DESIGN AND EVALUATION OF INTELLIGENT REWARD STRUCTURES IN HUMAN COMPUTATION GAMES

A Dissertation
Presented to
The Academic Faculty

By

Kristin A. Siu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing
School of Interactive Computing

Georgia Institute of Technology

August 2021

DESIGN AND EVALUATION OF INTELLIGENT REWARD STRUCTURES IN HUMAN COMPUTATION GAMES

Thesis committee:

Dr. Mark Riedl, Advisor
College of Computing
Georgia Institute of Technology

Dr. Betsy DiSalvo
College of Computing
Georgia Institute of Technology

Dr. Blair MacIntyre
College of Computing
Georgia Institute of Technology

Dr. Seth Cooper
Khoury College of Computer Sciences
Northeastern University

Dr. Brian Magerko
Ivan Allen College of Liberal Arts
Georgia Institute of Technology

Date approved: June 2, 2021

This dissertation is dedicated to Eric

who also undertook the graduate school journey.

We both made it.

ACKNOWLEDGMENTS

This document would not exist without the people that I have been fortunate to meet on this journey.

I would like to thank my advisor, Dr. Mark Riedl, for being an irreplaceable, unfailing presence on this long adventure. No current story generator can come close to crafting a narrative expressing how grateful I am for all the support you have given me. (And I hope that maybe someday, your research makes it possible.)

I would like to thank my thesis committee members: Dr. Blair MacIntyre, Dr. Brian Magerko, Dr. Betsy DiSalvo, and Dr. Seth Cooper for overseeing this process and providing invaluable feedback.

I would like to thank my past collaborators and co-authors: Eric Butler, Alexander Zook, Matthew Guzdial, Yunfei Bai, C. Karen Liu, Peizhao Zhang, Jianjie Zhang, and Jinxiang Chai for all of the publications and projects we worked on. While not all of these publications are reflected in this document, every one of them made me a stronger researcher.

I would like to thank my colleagues in the Entertainment Intelligence Lab, particularly those who were present while I was physically in the lab: Alexander Zook, Matthew Guzdial, Upol Ehsan, Chris Purdy, Lara Martin, Brian O'Neil, Boyang Li, Zhiyu Lin, Xinyu Wang, Brent Harrison, and Sasha Azad. Whether it was academic discussions or Pokémon doodles, your support made each day I came into the lab both fulfilling and fun.

I would like to thank my former colleagues in the Computer Graphics Group at Georgia Tech: Karthik Raveendran, Jie Tan, Sehoon Ha, Mark Luffel, Topraj Gurung, and Tina Zhuo for discussions and assistance that got me through my first years of grad school.

I would like to thank my former advisors in the Computer Graphics Group at Georgia Tech: Dr. C. Karen Liu and Dr. Jarek Rossignac for teaching me that convex optimization is hard, and that beautiful things can be built out of triangles and circles.

I would like to thank all of my study players, participants, and interviewees for their time and contributions to this work. Without you, this work would literally not have been possible.

I would like to thank my past and current industry colleagues for embracing that your fellow engineer was also an academic, and allowing me prove that academic research can be more that just seemingly impenetrable, esoteric, or intimidating.

I would like to thank four beloved hamsters—Amaterasu ("Ammy"), Artemis ("Missy"), Persimmon ("Minmin"), and Lilikoi ("Lili")—for keeping me company on long nights during paper and presentation deadlines. Rest in peace, my beloved *Phodopus roborovskii*.

Finally, I acknowledge that portions of this work were supported by the National Science Foundation under Grant No. IIS-1525967. Any opinions, findings, conclusions, and recommendations expressed in this document are those of the author and do not necessarily represent the official views or positions of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Despite the ubiquity of artificial intelligence and machine learning, some problems and procedures have no or few effective algorithmic solutions. Some of these problems—such as building commonsense knowledge understanding or generating creative works—are considered or straightforward for humans to solve, as they often rely on intrinsic human processes or capabilities that cannot be easily modeled by machines. *Human computation* is the process of taking these computationally-intractable problems and leveraging human problem solving with algorithmic aggregation, often by subdividing the overarching problem into smaller *tasks* whose solutions can be developed through human interaction, consensus, and/or optimization.

Human computation games (HCGs) are playful, game-based interfaces for tackling these kinds of crowdsourced problems. These games—also known as citizen science games, scientific discovery games, Games with a Purpose, etc.—have been used to tackle problems that were and still are considered complex for computational algorithms: accurately tagging images to power Google image search, designing real-life proteins to be synthesized, powering 3D reconstructions of buildings, and to generating vast data sets of creative artifacts (to name a few).

However, despite these successes, HCGs have not seen broad adoption compared to other types of serious (e.g., educational) digital games. Among the many reasons for this lack of adoption is the reality that these games are typically not seen as engaging or compelling to play, in part because designing an ideal HCG must optimize for *both* providing an engaging *player experience* and successfully solving the *underlying task*. Too often, existing HCGs and HCG research focus only on optimizing these games for the task with little to no consideration of an engaging player experience. This is in turn exacerbated by the fact that creating HCGs comes at a high development cost (compared to other online crowdsourcing platforms) to task providers who are typically not game design or develop-

ment experts. As a result, HCGs continue to languish, perceived as niche or unengaging experiences with their full potential woefully underexplored.

One potential solution to address this problem is to ensure that the necessary resources and design knowledge are there to help task providers build better HCGs. But while the digital games industry has no shortage of both anecdotal and empirically-validated, contemporary design and development resources (e.g., courses, conferences, frameworks, etc.), HCGs lack any sort of formalized design knowledge base or understanding of how existing game design resources may apply to building these games. Building such HCG-specific design knowledge is not straightforward, given that it must address and leverage both the understanding of making compelling *player experiences* and effectively solving the *underlying task*. This thesis is a one step in building and establishing this better understanding of what game elements make an HCG both engaging and effective.

