors note that point systems are the most-commonly utilized form of reward feedback, and while their discussion focuses primarily on extrinsically-motivated rewards, they note that the effect of extrinsic rewards on intrinsic motivation still remains unknown. How these existing theories might need to be modified in order to accommodate motivations unique to human computation is an open question. Unfortunately, only few attempts have been made to understand motivations in the context of HCGs. Using their game *Indagator*, Lee et al. [11] explore motivations for participating in mobile content-sharing using a model of player gratification. Similarly, in their analysis of *Foldit* [3], Cooper et al. report the results of a suvey asking a subset of users about motivations for playing the game. Their responses were categorized based on Yee's motivational components [28], amended with an additional "purpose" category to capture intrinsic motivations for participation (i.e., assisting with scientific discovery). Similar explorations appear in other serious game domains, such as educational games [12], which make adjustments to existing theories to accommodate for intrinsic motivations beyond those driven by gameplay.

Motivation in Crowdsourcing

Research in crowdsourcing, specifically in the context of paid crowdsourcing platforms, has also examined the effects of

motivation on worker performance, where extrinsic motivations are captured by financial compensation in addition to any intrinsic motivation workers may have for solving the task. Existing work shows that monetary reward may undermine the effects of intrinsic motivation in crowdsourced workers [15] and that increasing the amount of financial compensation may yield more results, though not necessarily those of higher quality [13]. Additionally, studies have examined the interchangeability between paid crowdsourcing platforms and HCGs [21, 19], suggesting that the quality of the completed work between the two is comparable. However, Sabou et al.[19] remark that maintaining player motivation in HCGs may be more difficult, suggesting that motivational findings in the context of financially-compensated crowdsourcing may not translate directly to HCGs. Thus is unclear whether how and if so, to what extent, rewards in HCGs compare with financial compensation.

EXPANDING ON REWARDS IN HCGS

Beyond point systems and leaderboards, we know very little about how other kinds of reward systems behave in human computation games. However, we know that all players are not necessarily motivated by point systems and leaderboards, but also for more immersive reasons which are not always encapsulated in the most-commonly used reward systems in HCGs. The diversity of reward and feedback systems in modern commercial games provides attractive alternatives, but how can these systems (such as customizable avatars or game narrative) be utilized in HCGs?

This raises the question of how to distribute rewards to players. If multiple reward systems are available, is it enough to randomly distribute rewards to players or allow them to pick which rewards they want? On the one hand, players who are incentivized to play for a particular type of reward may find themselves compelled to contribute for longer or faster in order to receive the rewards they prefer, at the risk of frustrating players who might not appreciate randomly-distributed rewards. On the other, giving players a choice of reward may allow players to enjoy the rewards they prefer and possibly also incentivize them to contribute better quality work, at the risk of running out of content for reward systems or distracting them from the underlying human computation task. Ideally, we desire a reward distribution system that is fair to the players (i.e., providing a quality *player experience*), but also respects any needs of the task (i.e., ensuring quality *task completion*) and the limitations of content within these systems.

To explore these questions, we built a game called *Cafe Flour Sack*. *Cafe Flour Sack* is a culinary-themed HCG that asks players to classify cooking ingredients for potential recipes. It contains four different reward systems (or reward *categories*) for players to interact with: global leaderboards, customizable avatars, unlockable narratives, and a global progress tracker. These systems were chosen to appeal to a broad audience of players and thus cover a variety of different motivations for play (e.g., such as those expressed in [28]), while remaining representative of reward systems in modern digital games. Leaderboards and customizable avatar systems have appeared in prior HCGs, while narrative was designed to address alternative motivations in a way that would not interact or interfere with the other rewards. Finally, the global tracker was added to accomodate a potential player population that derives motivation intrinsically by participating in learning or crowdsourcing, but not from extrinsic rewards.

Figure 1. The four reward systems in *Cafe Flour Sack*. Starting clockwise from the upper-left: the global leaderboards, the customizable avatar, the progress tracker, and the unlockable narratives.

We now describe these four reward systems:

- **Global Leaderboards** In the global leaderboards, "leaderboard" currency is automatically used to increase players' rank relative to other players. Figure 1 shows a screenshot of the leaderboards in the upper-left corner. After each round of tasks, players can check their leaderboard rank, which is represented as a medal (or badge) in the leaderboard menu. All players are added to the leaderboards by default, but players who do not receive leaderboard currency (or choose not to) remain at the default rank.

- **Customizable Avatars** In the customizable avatar system, players spend their "avatar" currency to purchase digital items that are used to customize a 2D avatar of a chef. These items include chef-themed clothing and culinary objects. While these kinds of virtual avatar systems are common in commercial games and content distribution platforms, they are rarely seen in HCGs (with one exception [5]). Figure 1 shows a screenshot of the customizable avatar in the upper-right corner.

