A Framework for Exploring and Evaluating Mechanics in Human Computation Games

Kristin Siu  
School of Interactive Computing  
Georgia Institute of Technology  
kasiu@gatech.edu

Alexander Zook  
School of Interactive Computing  
Georgia Institute of Technology  
a.zook@gatech.edu

Mark O. Riedl  
School of Interactive Computing  
Georgia Institute of Technology  
riedl@cc.gatech.edu

ABSTRACT

Human computation games (HCGs) are a crowdsourcing approach to solving computationally-tractable tasks using games. We outline a formal representation of the mechanics in HCGs, providing a structural breakdown to visualize, compare, and explore the space of HCG mechanics. We present a methodology based on small-scale design experiments using fixed tasks while varying game elements to observe effects on both the player experience and the human computation task completion. Ultimately, we wish enable easier exploration and development of HCGs, letting these games provide meaningful experiences to players while solving difficult problems.

CCS CONCEPTS

- Human-centered computing → Computer supported cooperative work;  
- Applied computing → Computer games;

KEYWORDS

human computation games; games with a purpose; scientific discovery games; game design; game mechanics

ACM Reference format:


1 INTRODUCTION

Games are everywhere. For human computation games (HCGs), games which harness the computational potential of the human crowd, this diverse, increasing audience of players presents new opportunities to solve complex, computationally-tractable tasks or generate data through gameplay. Already, HCGs—also known as Games With a Purpose (GWAPs), scientific discovery games, and citizen science games—have been used to solve a variety of problems such as image labeling, protein folding, and data collection.

However, one hurdle compounding HCG development compared with that of mainstream games for entertainment is that these games suffer the design problem of serving two different goals. On the one hand, an HCG must provide a sufficiently-engaging experience for its players. On the other hand, an HCG must enable players to successfully complete the underlying human computation task. Balancing these two goals is difficult, often resulting in conflicting design decisions. Unfortunately, very little design knowledge exists beyond a small number of simple patterns from examples or take-aways from successful games (e.g., [6, 25]). As a result, most HCGs to date are built around specific kinds of templates, leaving the space of possible HCG designs limited and relatively unexplored.

To facilitate broader adoption and ease of game development, HCG design needs the tools and frameworks to study and communicate about these games in a consistent manner. We need to understand precisely what game elements make certain HCGs successful, that is both effective at engaging players and solving tasks. A common language and structure for HCGs would allow us to talk about and explore the space of possible HCG designs.

In this desiderata, we describe a formal representation of HCG mechanics that provides us with a common vocabulary and structure to visualize, compare, and explore the space of game mechanics in HCGs. We advocate a methodology for building up HCG design knowledge, which uses small-scale, controlled design experiments on tasks with known solutions to understand how variations of game elements affect the player experience and the completion of human computation tasks. Further details and illustrative examples of how our framework enables the comparative study of existing HCGs and the exploration of novel HCG mechanics can be found in an extended version of this paper [21].

2 BACKGROUND

Human computation games have been developed as an alternative to traditional crowdsourcing systems, providing players an engaging gameplay experience while utilizing game mechanics to enable task completion. The original Game With a Purpose, the ESP Game, addressed the problem of labeling images [24]. Since then, HCGs have been used to annotate or classify other kinds of information, from music [2, 14], to relational information [12, 19], to protein function recognition [18]. Other HCGs have leveraged human players as alternatives to optimization functions for "scientific discovery" problems such as protein [5] and RNA folding [15], DNA multiple sequence alignment [10], and software verification [7]. Additionally, HCGs have been used to collect or generate new information, such as creative content or machine-learning datasets. Examples include photo collection [23], location tagging [3], and commonsense knowledge acquisition [13]. Comprehensive taxonomies [11, 17] detail a wide breadth of HCGs and their tasks.

HCG design has been primarily guided by examples of successful games. These include von Ahn and Dabbish’s templates for classification and labeling tasks [25] and the design anecdotes of Foldit [6] rather than systematic study of HCG elements. While these are
We specifically formalize game mechanics—the rules that define how a player can interact with the game systems—leaving other elements of HCG designs to future work. We divide HCG game mechanics into three types: action mechanics, verification mechanics, and feedback mechanics. As shown in Figure 1, this breakdown reflects the core gameplay loop of most HCGs. HCGs begin with players taking in-game actions, then compare task-relevant input from these actions through verification mechanisms, and finally use verification output to provide feedback or reward for players.

We now define and describe these three sets of mechanics in detail, illustrated using three successful HCGs spanning different tasks: the original ESP Game [24], Foldit [5], and PhotoCity [23]. Figure 2 shows the mechanical breakdown of these games into action, verification, and feedback mechanics. Further examples and discussion can be found the extended version of this paper [21].

3.1 Action Mechanics

Action mechanics are the interface for players to complete a human computation task through in-game actions or gameplay. These mechanics align with the process of solving the human computation task, often asking players to utilize skills necessary for solving the task during play. Such mechanics may be as simple as entering text input or as complicated as piloting a space ship in a virtual environment, and tend to vary based on the nature of the task.

Examples. In the ESP Game, players provide labels through text entry to solve the task of labeling given images. In Foldit, players are given a variety of spatial actions, such as handling or rotating components of a protein structure, to solve the task of “folding” a given protein into a minimal energy configuration. In PhotoCity, players navigate to a desired location and take pictures using their camera phones, which are later uploaded to a database and used to construct a 3D representation of the buildings in that location.

3.2 Verification Mechanics

Verification mechanics combine the output of player actions to compute task-relevant outcomes. These mechanics can support task completion outcomes including the quality, volume, diversity, and the rate at which the data are acquired.