In this thesis, I explore *reward mechanics* in human computation games. Game mechanics are rules defining player interactions with a game's systems and *reward mechanics* define the *reward systems* which are responsible for providing player feedback. Understanding *reward mechanics* is integral in HCGs due to their role and association with player motivation, player compensation, and task validation. I first propose a framework for understanding HCG mechanics and advocate for an experimental methodology considering both *player experience* and *task completion* metrics to understand variations in HCG mechanics. I then use these tools to frame and design three experiments that explore small-scale variations of *reward systems* in HCGs looking at three adjustable aspects of reward systems: reward functions, reward distribution, and reward personalization. These studies demonstrate that even small-scale variations in rewards (i.e., offering players the ability to choose the type of reward) may have significant positive effects on both *player experience* and *task completion* metrics. I also show that some variations (i.e., co-located, competitive reward scoring) may have both positive and negative tradeoffs across these metrics. Moreover, this work observes that existing, anecdotal design wisdom for HCGs may not always hold

(i.e., allowing players to verbally collude does not, in fact, lower task solution accuracy, but actually predicts higher task solution accuracy). Altogether, this thesis demonstrates that certain aspects of *reward systems* in HCGs can be varied to improve *player experience* without compromising *task completion*, and works to build more empirically-tested design knowledge for understanding and creating more engaging, effective HCGs.

CHAPTER 1
INTRODUCTION

1.1 Human Computation Games

Artificial intelligence and machine learning have been leveraged to solve a wide variety of difficult and complex problems, from autonomously operating vehicles to generating creative experiences. However, for certain problems, even the best algorithmic solutions remain intractable for computers to tackle. Meanwhile, other problems require training data that cannot be automatically generated or easily aggregated. One characteristic these types of problems share is that their solutions typically depend on knowledge that is intrinsic to humans and/or require complex models of human processes that lack straightforward, computational representations.

How might one tackle these problems when no direct algorithmic answers are available (or at best, require a domain-specific, knowledge-intensive, and potentially human-biased solution)? One alternative is to leverage human aptitude directly through an approach formally known as *human computation* [1]. As the name suggests, human computation is a paradigm that combines the skills of human problem solvers with algorithmic aggregation of their results, typically (but not necessarily) through a computer interface. This outsourcing of computational work to a human "crowd" is also commonly referred to as *crowdsourcing*.

The standard *human computation* approach to solving a difficult problem broadly consists of the following steps. First, the problem—referred to as the *task*—is subdivided or duplicated into smaller *subproblems*. Next, each subproblem is given an individual person—often referred to as a *worker*—to solve. Subsequently, each resulting solution is algorithmically recombined and then verified with all other individually solved results.

Finally, the aggregate combination of these results is used to reach a consensus about or determine a solution to the original task.

A classic example of applying human computation to a computer science problem is that of *image recognition* [2]—the (natural language) identification of the objects in a two-dimensional image. While myriads of other approaches to image recognition have been since been explored, the *human computation* approach leverages the human visual recognition process directly by asking workers to solve the problem (as opposed to architecting a complex computational system that mimics their problem-solving abilities), the results of which may be used directly or used as data to train other image recognition systems. An image is distributed to multiple workers, who then provide natural language words (i.e., labels, tags, descriptors) that describe its contents. The most commonly suggested words may then be taken as the labels or descriptors of its content. This straightforward verification process is effective because the domain of natural language words is large, making the likelihood of agreement on a label (by random chance) very low. But human computation is not limited simply to computer science problems. Other broad classes of problems addressed using human computation approaches include media classification [3], biological optimization [4, 5], and content generation [6, 7]—most of which rely on human problem solving skills or human creativity to accomplish.

Currently, the most common interfaces for human computation work are online services that give task providers a globally accessible platform to distribute tasks, recruit human workers, and compensate them for their work. Popular services include Amazon's *Mechanical Turk* and *CrowdFlower*, which provide online, end-to-end platforms for tackling human computation tasks. For these types of interfaces, compensation is primarily monetary, although other human computation efforts, particularly those addressing scientific discovery problems, rely on volunteer efforts motivated by altruistic goodwill or public recognition.

Games have been proposed as an alternative interface to online crowdsourcing services

and workflows. These games are known by a variety of names: *human computation games*, *Games with a Purpose*, scientific discovery games, citizen science games, and crowdsourcing games. Henceforth, I will refer to them as *human computation games*, or *HCGs*. HCGs span a wide range of experiences ranging from lightly gamified web forms to fully interactive, immersive games where the human computation task is solved through player actions during play. As such, these games share many commonalities with *serious* games designed for purposes beyond pure entertainment, such as education, training, self-improvement, or addressing real-world issues in politics and society.

So why might one advocate for using games as an alternative to traditional crowdsourcing platforms? One argument for the use of games is that certain game elements may align with the requirements of a particular task. For example, players' in-game actions may accurately capture the steps necessary to complete a given task. Additionally, flexible game elements can be used to direct or to encourage players to more appropriately solve tasks. For example, if a large number of task solutions is required, an HCG could reward players for completing as many tasks as possible, whereas if a small quantity of high quality task solutions is more desirable, the same HCG could instead reward players for more accurate results. In an ideal case, in-game feedback, such as the player's score, may help to inform the player (or task provider) how well the player may be completing a given task.

Another argument is that otherwise mundane or simple tasks might be made more engaging through gameplay—engaging enough that the gameplay experience might itself be considered a form of compensation and/or motivation for completing the task. For example, games used to collect real-world data (e.g., photos) can help to turn a mundane and repetitive process (e.g., manually taking many photos of certain objects, photos at very specific locations) into a more engaging experience. This makes games an attractive option if monetary compensation is unsuitable or prohibitively expensive for a particular task. At a minimum, an engaging game might assist in making tasks less burdensome or help to recruit and retain players who might not normally participate in crowdsourcing work.