- **Unlockable Narratives** In the unlockable narrative system, players use their "narrative" currency to unlock short stories set in the universe of the game. These stories are presented as conversational dialogue between the player and in-game characters, and are unlocked in sequencial order. Figure 1 shows a screenshot of the leaderboards in the bottom-left corner.

- **Global Progress Tracker** In the global progress tracker, players may view statistics showing their overall contribution to the tasks being completed by all players in the game.

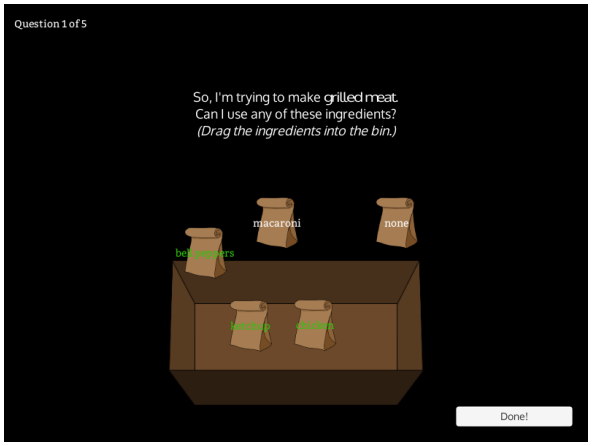

Figure 2. An example minigame from *Cafe Flour Sack*. Here, the player drags all ingredients that can be used in a corresponding recipe ("grilled meat") into a bin.

Figure 3. Screenshots of the reward selection screen between the two versions of the game. On the left, the *random* version selects a reward category (in this case, the avatar category) automatically. On the right, the *choice* version allows the player to click on their preferred category.

Figure 1 shows a screenshot of the progress tracker in the bottom-right corner. These statistics (number of players, recipes completed, etc.) are automatically updated each time a player completes a round. This system is meant to appeal to the intrinsic motivation of wanting to participate; consequently, it automatically increases when players complete tasks - and does not require any additional interaction. Instead, it exists merely to inform players of their progress relative the overall progress of the cooking task.

Cafe Flour Sack's cooking task is an artificial task with a known answer, which allows us to evaluate the efficacy of its reward mechanics without the complications of needing to simultaneously solve a human computation problem. This experimental approach of using an artificial task has been used successfully in prior HCG research in order to evaluate HCG design [14, 20]. We chose ingredient-recipe classification due to its similarity to other classification and commonsense-knowledge problems, as well as its simplicity (as players did not need actual culinary training, but merely knowledge of what ingredients could be used in classes of recipes). For this experiment, we used a gold-standard answer set containing 157 common cooking ingredients and 24 recipes. Each ingredient either belonged to a given recipe or not, and could belong to multiple recipes.

To ensure that the effects of each reward system could be measured independently of each other, each reward system has its own "currency" or point system. Currencies are not interchangeable between systems. Progression in one system does not impact progress in another - nor do any of the reward systems feed back into the gameplay of solving the task (e.g., players cannot purchase "powerups" to assist with the minigames).

METHODOLOGY

Cafe Flour Sack was released as an online game. Upon starting the game, players are placed into one of two versions of the game, *random* and *choice*, which serve as the two conditions in a between-subjects experiment. The game version changes how players will be assigned rewards. In the *random* version, the player is automatically assigned one of the three reward categories at the beginning of each round. In the *choice* version, the player is allowed to manually select one of the three categories at the beginning of each round. Visibly and interactively, the only difference between the two game versions is the reward selection screen, as shown in Figure 3.

Players solve tasks by completing small minigames in rounds of five. Each minigame presents the player with a recipe and four possible ingredients to select from (as either belonging to the recipe or not). Figure 2 shows an example of one of these minigames. After completing a round, players are awarded currency in one of the reward systems. The amount of currency a player receives ranges from zero to five, equivalent to the number of tasks successfully completed.

The game begins with a short tutorial round of five minigames, after which players are given currency in all three possible reward categories. They are then instructed to view each of the reward menus in order to use their points, thus introducing them to all of the reward systems, before progressing further in the game. Players are then allowed to complete as many tasks as they desire throughout the duration of the experiment. At the end of the experiment, players are asked to fill out a post-game survey. Throughout gameplay, the game continually logs data for both tasks and player actions.

