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-intractable 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:
Kristin Siu, Alexander Zook, and Mark O. Riedl. 2017. A Framework for Exploring and Evaluating Mechanics in Human Computation Games. In *Proceedings of FDG'17, Hyannis, MA, USA, August 14-17, 2017,* 4 pages.
DOI: 10.1145/3102071.3106344

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-intractable 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 takeaways 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

FDG'17, Hyannis, MA, USA
 978-1-4503-5319-9/17/08…$15.00
DOI: 10.1145/3102071.3106344

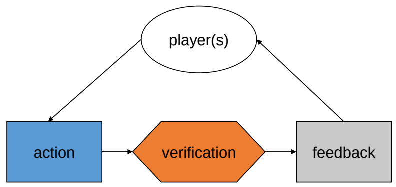

Figure 1: Breakdown of HCG mechanics. Players provide inputs to take actions (in blue), which are verified (in orange), and receive feedback (in gray) from the game. Solid lines represent transitions through the gameplay loop.

useful, we do not understand what specific elements of these particular design choices work and how to appropriately generalize them or consider new alternatives. Confounding this issue is the fact that HCG research remains divided on how game elements, in particular game *mechanics*, can ensure both engaging player experiences and successful completion of tasks. Some argue that HCG game mechanics should be *isomorphic* or *non-orthogonal* to the underlying task [9, 22] while others argue that incorporation or adaptation of game mechanics from successful digital games designed for entertainment can keep players more engaged [12].

Controlled studies utilizing quantitative and qualitative methods to study the influence of general game design elements have been widely used in analogous, dual-purpose domains such as education [1, 16]. It is only recently that researchers have conducted similar studies on specific game elements of HCGs that jointly address aspects of the player experience and the completion of the human computation task [8], and advocated for their use [20]. Combined with formal crowdsourcing research, we posit that these approaches can enable a formal study of HCG design.

3 FORMALIZING HCG MECHANICS

We outline a formal representation of the mechanics of human computation games. This representation serves three core functions:

(1) Provides a common vocabulary and visual organization of HCG elements
(2) Enables formal comparison of existing HCGs to understand the space of HCG designs and their consequences
(3) Facilitates the formulation of controlled design experiments of HCG elements to build further, generalizable knowledge of HCG design

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.

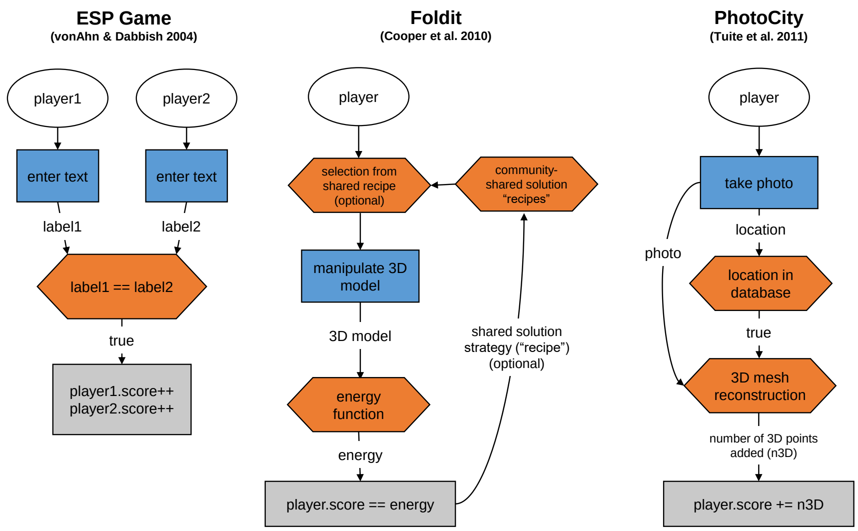

Figure 2: Examples of HCGs [5, 23, 24] subdivided into *action*, *verification*, and *feedback* mechanics. Arrows from feedback to players have been omitted for clarity.