As distribution platforms have increased and hardware costs have lowered, games have become a more pervasive part of people's everyday lives. Accompanying these changes, tools for game development have become more sophisticated and accessible for non-technical developers. This might suggest that task providers and developers of human computation games would wish to take advantage of these growing audiences and better development tools. However, beyond a few successful instances, HCGs have not seen wide adoption, let alone appeared to have benefited from this growing ubiquity of play and ease of creation.

One possible explanation for the lack of adoption is that despite these improvements in game development technology, the cost of developing a game still exceeds that of exposing a simple crowdsourcing task on an existing platform such as Amazon's *Mechanical Turk*. Putting together a web form and utilizing an existing community of crowdsourced workers requires substantially less effort and resources than developing (or contracting someone to develop) an interactive experience with a time-consuming process for design, engineering, and quality assurance. As a result, task providers who are typically not expert game developers might be reluctant to turn to games as the interface of choice in order to accomplish their crowdsourced needs.

However, even when tools and development assistance are readily available, human computation games present a nonstandard design challenge. Unlike other games designed primarily to entertain, HCGs have an additional purpose: completing the underlying task. Thus, HCGs have two, potentially conflicting design goals: completing a human computation task and entertaining the game's players, which reflect the needs of the task providers and players respectively. On the one hand, task providers may not care about the details of the gameplay experience, but are concerned with the quality and quantity of the crowdsourced results. On the other hand, players may not even care about the task itself, but are concerned with the quality of the experience and entertainment resulting from their interaction with the game. A *truly effective* HCG must address both issues. It must enable players to complete the task correctly, but must be beneficially entertaining enough to motivate

players to interact with the game at all.

While there is a wealth of literature and readily available resources on general game design, very little of it is tailored specifically for the design of human computation games. Given that HCGs must also optimize for the task completion in addition to an engaging player experience, it is unclear how existing design knowledge might even apply to HCGs. Existing design knowledge for HCGs is currently limited to a handful of papers and ad-hoc guidelines suggesting game templates or providing anecdotes from a few successful examples. While these results provide useful starting points for the amateur HCG developer or task provider, it is not always clear *which* specific game elements or combination of them are responsible for ensuring both successful task completion and a positive player experience, let alone how they might apply to a novel problem or unexplored considerations of a potential HCG. Which game elements can be changed without negatively impacting task completion and/or the player experience? Furthermore, how will these guidelines generalize to expanding player audiences, and changes in technology? These are just two of the many questions that remain unanswered in the existing space of HCG design.

Therefore, with little existing design knowledge to work with and no strong guarantees to know whether or not a task could even be completed effectively using a game, it is no wonder why task providers might be reluctant to invest the time and resources into game development when other, more straightforward options are available. As a result, the design space of human computation games remains woefully underexplored, the full potential of these games remains undetermined, and the perception persists that these games are both ineffective and unengaging.

Ideally, formalized design knowledge for human computation games would exist and task providers would not necessarily need to be expert game developers. Instead, they would be armed with tested principles, models, and design guidelines—a validated design knowledge base tailored for HCGs—that would help them select game elements appropriate for their given task while crafting a compelling gameplay experience. The long-term,

overarching goal of this dissertation work is to help enable such a foundation of HCG design.

1.2 Reward Mechanics

Creating a human computation game begins much like the standard development of any game designed to entertain: by making a myriad of design decisions related to the elements that compose the experience—gameplay mechanics, visual and aural aesthetics, to state but a few. However for HCGs, this design process is complicated by the fact that any single element may have potential effects on both the desired player experience and the task completion results. Understanding the effects of all possible game design decisions for HCGs is considerably intractable, so the question then becomes: what game elements should be prioritized? It would be ideal to focus on game elements that have a noticeable impact on both aspects of the player experience and the underlying task. For this dissertation, I propose focusing on game *mechanics*, specifically *reward mechanics*. In games, the *mechanics* are the rules and systems associated with player interaction. *Reward mechanics* are the game mechanics responsible for providing player feedback, which encompass common game systems (or parts of such systems) such as scoring, leaderboards, and item acquisition.

So why should rewards be emphasized? In games, rewards are the currency of positive feedback. More formally, *reward mechanics* are the rules of the underlying systems that provide (primarily positive) feedback to players. These reward systems can take many forms, but can be distinguished from each other by the currency or type of feedback they provide to players. Common reward systems expose numerical scores (e.g., points, ranks) or collectible digital objects (e.g., badges, equipment), which are *explicit* nominal metrics distributed based on functions that measure aspects of player performance or progression. For example, in the classic action game *Super Mario Bros.*, the player encounters various types of reward systems. In the game, players progress through increasingly dangerous 2D

levels; each level has a numerical score which increments when the player accomplishes certain actions, such as collecting coins strewn throughout the level, defeating enemies, and completing the level within the time limit. Additionally, players may seek out powerups (often placed in challenging-to-access places), which reward the player with extra protection when traversing the level or defeating enemies. Meanwhile, other rewards cannot be as easily quantified and may be earned *implicitly* as a consequence of other mechanics or extradiegetic game elements, such as the social prestige based on competitive game rankings or the subjective enjoyment derived from experiencing a game's narrative. In *Super Mario Bros.*, one such implicit reward may be the sense of satisfaction may feel upon reaching the end of the game, whereupon they may view a cutscene showing the successful rescue of an in-game character. [1]

In game design, reward mechanics are considered to be some of the most important game elements because they play a pivotal role in providing feedback to players [8]. A player's experience with a game is typically driven by factors such their preferences for particular game elements and motivations for play. Reward systems are often the most powerful communicators of the intended experience—and thus are also powerful motivators for play. Ultimately, these systems influence whether or not a player has a positive (or negative) experience with a game—and determine the likelihood that they will continue playing it or not.