We recruited players (participants) from two populations. The first was through Amazon Mechanical Turk. Previous work has successfully explored the use of paid crowdsourcing platforms, such as Amazon Mechanical Turk, for distributing HCGs [14, 21]. Cafe Flour Sack was made available as a task (HIT) on Amazon Mechanical Turk's online portal where workers were compensated for playing the game, then answering the post-game survey. The second group was recruited through an undergraduate computer science class and were compensated for course credit for writing a report on the game (again, after playing the game and taking the same post-game survey).

The Amazon Mechanical Turk workers represent a group of players who are highly-skilled at crowdsourcing work, but are performing it through a monetarily-compensated interface (and thus not necessarily through HCGs). The student population represents an audience likely to be familiar with

games, but not necessarily crowdsourcing work. Thus, when compared with university students, Amazon Mechanical Turk workers may be considered crowdsourcing experts and are likely to encompass a wider range of demographics (such as age range). Part of our long term goal is to broaden the accessibility of HCGs, so we deliberately chose to not only evaluate our work across two different experimental conditions, but two different audiences as well — something that has not been done in prior HCG research.

Because we are interested in understanding engagement in the context of rewards, we took some additional steps to account for the fact that players might have extrinsic motivations for completing the task quickly. First, we required all participants to play for at least 20 minutes, during which they were allowed to freely allocate their time between interacting with the reward systems and completing tasks (and thus yielding additional currency for the reward systems). This was meant to ensure that players would not be incentivized to rush through the experiment as quickly as possible, in which case it would be optimal to avoid interaction with the reward systems at all. Similarly, we also did not require that players complete a certain number of tasks.

Second, we introduced a button in the game's main menu, which we refer to as the "boredom" button. Players were explicitly asked to press the button when they would have considered quitting the game under non-experimental conditions (i.e., had they been playing the game without time enforcement or financial compensation). Pressing the button was optional and did not have any impact on whether or not players on Amazon Mechanical Turk were compensated.

Finally, we wished to ensure that players who completed the study later would not be biased by the presence and progression of earlier players in reward systems with visible social elements — namely the leaderboard and the progress tracker. In order to preserve the social elements of the study while maintaining consistency across all players, we simulated both the leaderboards and progress tracker using a set of fake players and results. After each round of the game, these players were updated (including the addition of new fake players to the game) with artificial progress in both in the leaderboards and the progress tracker.

RESULTS

The study was conducted over the course of several weeks, during which the game was made available online both to workers on Amazon Mechanical Turk and a university student population. We report on results from 78 players who took part in the study. 40 players were placed in the *random* condition and 38 were placed in the *choice* condition. 39 players were workers from Amazon Mechanical Turk (randomly selected from a larger population of 59 workers) and 39 players were students.

In total, 24 players self-reported as female and 54 players self-reported as male. Most players reported themselves as 18-40 years old. Additionally, most players reported prior gaming experience (around 80%); however only 15 players (around 20%) reported any prior experience with HCGs.

	Random	*Choice*
AMT Workers	0.725	0.722
Students	0.670	0.725
Total	0.696	0.724

Table 1. Mean task scores split by experimental condition, first broken down into separate player audiences and then shown in total.

Our evaluation focuses on both the results of the task (*task completion*) and the *player experience*. We investigate differences between the experimental conditions of *random* and *choice*. Additionally, we investigate differences between the two populations of our player audience — Amazon Mechanical Turk workers and students — while accounting for interaction effects with experimental condition. The majority of our dependent variables had nonparametric distributions. To measure differences and interactions between the conditions, unless otherwise stated, we used two-way ANOVAs with aligned rank transforms [27] to account for the nonparametric nature of the data. Below, we report our results; we then discuss them in the subsequent section.

Task Completion

To evaluate the *task completion*, we considered three metrics: the answer correctness, the number of tasks completed, and the timing of task completion. These metrics reflect the design considerations of task providers. For an actual human computation task, different metrics might be prioritized over others depending on the task requirements; here, we observe all metrics equally.

Correctness of Completed Tasks

To verify answer correctness, each task — the pairing of four cooking ingredients with a recipe — was assigned a score. This score was computed using our gold-standard answer set and is the ratio of correctly-assigned ingredients to the total number of ingredients in the task. A task was considered correct if 75% (a corresponding score of 0.75) or more of its ingredients belonged to the given recipe.

The results show that both experimental condition and player audience had significant effects on answer correctness. Players in the *choice* condition had higher mean scores than players in the *random* condition, 0.723 vs. 0.696 ($F = 9.474, p < 0.01$). Amazon Mechanical Turk players had higher mean scores than student players, 0.724 vs. 0.692 ($F = 9.072, p < 0.01$).