Examples. For many human computation tasks, consensus on player input often serves as verification. The ESP Game (and other structurally-similar games) verify using an online agreement check that filters correct answers from incorrect answers using agreement between players (Figure 2). The ESP Game later added “taboo word” mechanics to promote data diversity through banning words once consensus on existing data was reached.

By contrast, both Foldit and PhotoCity accomplish verification through task-based evaluation functions. Foldit’s protein configuration energy function determines the quality of player solutions online. In PhotoCity, the game does not explicitly evaluate the provided photos; photos are instead processed on an offline server and then player feedback is based on the resulting alterations to a constructed 3D mesh of the world.

Foldit also makes use of social mechanics, such as allowing players to share solution procedures (called "recipes") through its community interfaces, as an additional (but optional) instance of verification [4]. Players can utilize existing recipes uploaded by other players as a starting point for solving tasks, thus validating and iterating on pre-existing, partial solution strategies.

3.3 Feedback Mechanics

Feedback mechanics provide players with information or digital artifacts based on the results of player actions in terms of partial or full task completion. These mechanics commonly encompass gameplay elements such as rewards and scoring, and can also be mapped to evaluation metrics for the underlying task, thus allowing both researchers and designers to assess player performance at both the completion of the task and progression through the in-game experience.
Examples. For all of the games shown in Figure 2, players receive feedback in the form of a score. However, the scale of the scoring mechanics themselves are unique to the tasks performed. The ESP Game rewards players with points for agreement on an image label. By contrast, Foldit rewards players with points for minimizing an energy function describing the protein structure. PhotoCity rewards players for the number of points their photo choices add to the reconstructed 3D mesh. These examples are all similar in that the feedback “currency” is nominal—points contributing to a numerical score—but vary in what players are rewarded for.

4 A METHODOLOGY FOR HCG DESIGN

Our mechanics representation provides a breakdown of the different kinds of mechanics in human computation games. This enables us to identify where we can focus our explorations of the HCG design space, but not how we should explore the space in order to build up generalizable design knowledge.

We highlight a methodology of controlled A/B design experiments that explore the space of HCG designs, using formal representations for game elements, tasks, and audiences. In the context of HCG mechanics, this manifests as between-subjects (alternatively, within-subjects) experiments comparing separate versions of HCGs with different mechanical variations.

These design experiments should (1) implement a task with a known solution, while (2) focusing on a single element of a HCG’s design. First, testing with a known solution allows us to evaluate task-related metrics objectively without simultaneously solving a novel problem. Such known solutions may be the result of presolved human computation problems (e.g., image labeling datasets) or simpler tasks that are analogous to existing problems. Second, focusing on one particular element of an HCG’s design allows us to understand exactly what kind of impact an element may have on both players and the task with minimal interaction effects. Our mechanics representation can be used to assist us in understanding where and how the introduction of an element might affect the HCG game loop.

These experiments should simultaneously evaluate how design decisions meet the needs of players and tasks. Optimizing only for the player may result in a game with engaging mechanics that do not effectively solve the human computation task. Optimizing only for the task may result in a game that players do not find engaging enough to play even if the human computation task can be solved effectively. We refer to these two axes of metrics as the player experience and the task completion.

Player experience encompasses metrics such as:

- **Engagement**: how players interact with the game or rate their experience with it
- **Retention**: how likely are players to continue playing
- **Quality**: correctness or accuracy of task results
- **Volume**: amount of completed tasks
- **Diversity**: the variation or breadth of task results
- **Rate of Acquisition**: how quickly tasks are completed

Task completion refers to the task-related metrics such as:

- **Quality**: correctness or accuracy of task results
- **Volume**: amount of completed tasks
- **Diversity**: the variation or breadth of task results
- **Rate of Acquisition**: how quickly tasks are completed

The exact metrics to test for often depend on the nature of the human computation task and the HCG’s target player audiences. For example, HCGs with tasks requiring trained players to solve them effectively may consider metrics such as player retention much more important than HCGs for simpler tasks where maintaining a skilled player base is not a priority.

We note that this methodology is not new, as similar experimental approaches have recently been applied to HCGs. Here, we cite two such examples. Goh et al. [8] compared a non-gamified...
control application for image labeling against two versions of the ESP Game, one using collaborative scoring mechanisms and the other using competitive scoring mechanisms. Similarly, Siu et al. [20] conducted an experiment with the game Cabbage Quest, utilizing a task with a known solution—categorizing everyday objects with purchasing locations—to compare two variants of scoring mechanisms: one collaborative and one competitive. Both of these experiments follow our proposed methodology of taking a problem with a known solution or gold-standard answer set, testing design elements by treating a set of game mechanics as independent variables, and measuring aspects of both the player experience and task completion.

5 CONCLUSIONS

In this paper, we outline a framework for designing and studying human-computation games. We described a formal representation of HCG mechanics into three types: action, verification, and feedback, and illustrated it with several examples. Additionally, we highlighted a methodology of running design experiments on known tasks that measure both player experience and task completion.

Human computation games have demonstrated the potential to solve complex and difficult problems, but must be both engaging experiences for players and effective at solving their tasks. As games become more pervasive, HCGs must compete for players’ attention, and thus must remain relevant and consistent with player expectations. To ensure this, we need to understand how HCGs work, to build better and broader generalizable design knowledge that can adapt to new games, tasks, and audiences, especially when HCG developers do not typically have the training or resources of professional game studios. Our framework is designed to explore and evaluate HCG mechanics so that it will be easier to design and develop successful, effective HCGs. In doing so, we hope to work towards a future where HCGs are engaging, effective, ubiquitous, and empowering.

ACKNOWLEDGMENTS

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