This is no different for human computation games; if anything, the role that reward mechanics serve may be even more critical. HCGs rely on players to ensure that a task is completed at all. Since reward mechanics affect the likelihood that a player will continue to interact with the game and thus complete the task(s), I argue that the HCG research community would benefit from the study of reward mechanics and their effects on task results both the player experience and the task results. I now highlight two additional

[1] As mentioned, whether or not an implicit reward can be considered "rewarding" is entirely subjective. For example, I do not derive any sense of satisfaction upon viewing this scene of Princess Peach's rescue due to an outright dislike of "save the princess" narratives in games.

reasons why reward mechanics are both complicated and sensitive—and thus imperative to study—in the context of HCGs.

First, motivations for players to participate in human computation games are complicated by the addition of the human computation task. Unlike in other games, players may also derive implicit reward via a sense of altruism or satisfaction from contributing to the human computation or problem solving process. This means that some players may be intrinsically motivated by the task itself whereas other players might only be compelled to participate provided that the game mechanics (including the rewards which provide feedback on their performance) provide a satisfying experience [9, 10]. It is well-established in general human computation and crowdsourcing research that worker motivation and compensation type affect the completion of tasks. For example, intrinsically motivated workers are known to be turned off by extrinsic (i.e., monetary) compensation for crowdsourced work [11], making them less effective in certain scenarios. However, little is known about whether or not this holds for HCGs, where typical reward systems such as points and associated leaderboards are treated as extrinsic compensation. Therefore, understanding how reward systems in HCGs correspond to player motivations and which kinds of rewards are effective given different kinds of players is necessary to improve the state of HCG design.

Second, in-game rewards are commonly compared to or considered alternatives to other forms of compensation for human computation and crowdsourcing work. Approaches to crowdsourcing as well related applications of gamification (i.e., applying game elements to a process normally absent of play) have both raised concerns regarding the ethical treatment of workers and players and the fairness of compensation. HCGs share many similarities to these applications in design goals and public perception. Therefore, it is morally necessary to ensure that reward mechanics provide both commensurate and appropriate feedback to players through a positively-received player experience. This requires a deep understanding of how reward mechanics function in HCGs and how they affect the players who interact with them, not just their effects on the completion of the human computation task.

To reiterate, human computation games and their mechanics, particularly rewards, must respect the needs and desires of the players involved. Failure to do so—by blindly optimizing only for the human computation task at any serious expense to the player experience—has long term consequences for HCGs as a whole. In the worst case, HCGs would garner a reputation as being disrespectful of players' time and effort. Without an engaged player base, task providers are even more likely to decry these games as ineffective. This would make HCGs an unappealing option for solving human computation tasks—long before we can explore and understand their full potential. In order to prevent this, the HCG research community needs to understand how game elements of human computation games truly work, and in doing so, enable the creation of more effective HCGs.

1.3 Thesis

> Varying properties of reward systems in human computation games can improve the player experience while also improving or maintaining the quality, quantity, and acquisition rate of the human computation task results.

In this dissertation, I begin by presenting a framework and methodology for conducting human-subjects studies on human computation games to evaluate both *task* and *player experience* metrics. I then use this framework and methodology to conduct three human-subjects studies on HCG reward systems to support this statement. Table 1.1 provides a summarizing chart of the questions, methods, data collection, and contributions of these experiments.

Table 1.1: Summary of the research questions explored in this dissertation and their contributions. The **metrics** term refers to the combination of *task completion* and *player experience* metrics as defined by the methodology in Chapter 3.

Research Questions	**Methods**	**Data Collected**	**Contributions**
Chapter 4			
In *multiplayer*, what effect will *collaborative* versus *competitive* reward scoring have on **metrics**?	A within-subjects study with two conditions: *singleplayer* versus *multiplayer*. In the *multiplayer* condition, a between-subjects study for two further conditions: *collaboration* versus *competition*.	Gameplay telemetry, Likert-like/rating post-round/post-game survey results, and manual tracking of multiplayer *collusion*.	*Competitive* multiplayer was perceived as more compelling than *collaborative* multiplayer, but also resulted in less accurate task results. Overall, participants strongly preferred both *multiplayer* conditions over *singleplayer*.
Would *collusion* (i.e., direct communication between players regarding the task) result in lower **metrics**?			*Collusion* between participants was a predictor of higher task accuracy.
Chapter 5			
What effect will *randomly assigning* versus *letting participants choose* reward type have on **metrics**?	A between-subjects study with two sets of conditions: reward assignment—*random* reward assignment versus player reward choice—and audience—*experts* versus *amateurs*.	Gameplay telemetry (with a focus on reward system interaction and self-reported boredom) and Likert-like/free response post-game survey results.	Participants able to *choose rewards* demonstrated higher task and more positive experience metrics.
What effect will using an *expert* crowdsourced worker audience versus *amateur* (student) worker audience have on **metrics**?			*Experts* were better than *amateurs* at all task metrics. However, *amateurs* performed comparably to *experts* when given the ability to *choose rewards*.
Chapter 6			
For more time-consuming tasks, what effect will *not-personalizing* versus *personalizing* the type of reward have on **metrics**?	A between-subjects study with two conditions: *no personalization* versus *personalization*.	Gameplay telemetry (with a focus on reward system interaction and more time-consuming task completion) and Likert-like/free response pre/post-game survey results.	No differences were detected between the *non-personalized* and *personalized* conditions.

CHAPTER 2
BACKGROUND

In this chapter, I provide a general overview of human computation games. Each subsequent chapter will review its own relevant and background material:

1. Chapter 3 presents an overview of game design frameworks, existing game design knowledge for HCGs, and an overview of experimental methods in games research as part of a separate background section.

2. Chapter 4 discusses the study of relevant game mechanics—singleplayer versus multiplayer and collaboration versus competition—in both entertainment-oriented games as well as HCGs—as part of its introduction.