The player audience $\times$ experiment condition interaction was significant ($F = 28.648, p < 0.001$). Table 1 shows the mean task scores split across experimental condition and player audience. Amazon Mechanical Turk players in the *random* condition demonstrate the highest mean scores (0.7254) with student players in the *choice* performing closely behind (0.7245). Meanwhile, student players in the *random* condition demonstrated the lowest mean scores (0.670).

Number of Completed Tasks

We also looked at the number of tasks completed per player across both experimental condition and player audience. We broke down our observations into three categories: the *total* number of tasks completed, the number of *correct* tasks

	Random	*Choice*
AMT Workers	8.382	9.129
Students	14.229	12.753
Total	11.492	10.507

Table 2. Mean task completion times (in seconds) for total tasks split by experimental condition, first broken down into separate player audiences and then shown in total.

completed, and the number of *incorrect* tasks completed. On average, Amazon Mechanical Turk players provided significantly more total answers (82.308 answers) compared with student players (70.128 answers) ($F = 5.083, p < 0.05$). Additionally, when looking only at correct answers, Amazon Mechanical Turk players also provided significantly more correct answers (57.410 answers) compared with student players (44.590 answers) ($F = 5.083, p < 0.05$). No other significant effects were observed across experimental conditions and player audience.

Timing of Completed Tasks

For our final task completion metric, we looked at the time (in seconds) it took players to complete tasks. Similarly to our observations of number of tasks completed, we evaluated these results across *total* tasks, *correct* tasks, and *incorrect* tasks.

When it came to the number of seconds it took players to complete all (total) tasks, both experimental condition and player audience had significant main effects. Players in the *choice* condition showed faster mean times for total task completion than players in the *random* condition, 10.507 seconds vs. 11.492 seconds ($F = 8.228, p < 0.01$). Meanwhile, Amazon Mechanical Turk players showed faster mean times for total task completion than student players, 8.788 seconds vs. 13.652 seconds ($F = 281.682, p < 0.001$).

There were also interaction effects. When accounting for experiment condition $\times$ player audience interaction across all tasks, we also found a significant effect ($F = 40.875, p < 0.001$). Table 2 shows the mean task completion times for all tasks split across experimental condition and player audience. Overall, Amazon Mechanical Turk players in the *random* condition demonstrated fastest mean times (8.382 seconds), and are slightly slower in the *choice* condition (9.128 seconds). This result is flipped across conditions for student players, who demonstrated faster mean times in the *choice* condition (12.753 seconds) compared with the slowest mean times in the *random* condition (14.229 seconds).

Next, when looking only at the times it took players to complete tasks correctly, we found that once again, both experimental condition and player audience had significant effects (however no interaction effects were observed). Players in the *choice* condition were faster at completing tasks correctly than players in the *random* condition, 9.773 seconds vs. 10.820 seconds ($F = 5.809, p < 0.05$). Meanwhile, Amazon Mechanical Turk players were faster at completing tasks correctly than student players, 8.348 seconds vs. 12.780 seconds ($F = 190.930, p < 0.001$).

		Leaderboards	*Avatar*	*Narrative*	*Tracker*
Random	AMT Workers	4	8	5	0
	Students	13	6	2	2
Choice	AMT Workers	14	2	6	0
	Students	8	3	5	0
Total		39	19	18	2

Table 3. Counts of players' favorite rewards across both experimental condition and player audience.

Similarly, when looking only at the times it took players to complete tasks incorrectly, both experimental condition and player audience had significant effects. Players in the *choice* condition were slightly faster at completing tasks incorrectly than players in the *random* condition, 12.262 seconds vs. 12.726 seconds ($F = 10.222, p < 0.01$). Again, Amazon Mechanical Turk players were faster at completing tasks incorrectly compared to student players, 9.802 mean seconds vs. 15.174 mean seconds ($F = 21.868, p < 0.001$). Significant effects for experimental condition *x* player audience interaction were also observed ($F = 43.596, p < 0.001$). Amazon Mechanical Turk players were faster overall (at 9.371 mean seconds in the *random* condition and 10.167 mean seconds in the *choice*). Student players were slower (at 14.991 and 15.531 mean seconds in the *random* and *choice* conditions respectively).

In summary, players in the *choice* condition had faster mean times for task completion than players in the *random* condition. Additionally, Amazon Mechanical Turk players were significantly faster than completing tasks than student players at completing tasks. These findings were observed not just for all tasks, but also for tasks answered correctly and tasks answered incorrectly. For total tasks, Amazon Mechanical Turk players in the *random* condition were the fastest at completing tasks, while students in the *random* condition were the slowest. For incorrectly-answered tasks, Amazon Mechanical Turk players in the *random* condition were the fastest, while students in the *choice* condition were the slowest.