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 pre-solved 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
- Other subjective measures related to how players interact and perceive the game (e.g., preferences, unstructured self-reported feedback)

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

[1] Erik Andersen, Yun-En Liu, Richard Snider, Roy Szeto, Seth Cooper, and Zoran Popović. 2011. On the Harmfulness of Secondary Game Objectives. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG '11)*. ACM, New York, NY, USA, 30–37.

[2] Luke Barrington, Douglas Turnbull, and Gert Lanckriet. 2012. Game-powered machine learning. *Proceedings of the National Academy of Sciences* 109, 17 (2012), 6411–6416.

[3] Marek Bell, Stuart Reeves, Barry Brown, Scott Sherwood, Donny MacMillan, John Ferguson, and Matthew Chalmers. 2009. EyeSpy: Supporting Navigation Through Play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 123–132.

[4] Seth Cooper, Firas Khatib, Ilya Makedon, Hao Lu, Janos Barbero, David Baker, James Fogarty, Zoran Popović, and Foldit players. 2011. Analysis of Social Gameplay Macros in the Foldit Cookbook. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG '11)*. ACM, New York, NY, USA, 9–14.

[5] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and others. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[6] 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. 2010. The Challenge of Designing Scientific Discovery Games. In *5th International Conference on the Foundations of Digital Games*.

[7] Werner Dietl, Stephanie Dietzel, Michael D. Ernst, Nathaniel Mote, Brian Walker, Seth Cooper, Timothy Pavlik, and Zoran Popović. 2012. Verification Games: Making Verification Fun. In *14th Workshop on Formal Techniques for Java-like Programs*.

[8] Dion Hoe-Lian Goh, Rebecca P. Ang, Chei Sian Lee, and Alton Y. K. Chua. 2011. Fight or Unite: Investigating Game Genres for Image Tagging. *J. Am. Soc. Inf. Sci. Technol.* 62, 7 (July 2011), 1311–1324.

[9] Peter Jamieson, Lindsay Grace, and Jack Hall. 2012. Research Directions for Pushing Harnessing Human Computation to Mainstream Video Games. In *Meaningful Play*.

[10] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, Jérôme Waldispühl, and Phylo Players. 2012. Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One* 7, 3 (2012), e31362.

[11] Markus Krause and Jan Smeddinck. 2011. Human computation games: A survey. In *2011 19th European Signal Processing Conference*. IEEE, 754–758.

[12] Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. Frontiers of a Paradigm: Exploring Human Computation with Digital Games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 22–25.

[13] Yen-ling Kuo, Jong-Chuan Lee, Kai-yang Chiang, Rex Wang, Edward Shen, Cheng-wei Chan, and Jane Yung-jen Hsu. 2009. Community-based Game Design: Experiments on Social Games for Commonsense Data Collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, New York, NY, USA, 15–22.

[14] 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 (CHI '09)*. ACM, New York, NY, USA, 1197–1206.

[15] Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Sungroh Yoon, Adrien Treuille, Rhiju Das, and EteRNA Participants. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2122–2127.

[16] Derek Lomas, Kishan Patel, Jodi L. Forlizzi, and Kenneth R. Koedinger. 2013. Optimizing Challenge in an Educational Game Using Large-scale Design Experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 89–98.

[17] 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).

[18] Mark Peplow. 2016. Citizen science lures gamers into Sweden's Human Protein Atlas. *Nature Biotechnology* 34, 5 (2016), 452–452.

[19] Katharina Siorpaes and Martin Hepp. 2008. Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems* 23, 3 (2008), 50–60.

[20] Kristin Siu, Alexander Zook, and Mark O. Riedl. 2014. Collaboration versus Competition: Design and Evaluation of Mechanics for Games with a Purpose. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

[21] Kristin Siu, Alexander Zook, and Mark O. Riedl. 2017. A Framework for Exploring and Evaluating Mechanics in Human Computation Games. (2017). arXiv:1706.03311

[22] Kathleen Tuite. 2014. GWAPs: Games with a Problem. In *Proceedings of the 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1383–1392.

[24] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 319–326.

[25] Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM* 51, 8 (2008), 58–67.