3. Chapter 5 provides an overview of rewards in games, research surrounding rewards, and how compensation affects crowdsourcing as part of its introduction.

4. Chapter 6 presents an overview of personalization and adaptive systems, which dynamically-adjust game elements based on player interaction in games, as well as the player modeling work which contributes to these systems, as part of its introduction.

Human computation games are an alternative, interactive interface for approaching and solving crowdsourcing problems. These games are known by a variety of different names [12]: *Games With a Purpose* (GWAPs), scientific discovery games, citizen science games, crowdsourcing games, and knowledge games. In this dissertation, I refer to them as *human computation games*, abbreviated as *HCGs*. The choice of "human computation games" over other phrases is intended to encompass the broad set of games where humans participate in the task or problem solving process (i.e., computation) through gameplay. Schrier's aforementioned survey outlines the various (and potential) pros and cons of the

term “HCGs” versus others. My choice of term was independent of her work, but I have eschewed other names for reasons similar to her observations. While my early publications used the term “Games with a Purpose,” I have since transitioned from the use of this term to avoid conflation with other serious games—almost all of which have goals or intended purposes beyond entertainment. Arguably, all serious games are thus “games with a purpose” (albeit not in the narrow definition originally utilized by the human computation community [2]). Meanwhile, “scientific discovery,” “citizen science,” and “crowdsourcing” games are associated with specific subsets of the broader collection of games in which humans solve tasks through gameplay. I will discuss these subsets below.

Here, I provide an overview of existing human computation games, focusing on both the tasks they have tackled and the interactive experiences they have provided players. For convenience, I split these games into three broad categories based on the objective output and the problem-solving process required to complete the task:

1. **Classification**: augmenting existing data with annotations, labels, or categories

2. **Data Manipulation and Optimization**: developing solutions to or refining partial solutions of (often scientific) optimization problems

3. **Data Collection and Artifact Creation**: aggregation or collection of raw data, possibly refined or filtered during the human computation process

This categorization is not intended to be exhaustive or prescriptive. Detailed taxonomies such as those of Law [1], Krause and Smeddinck [13], Thaler *et al*. [14], and Pe-Than *et al*. [15] provide detailed breakdowns and organizations of HCGs, also based on task type and similarities of game elements. These taxonomies are unfortunately limited to older examples of HCGs (i.e., those developed prior to the year 2013) and thus do not take more recent games or developments into account. Similar to these prior taxonomies, this overview groups these games by the tasks they have tackled (although into the three broad

categories above, as opposed to a narrow focus on the individual tasks). Additionally, I dedicate attention to recent HCGs not fully covered by existing taxonomies.

2.1 Classification

The earliest published instances of using games as an interface to tackle crowdsourcing problems began with games that relied on human players to classify or annotate data. Structurally, these tasks begin with unsorted or noisy data as input, then ask workers or players to categorize or provide additional information regarding that input. (In this context, terms such as "annotating," "labeling," and "tagging" may be considered synonymous for generality.) The output domain of acceptable solutions is often impossibly large (e.g., the set of all natural language strings), or sufficiently complex to enumerate or describe. In tandem, modeling the necessary or relevant human decision processes to solve these problems algorithmically is often challenging if not infeasible. Instead, human computation on these *classification* tasks seeks to solve these problems by leveraging human capabilities (e.g., the human visual cortex) or intrinsic knowledge with efficient algorithmic aggregation (e.g., commonsense knowledge acquired over years of existing in modern society).

The earliest—and possibly most seminal—example of a game used to classify data was the *ESP Game*, an online guessing game responsible for popularizing the term *Games With a Purpose* [2]. In the *ESP Game*, players worked in pairs to assign text labels which described (and thus annotated) the content of photographic images. In each instance of the game, a pair of players was networked over the internet, but remained anonymous to each other and thus unable to communicate with their partner. Both players were then shown an identical image and given a time limit. Each player was then responsible for typing as many potential image labels into a text box and whenever both players entered the same word (a match), they were rewarded with in-game points. Every image was distributed to many rounds of players interacting with the game and for every round, the system tracked the most frequently provided matches. The set of matches with the highest number of

occurrences were determined to be labels for the image.

This anonymous, two-player arrangement was designed to encourage players to provide labels that most accurately described the image in question. Players reaching consensus on the same word acted as a means of verifying that label's correctness (as the probability that two arbitrary people might suggest the same word given the entire set of words in the English language is extraordinarily small [1]). Additionally, the *ESP Game* alleviated some of the burden required to build labeled datasets for image labeling algorithms, by turning what was once a time-consuming manual effort on the part of those implementing and hand-tuning image labeling algorithms (e.g., a single or several persons) into a broader distributed effort across many others.

The success of the *ESP Game* encouraged the development of similar games: image-object identification in *Peekaboom* [16], music tagging in *Tag-a-Tune* [3], and image ranking [17]. These were followed up by a variety of other games such galaxy classification in *Galaxy Zoo* [18] and music-tagging in *Herdit* [19]. *Galaxy Zoo* in particular is a notable example, as its initial success led to the development of the *Zooniverse* [20] platform for crowdsourcing that has since gone onto tackle a myriad of human computation problems such as biodiversity tracking and architectural identification.

Other well-explored "classification" problems include relational information acquisition, or the construction of ontologies. In 2008, Siorpaes and Hepp introduced *Ontogame* [21], followed by Krause and *et al*.'s *OntoGalaxy* [22]. Additionally, similar solving of common-sense knowledge problems has also been accomplished for tasks such as object-relationship understanding [23].

Currently, the use of human computation games for classification problems has been largely outstripped by improvements in machine learning algorithms and the convenience of other non-game crowdsourcing platforms such as Amazon Mechanical Turk and Crowdflower. However, a novel trend is to integrate human computation tasks into existing, high-profile games for entertainment, or to use such games as a platform for human computation.