Player Experience

Our evaluation of the player experience consists of observations of player interaction, combined with player responses to questions on the post-game survey. In particular, we are interested in understanding how players engaged with the reward systems, as well as why they may have become disengaged with these systems. We first on report player survey responses regarding their favorite and least favorite reward systems in *Cafe Flour Sack*, and a question of whether or not players perceived they had a choice of reward systems. Next, we report on their their interaction time within each of the reward systems. Finally, we report their interaction with the boredom button in order to understand why they would have disengaged with the game — and if our reward systems were responsible.

Reward Preference

First, we were interested to know how players responded to each of the different reward systems available. In the post-game survey, players were asked to provide their favorite

		Leaderboards	*Avatar*	*Narrative*	*Tracker*
Random	AMT Workers	3	4	7	3
	Students	3	4	13	3
Choice	AMT Workers	3	4	8	7
	Students	2	6	7	1
Total		11	18	35	14

Table 4. Counts of players' least favorite rewards across both experimental condition and player audience.

reward system and their least favorite system in *Cafe Flour Sack*. For players' favorite reward system, 39 players selected the leaderboards, 19 players selected the narrative rewards, 18 players selected the customizable avatar, and 2 players selected the progress tracker. Table 3 shows the exact breakdown of players' favorite rewards across the experimental condition and player audiences.

Meanwhile, regarding players' least favorite reward system, 35 players selected the narrative, 18 players selected the customizable avatar system, 14 players selected the progress tracker, and 11 players selected the leaderboards.

We found no differences or effects on task performance based on players' favorite and least favorite reward systems.

Perception of Choice

We looked at whether or not players perceived they had a choice of rewards available, which we will refer to as "perception of reward choice". In the post-game survey, players were asked to rate the statement "I was able to choose which rewards I wanted." on a Likert-like scale from 1 to 5 (1 corresponding to "Strongly Disagree", 5 corresponding to "Strongly Agree").

Both experimental condition and player audience had significant main effects on players' "perception of reward choice." In the *choice* condition, players reported significantly higher "perception of reward choice" than in the *random* condition ($F = 73.631, p < 0.001$). Amazon Mechanical Turk players reported higher perception of reward choice than student players ($F = 5.548, p < 0.05$). No significant interaction effects were detected.

Duration of Play

As previously mentioned, interaction within the game was limited to 20 minutes. For players who were participating in this study through Amazon Mechanical Turk, it is likely that were already incentivized to participate for financial reasons. (Amazon Mechanical Turk also imposes a time limit for submitting task results, so players would have been unlikely to continue playing under this additional time pressure.) Under these limitations, we cannot look at total duration of play as an indication of engagement or retention.

Instead, we look where *how* players spent their time during those 20 minutes of play. In particular, we are interested to see how long players spent in each of the different reward systems. Each system had its own dedicated interface and we recorded how long players spent in these interfaces. Some of these systems, in particular, the leaderboards and the progress tracker, show very short durations, as interaction is limited to viewing information such as leaderboard rank or task progress. In comparison, the narrative system required players to read and actively click through character dialogue. Table 5 shows the mean time spent in each reward menu, broken down by experimental condition and player audience.

Leaderboards	*Random*	*Choice*
AMT Workers	10.174	7.828
Students	12.829	10.174
Customizable Avatar	*Random*	*Choice*
AMT Workers	11.157	10.939
Students	12.039	12.694
Narratives	*Random*	*Choice*
AMT Workers	45.093	60.980
Students	47.070	58.036
Global Tracker	*Random*	*Choice*
AMT Workers	6.068	6.808
Students	10.278	7.654

Table 5. Mean duration (in seconds) for spent in all four reward systems across both player audience type and experimental condition.

In the leaderboards, both experimental condition and player audience had a significant main effect on the duration of interaction. Players in the *random* condition spent longer in the leaderboards than players in the *choice* condition ($F = 7.319, p < 0.01$), with a mean time of 11.904 seconds vs. 8.868 seconds. Student players spent much longer in the leaderboards than Amazon Mechanical Turk players ($F = 7.265, p < 0.01$), with a mean time of 11.795 seconds vs. 8.650 seconds. No interaction effects were observed.

No significant differences in duration of interaction were observed between experimental conditions and player audience for the remaining reward systems: the customizable avatar, the unlockable narrative, and the global progress tracker.

Boredom

62 of the 78 players pressed the boredom button. Of these players, 32 were in the *random* condition (80% press rate) and 30 were in the *choice* condition (79% press rate). 34 of these players were Amazon Mechanical Turk players and 28 were student players. When looking at the times (since the start of the game) at which the boredom button was pressed, no significant differences were detected between the experimental conditions and the player audience.