In the classification space, these efforts include *Project Discovery* [24], a high-profile human computation game, due in part to its unique integration with an existing game: the massively-multiplayer online strategy game *Eve Online*. Existing *Eve Online* players may participate in a separate minigame, thematically set in the *Eve Online* universe while earning in-game currency for classifying images for a protein-function recognition task. Additionally, extradiegetic elements of the real world, such as the scientists involved with the project, are integrated as digital characters or elements within the *Eve Online* universe.

2.2 Data Manipulation and Optimization

Beyond asking players to act as human data classifiers, human computation games have also leveraged player skills to tackle scientific optimization problems. Correspondingly, these games are often referred to as *scientific discovery* or *citizen science* games. Typically, the input to these tasks is existing scientific data that must be refined or manipulated to meet a desired goal (which can be occasionally measured using an objective function grounded in existing theories or models of scientific processes). Many of these tasks rely on reasoning or abilities that are difficult to represent algorithmically such as spatial reasoning.

One of the earliest and most well-known examples of *scientific discovery* games is *Foldit* [4], which tasks players with the problem of "folding" proteins structures into their lowest energy configurations. Such low-energy configurations for proteins are considered the most natural, stable solutions for scientists to potentially synthesize, but tackling this algorithmically requires computing a large, multi-dimensional optimization problem. Instead, *Foldit* relies on intrinsic human spatial reasoning skills, presenting players with a 3D model of a potential protein structure. Players then nudge, tweak, and drag molecular components using their mouse and keyboard; all the while, an optimization function computes the energy configuration value based on the components and their structure, providing real-time feedback to players as they work on the task. The results from *Foldit* were then (and continue to be) used to synthesize actual proteins as part of a collaboration with

biochemistry researchers.

Other scientific tasks include RNA folding [5] in *Eterna*, DNA multiple sequence alignment [25] in *Phylo*, mapping dataflow diagrams onto hardware architectures [26], and multigraph maximal-clique discovery [27]. For many of these tasks, players have uncovered noticeably better solutions over existing algorithms, with players' solution methods often serving as guidance to develop new algorithms and optimizations. More recently, *scientific discovery* games have been designed to act in conjunction with computational and/or machine learning algorithms to build or improve existing datasets. For example, in the game *Mozak* [28], players trace neurons in brain images to reconstruct neural pathways. The game is but one form of verification in a larger system, which incorporates both algorithmic optimization and non-player expert verification of solutions to rapidly converge on solutions.

Several data optimization problems have also been integrated into high-profile games intended primarily for entertainment, or utilize these games' mechanics as a platform for human computation. One such example is the game *Phylo*, which after its initial introduction in 2012, expanded to address various other DNA multiple sequence alignment problems (e.g., Ebola virus sequences [29]). Most recently, *Phylo* was integrated into the high-profile game *Borderlands 3* as a minigame called *Borderlands Science* [30], wherein players can tackle multiple sequence alignment tasks under the auspices of an existing in-game character, and earn special in-game currency and loot for their contributions. Another recent example uses the popular, open-world game *Minecraft* to tackle the problem of developing effective compounds for treating chemotherapy-resistant cancers [31]. Players in *Minecraft* are provided in-game block structures representing visualizations of partial compound solutions, and by placing in-game blocks down, work towards solutions that are analyzed in real-time. These games, as well as the aforementioned *Project Discovery* platform, demonstrate how collaborations between the game industry and task providers from research institutions can fruitfully leverage the player audiences of popular games for

human computation solutions.

2.3 Data Collection and Artifact Creation

In addition to classifying and manipulating existing data, human computation games have also tapped into crowdsourcing audiences to collect or generate new data or artifacts. These tasks are similar to *classification* tasks in that players are asked to provide new information (rather than refine or optimize existing solutions) and that the output of both types of tasks is frequently used to train machine-learning systems. However, the key difference is that the output of *classification* tasks is typically metadata (e.g., annotations/labels/mappings) to existing input data, while the output of *data collection* tasks is the wholly new data themselves. This distinction between *classification* and *data collection* tasks may appear subtle, but because the creation of new data is the objective of *data collection* tasks, *data collection* HCGs often require game interfaces and game mechanics specific to generating that data and do not fit neatly into the pre-existing patterns for *classification* HCGs.

One example of a *data collection* game is the mobile HCG *Photocity* [6]. In *Photocity*, players are asked to take pictures of various buildings and environments in order to reconstruct 3D models of the environment. Players accomplish this by physically navigating to a real-world location designated by the game and then taking pictures with their camera phones. These pictures are then uploaded to a separate processing server that utilizes the pictures to reconstruct 3D buildings using structure from motion algorithms. Players are then awarded points based on the number of new vertices added to these 3D models.

Early examples of *data collection* games include those which gather artifacts such as the aforementioned photographic images used for 3D environment reconstruction [6, 32], provide detailed location tags [7], and commonsense knowledge [33]. More recently, data collection involves the creation of artifacts that require players to perform creative inference or require procedures that are difficult for machines to generate naturally. One such example is *Quick, Draw!* [34], wherein players are playing a guessing game against an AI

agent trained on simple, hand-drawn images of nouns. Players' drawings are then subsequently incorporated into a dataset used to train *sketch-rnn* [35], a recursive neural network that can recreate simple hand-drawn images (which in turn, improves the AI agent used in *Quick, Draw!*). Another example—though it may be argued that it is more of a playful interface rather an HCG in the traditional sense—is the *Amino Acid Synthesizer* [36], which maps the problem of protein synthesis to music generation and provides users with 20 musical tones (corresponding to the the 20 amino acids) that can be strung together into compositions. These user-generated music compositions are then used to help train the protein synthesis models.

Again, these tasks, despite surface differences and widely varying game interfaces, share a common structure of having players provide artifacts, which are then checked against a model of how that information should function to provide diverse inputs and coverage of the desired domain. One metric of particular interest in these domains is the diversity of provided data, since these tasks are often open-ended, ambiguous, or have definitions of correctness that are open to player subjectivity.

CHAPTER 3
A FRAMEWORK FOR EXPLORING AND EVALUATING GAME MECHANICS IN HCGS

3.1 Introduction

Creators of digital games oriented towards entertainment have no shortage of resources when it comes to the design and development process. These options span everything from professional online tutorials to libraries of published textbooks to anecdotal postmortem videos to comprehensive educational programs (from primary to post-graduate). Typically, these resources focus on how to craft interactive artifacts which optimize for a positive player experience, focusing on elements and principles that encourage player engagement.

However, when considering specific kinds of games such as serious and educational games, these resources become scarcer. One difference between serious and educational games compared with games designed purely for entertainment, is the addition of a *second*, often equally-important design focus or *goal*. This design focus might be the conveyance of an extradiegetic message, the ensurance of player self-improvement (e.g., educational goals), or the solution to a specific problem through gameplay (e.g., a human computation task). Whereas the typical digital game designed by an entertainment company, independent developer, or hobbyist may only need to consider how well the game engages its players, creators of serious games are burdened with the additional complication of ensuring that their games must both accomplish their additional design goals while still remaining faithful to promise of an entertaining experience.

Human computation game development must contend with this complication of needing to ensure two different design 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, even for expert game developers, as the occasionally mundane or repetitive nature of a human computation task does not always map cleanly to engaging game elements. This tension (or worse, unawareness of the problem altogether) often results in games which prioritize one goal over the other (typically the *task*, as these games are more often created by task providers rather than entertainment companies).

To compound this dilemma, very little design knowledge exists beyond a small number of simple patterns from examples or specific takeaways from successful games. Most documentation about the design and development process of human computation games comes from the research publications written by the task providers (i.e., researchers) interested in solving a particular human computation task. As previously noted above, researchers typically prioritize only the task and develop games that function just long enough to attain some semblance of a solution. Examples of failed efforts or games in this space are virtually nonexistent due to an unwillingness to publish or document negative results. The most that one can often take away from many other publications is that a game was the novel interface for solving a human computation task and that some (often arbitrarily-selected) combination of game-like elements were implemented to enable the completion of the task. Rarely do these efforts focus on deeply understanding *what* combinations of game elements contributed to that success, how these elements may have impacted both the quality of results and the experiences of their players, and optionally, how this might generalize to the design and development of future HCGs.

This is not unexpected. Task providers and novice developers may not always have the necessary backgrounds or experience in game design and development. A task provider may already find it difficult to justify the risk of an expensive game development process, particularly one with limited design resources that cannot guarantee that the final game may even be successful at solving their desired human computation task. As a result, most HCGs to date are built around specific kinds of templates or mechanics (and often only when tasks

are similar enough to those for which such design knowledge is available). This conservative approach to HCG development leaves the space of HCGs woefully unexplored. How what little HCG design knowledge exists might transfer to new tasks, new generations of game interfaces (i.e., both hardware and software platforms), and the ever-changing player appetites for certain kinds of games is an equally unexplored problem.

To facilitate broader adoption of games as an interface for human computation and to enable greater ease of their development, human computation game design needs the tools and frameworks to study and communicate about these games in a consistent manner. HCG developers need to understand precisely what game elements—mechanics, aesthetics, narratives, and more—make certain HCGs successful, that is *both* effective at engaging players *and* solving tasks.

Contemporary, entertainment-oriented game design has been built up through deliberate and dedicated efforts to consolidate design knowledge into formal representations and vocabularies. I argue that the same must be done for HCGs, and in a way that incorporates *human computation specific* concepts such as *tasks* and *solution verification*, while still considering general game design concepts such as *player audience* and *game mechanics*. A consistent, common language and representations for HCGs would then allow developers to discuss and explore the space of possible HCG designs. Not only might these constructs help to ensure that HCGs are as engaging and effective as possible, but they might also enable current and future game developers to explore and diversify the space of HCGs.

In this chapter, I describe the *framework* I developed for defining and understanding the *mechanics* of human computation games. This framework is motivated by the lack of a formal language or any generalized design tools for defining HCGs. This framework aims to break down HCGs into a common vocabulary and structure that allows for visualization, comparison, and exploration of the space of HCG game mechanics. Alongside this framework, I advocate for a *methodology* of building up HCG design knowledge, which uses small-scale, controlled design experiments on tasks with known solutions to under-

stand how variations of game elements may affect both the *player experience* and the *task completion*. I utilize both this mechanics framework and the methodology to frame the contributions of this dissertation consistently across its subsequent chapters.

This chapter consists of four parts:

1. A brief background on the existing game design literature for human computation games.

2. The mechanics framework for human computation games, utilizing three successful HCGs as illustrative examples.

3. The experimental methodology meant to accompany this framework, proposed as a means to develop design knowledge for HCGs.

4. A case study exploring image-labeling HCGs using this framework and methodology to compare and contrast the mechanics of these games.

As a reference, the peer-reviewed version of this work was published as a short paper accompanied with a poster at the Foundations of Digital Games Conference in 2017 [37]. The extended version of this work can be found on arXiv [38].

3.2 Background

3.2.1 Contemporary Game Design and Development

In the field of game design, a myriad of design resources exist to help guide the design process. These range from classic game design tomes such as Salem and Zimmerman's rules of play [39], Fullerton's playcentric approach to game development [40], and Schell's lenses of game design [41]. These resources are often grounded in past game development experience, providing design wisdom through anecdotal examples and examinations of seminal games. Other approaches to design focus on formalizing and taxonomizing patterns of game design. These include Björk and Holopainen's game design patterns [42]

and Falstein's rules for game design [43]. Additionally, online resources, such the aggregation of design articles on the website Gamasutra [44] or the archives of professional design talks hosted on the GDC Vault [45], also provide useful documentation and insight into best game design practices in an often-opaque software engineering industry.