Additionally, players were asked to clarify why they had pressed the boredom button (if they had chosen to do so). Overall, 26 players (around 42% of players) described their primary reason for pressing the boredom button as due to the repetitive nature of tasks (i.e., lack of variety in the tasks or tasks that were too similar). 10 players described their main reason as due to finishing or running out of reward content. Other reasons included a lack of interest in the task and game overall (10 players), general confusion or unfamiliarity with

certain ingredients (4 players), a lack of challenge (3 players), and a lack of purpose and/or learning (3 players).

Given that the task was repetitive in nature (and addressing these issues for boredom would involve looking at gameplay mechanics beyond the scope of this study), we looked more closely at the 10 players who described boredom due to finishing and running out of reward content, as this is directly related to reward systems. Of these players, 4 were in the *random* condition and 6 were in the *choice* condition, while 8 players were Amazon Mechanical Turk players and 2 were student players. A majority of these players (6 of 10) listed their favorite reward as the unlockable narrative, with 2 more preferring the customizable avatar, and the last 2 preferring the leaderboards.

DISCUSSION

What considerations for the design of reward systems in human computation games can we draw from our results?

With multiple rewards systems, offering players the choice of reward is both effective and engaging.

Overall, players in the *choice* condition demonstrated higher task correctness and were faster at completing tasks. Additionally, players in the *choice* condition perceived they had had more choice of rewards. This however, did not appear to significantly affect interactions with the reward systems themselves as we found no differences in the duration of interaction, suggesting that the lengths of player experiences were similar. The only exception to this was that players in the *random* condition spent longer in the leaderboards, but these differences, while significant, were only on the order of several seconds. We conclude that offering players the choice of reward benefits both *task completion* and the *player experience*. While other explorations of mechanics in HCGs have shown potential trade-offs in *task completion* and *player experience* [20] (and thus may require balancing design decisions for maximizing one aspect of HCGs over the other), the *choice* condition showed benefits for both.

Adjusting reward mechanics can make certain player audiences perform more effectively.

Overall, Amazon Mechanical Turk players performed significantly better than student players at all task completion metrics (task correctness, number of tasks completed, and rate of task completion), which is unsurprising given that Amazon Mechanical Turk players are considered crowdsourcing experts. As previously mentioned, Amazon Mechanical Turk players in the *random* condition were the most effective players overall, significantly so when it came to both task correctness and rate of task completion. However, these differences in *task completion* metrics, compared to the next most effective population, are significant but small. When separating students by experimental condition, students in the *choice* condition have *task completion* metrics more comparable to those of Amazon Mechanical Turk players. This is not the case with the *random* condition, where the difference in *task completion* metrics is much larger. So while our two player audiences performed very differently on *task completion* in one experimental condition (students significantly lower than Amazon Mechanical Turk players in *random*), they were comparable in the other (*choice*). Our findings are limited because our task was selected for its simplicity, relying on primarily on commonsense knowledge without additional training. However, for more complicated tasks, such improvements could be very valuable. Combined with the previous consideration, this suggests that design decisions such as offering players choice of multiple rewards have the potential to greatly improve *task completion* metrics without negatively affecting the *player experience*.

Small changes in the design of reward mechanics can have large impacts on task completion and the player experience.

A design concern unique to HCG design is determining which gameplay elements have the most significant effects on both the *task completion* and the *player experience*. The difference between the *random* and the *choice* versions of the game was a single screen that assigned or allowed players to choose their reward before completing a round of gameplay. In this study, we showed this fairly simple design change for in the presentation and acquisition of rewards could had significant affects on both *task completion* and the *player experience*, in particular managing to improve results for a non-expert player audience. At the same time, the interaction effects between how we reward players and player audience highlight the importance of paying attention to the target player audience. This appears to be especially true in the context of reward systems and their mechanics. To the best of our knowledge, existing HCG research has not deeply examined how different player audiences might affect HCGs, not to mention tailoring subsets of HCG game mechanics within a single game to different audiences. This study helps to confirm the importance of reward mechanics to both *task completion* and the *player experience*.

LIMITATIONS AND FUTURE DIRECTIONS

Our study is limited by the number of users. This is due in part to the fact that conducting studies on Amazon Mechanical Turk is also prohibitively more expensive (both financially and logistically compared to the majority of tasks on the platform with extremely short durations). Additionally, many steps were taken to address factors in the study confounded by financial or academic compensation, which possibly affected aspects of gameplay interaction players would have had in a non-experimental setting. For example, we simulated the presence of social elements (artificial players) to avoid bias, but it is not clear how this compares to the presence of real social elements. Finally, reward systems in many digital games often contain interacting elements (e.g., exchangeable reward currencies) or are entangled with other game mechanics. Our setup necessitated keeping the systems separate to observe experimental effects, thus possibly limiting the kinds and implementations of reward systems.

Going forward, we believe there are many possible investigations enabling a better understanding of reward systems in human computation games. We utilized multiple reward systems in *Cafe Flour Sack*, some of which are present in existing HCGs and others which have never been examined before. While leaderboards were the preferred reward in *Cafe Flour Sack*, many players also expressed preferences for other underutilized systems. We also found no correlations between

players who selected leaderboards with higher *task completion* or *player experience* metrics, suggesting that other reward systems might be viable for inclusion in HCGs. This raises questions such as whether leaderboards are the most effective reward system for all tasks and all audiences? Was a dislike of the unlockable narrative due to its particular implementation or because these particular audiences were unengaged by the content in this context?

Answering such questions would require undertaking a direct comparison of the different reward systems (including others not explored in this study) and seeing their effects on *task completion* and the *player experience*. Based on our explorations in this study, investigating leaderboard alternatives might focus on more neutrally-favored systems (e.g., the customizable avatar) over more polarizing systems (e.g., the unlockable narrative). This, however, comes with some considerations. Implementing many or multiple kinds of reward systems puts an additional burden on HCG developers, not just for their implementation, but generation of content as well. While the most frequently-cited reason for player boredom with the game was due to the repetitive nature of the tasks, we note that the next-most identified reason for boredom (affecting 12% of players) was due to running out of or finishing reward content. These players showed a clear preference for reward systems with finite content (the unlockable narrative and the customizable avatar), suggesting that a population of players was in fact deeply-engaged with these systems and performed enough work to exhaust all of the content in them. In order for these systems to be effective for potential player populations such as this, the amount of available reward content must match the amount of desired (or estimated) human computation work required per player, something that is of concern to HCG developers.

Other aspects of rewards, such as reward contingencies (what players were rewarded for) and schedules (when rewards were received), were kept constant for this study to reduce the number of variables, but also merit separate investigation for their effects on task performance and player engagement. Additionally, while our setup prohibited us from conducting fully qualitative interviews (i.e., not violating Amazon Mechanical Turk's Terms of Service), a deeper, detailed understanding of what motivates players to engage with HCGs — and how these findings fit within existing motivational and crowdsourcing frameworks for compensation — is imperative to making more effective and engaging HCGs based on player feedback.

CONCLUSIONS

In this paper, we explored the use of multiple reward systems in human computation games and the effect of changing how these rewards are distributed to players. Studying the impact of design decisions in HCGs is crucial to helping scientists, researchers, and game developers create more effective and engaging games. We ran a study comparing two versions of a cooking-themed HCG, *Cafe Flour Sack*, containing multiple reward systems. One version of the game (*random*) randomly distributed rewards to players and the other version of the game (*choice*) that allowed players to choose between possible reward systems. We released this game to two different player audiences and studied the effects of the conditions as they relate to the two main design considerations of HCGs: *task completion* and the *player experience*.

We observed several main and interaction effects, such as that players in the *choice* condition solved tasks more correctly and perform tasks quicker. Unsurprisingly, we also found that Amazon Mechanical Turk players proved to be significantly better at solving tasks than student players. Overall, Amazon Mechanical Turk players in the *random* condition had the highest *task completion* metrics, but all other players in the *choice* condition were not far behind (with student players in the *random* condition demonstrating significantly lower *task completion* metrics). When it came to aspects of the *player experience*, we found that players in the *choice* condition perceived they had more choice of rewards, but there were few differences in their interaction with the reward systems (with leaderboards being the only exception). Additionally, student players were more engaged along these metrics than Amazon Mechanical Turk players.

Based on our results, we suggest that offering players *choice* of rewards leads to better *task completion* and a more engaged *player experience*. Interaction effects suggest that reward mechanics are sensitive to both our experimental conditions and player audiences, but we can leverage reward mechanics to improve *task completion* without negatively affecting the *player experience* of one audience (students) compared to another (Amazon Mechanical Turk workers). Finally, we discuss our limitations and future work in reward mechanics for HCGs. Ultimately, our goal is to help HCGs become more effective and engaging for both task providers and players, and that our investigations from this study help to clarify the design space of reward systems in these games.

ACKNOWLEDGMENTS

We thank members of the Entertainment Intelligence Lab for providing feedback on the study and the game. We also thank Eric Butler and Eleanor O'Rourke for valuable feedback and assistance.

This material is based upon work supported by the National Science Foundation under Grant No. 1525967. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Richard Bartle. 1996. Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research* 1, 1 (1996), 19.
2. Dongseong Choi and Jinwoo Kim. 2004. Why people continue to play online games: In search of critical design factors to increase customer loyalty to online contents. *CyberPsychology & behavior* 7, 1 (2004), 11–24.
3. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and others. 2010a. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

4. Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver-Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, Zoran Popović, and Foldit Players. 2010b. The Challenge of Designing Scientific Discovery Games. In *5th International Conference on the Foundations of Digital Games*.
5. Dion Hoe-Lian Goh, Chei Sian Lee, Alton YK Chua, Khasfariyati Razikin, and Keng-Tiong Tan. 2011. SPLASH: Blending Gaming and Content Sharing in a Location-Based Mobile Application. *Social Informatics* (2011), 328.
6. Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, and Kuan-Ta Chen. 2009. KissKissBan: A Competitive Human Computation Game for Image Annotation. In *ACM SIGKDD Workshop on Human Computation (HCOMP '09)*.
7. Markus Krause and Jan Smeddinck. 2011. Human computation games: A survey. In *Signal Processing Conference, 2011 19th European*. IEEE, 754–758.
8. Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. Frontiers of a paradigm: exploring human computation with digital games. In *ACM SIGKDD Workshop on Human Computation*.
9. Edith Law and Luis von Ahn. 2009. Input-agreement: A New Mechanism for Collecting Data Using Human Computation Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1197–1206.
10. Robin D Laws. 2002. *Robin's laws of good game mastering*. Steve Jackson Games.
11. Chei Sian Lee, Dion Hoe-Lian Goh, Alton YK Chua, and Rebecca P Ang. 2010. Indagator: Investigating perceived gratifications of an application that blends mobile content sharing with gameplay. *Journal of the American Society for Information Science and Technology* 61, 6 (2010), 1244–1257.
12. Brian Magerko, Carrie Heeter, and Ben Medler. 2010. Different Strokes for Different Folks: Tapping Into the Hidden. *Gaming and Cognition: Theories and Practice from the Learning Sciences: Theories and Practice from the Learning Sciences* (2010), 255.
13. Winter Mason and Duncan J Watts. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11, 2 (2010), 100–108.
14. Winter Mason and Duncan J Watts. 2011. Collective problem solving in networks. *Available at http://dx.doi.org/10.2139/ssrn.1795224* (2011).
15. Kou Murayama, Madoka Matsumoto, Keise Izuma, and Kenji Matsumoto. 2010. Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences* 107, 49 (2010), 20911–20916.
16. Ei Pa Pa Pe-Than, Dion Hoe-Lian Goh, and Chei Sian Lee. 2013. A typology of human computation games: an analysis and a review of current games. *Behaviour & Information Technology* (2013).
17. Andrew K Przybylski, C Scott Rigby, and Richard M Ryan. 2010. A motivational model of video game engagement. *Review of general psychology* 14, 2 (2010), 154.
18. Ganit Richter, Daphne R. Raban, and Sheizaf Rafaeli. 2015. *Gamification in Education and Business*. Springer International Publishing, Chapter Studying Gamification: The Effect of Rewards and Incentives on Motivation, 21–46.
19. Marta Sabou, Kalina Bontcheva, Arno Scharl, and Michael Föls. 2013. Games with a Purpose or Mechanised Labour?: A Comparative Study. In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (i-Know '13)*. ACM, New York, NY, USA, Article 19, 8 pages.
20. Kristin Siu, Alexander Zook, and Mark O. Riedl. 2014. Collaboration versus Competition: Design and Evaluation of Mechanics for Games with a Purpose. In *9th International Conference on the Foundations of Digital Games*.
21. Stefan Thaler, Elena Simperl, and Stephan Wolger. 2012. An experiment in comparing human-computation techniques. *Internet Computing, IEEE* 16, 5 (2012), 52–58.
22. Kathleen Tuite. 2014. GWAPs: Games with a Problem. In *9th International Conference on the Foundations of Digital Games*.
23. Kathleen Tuite, Noah Snavely, Dun-Yu Hsiao, Nadine Tabing, and Zoran Popović. 2011. PhotoCity: training experts at large-scale image acquisition through a competitive game. In *ACM SIGCHI Conference on Human Factors in Computing Systems*.
24. L. von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319–326.
25. Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM* 51, 8 (2008), 58–67.
26. Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, 55–64.
27. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. DOI: **http://dx.doi.org/10.1145/1978942.1978963**
28. Nick Yee. 2006. Motivations for play in online games. *CyberPsychology & behavior* 9, 6 (2006), 772–775.