Overall, these design resources, anecdotes, and heuristics are extremely useful. However they have seen little empirical evaluation, except for rare instances [46, 47]. Moreover, all of these frameworks and guidelines are generally intended for commercial or entertainment-oriented games. It is unclear how this knowledge would transfer to human computation games, or other games that have an alternative purpose beyond providing entertainment.

3.2.2 Game Design and Development for Human Computation Games

When it comes to games intended primarily for entertainment, there are a wide variety of frameworks and tools available for game designers and developers. However, there are few such affordances targeted specifically at human computation games. Existing serious games literature [48] often focuses primarily on games intended to convey information or messaging beyond or orthogonal to the entertainment provided by the game. Given that human computation games are complicated by the addition of a secondary goal—solving the human computation task—principles from serious games may be applicable. However these resources often omit human computation games entirely [49] despite their similarities to other serious games domains such as education.

Instead, human computation game design has been guided by examples of successful games, rather than systematic study of HCG elements. Among the most utilized examples are von Ahn and Dabbish's templates [50] in the context of classification and labeling tasks. These templates define three game structures: *output-agreement*, *inversion*, and *input-agreement*, which were based on the authors' experiences developing games such as the *ESP Game* [2], *Peekaboom*[16], and *TagATune*[3] respectively. Of these three tem-

plates, *output-agreement* is perhaps the most well-known, given the seminal nature of the *ESP Game* [2]. Similarly, in the domain of scientific discovery games, the work of Cooper *et al*. [9] emerges as one of the most utilized resources due to a thorough discussion on the design and development process of the successful, popular protein-folding game, *Foldit*. Cooper *et al*. describe the consideration of two design goals: correctly solving the protein-folding task, and making interaction "intuitive and fun" for the players. Additionally, their work details their iterative development process, and provides both an evaluation of the game's usability as well as insights into the *Foldit* player audience.

Unfortunately, while it is clear that these combinations of design choices do work (as evidenced by the relative success of these respective games), it is unclear which specific gameplay elements are responsible, or how specific game elements influence the player experience and the completion of the task at a more atomic level. This makes it difficult to appropriately generalize these design anecdotes or consider new alternatives when it comes to new or different tasks, not to mention new or different player audiences.

Beyond using anecdotally-developed HCG design patterns, several efforts have explored adapting existing design knowledge, patterns, and theories to the development of HCGs. For example, Carranza and Krause [51] explore the application of the Mechanics-Dynamics-Aesthetics (MDA) [52] framework to the development of the game *OnToGalaxy*. A more recent example is the work of Miller *et al*., who use Self-Determination Theory (SDT) and Cognitive Load Theory (CLT) as a design lens to successfully improve in-game elements in *Foldit* [53]. These projects demonstrate how more general game design knowledge or motivational theories might translate and apply to HCGs, but such work is rare and otherwise unexplored.

Additionally, there remain unanswered questions about human computation design in general. One such question is that of whether or not certain gameplay elements map better to certain tasks than others (i.e., are some kinds of gameplay elements better for certain human computation tasks). The previously-cited HCG taxonomies, especially that of Pe-Than

et al. [14, 15], highlight the similarities between the kinds of human computation tasks and corresponding gameplay elements. These aggregations might suggest that certain kinds of game elements are more effective for solving of certain kinds of human computation tasks. Alternatively these similarities may simply be the result of existing games copying from a limited set of templates and prior exemplars, rather than a thorough examination of the human computation design space.

Discussion on how to best map gameplay elements to human computation tasks focuses primarily on game *mechanics*. Jamieson et al. [54] introduce the term "isomorphs" to describe the potential for games with similar underlying problems or tasks to be mapped to different surface elements (e.g., game interfaces or genres). Meanwhile, Tuite [55] uses the term "orthogonal mechanics" to describe gameplay mechanics that detract or divert the player's attention from solving the task, and advises against the inclusion of these elements. However, others argue that the use of such "orthogonal mechanics" might be used to attract alternative audiences or improve player engagement. Krause et al. [22] suggest that the use of mechanics unrelated to the underlying task, but familiar in entertainment-oriented games, may attract skilled players otherwise uninterested in HCGs. Ultimately, it is unclear just if and to what extent the relationship between gameplay elements and the underlying human computation task should be for a given HCG. Answering this question necessitates establishing how certain gameplay elements affect both task completion and the player experience through actual, empirical evaluation.

Finally, when it comes to the development process, reusable tools for developing and publishing HCGs remain limited. While game engines and development tools are more readily available and accessible than ever, virtually no attempts have been made to develop dedicated game engines or game distribution platforms specifically for HCGs (e.g., there is no equivalent to Amazon's Mechanical Turk or Crowdflower for distributing HCGs). Typically, HCGs are developed on a per-need basis using free or easily-available game development tools and development environments. Exceptions include the work of Cam-

bria *et al*. [56], who describe the development of a game engine for building commonsense knowledge games. Most recently, the Massively Multiplayer Online Science (MMOS) platform [57] was developed to help support the integration of HCGs into larger-scale games; this platform currently supports ongoing initiatives such as the aforementioned *Project Discovery* and *Borderlands Science*.

3.2.3 Experimental Evaluation of Games

While game design knowledge provides initial good practices and intuition about how to build games, empirical or systematic evaluation of game elements and designs is required to verify how successful these might generalize or how effective these elements and designs may be for a given scenario. Prior research in games and human computation interaction has shown that both qualitative and quantitative research methods can be used to study game design. Broad approaches include methods from game usability [58], game analytics [59], and visual analysis [60].

In particular, between-subjects studies (or "A/B testing") have been commonly applied to test variations of design elements in games. Controlled studies have proved successful for understanding and analyzing game design elements such as difficulty [61, 62], controls [63, 64], and tutorials [65, 66, 67, 68]. Many of these studies (i.e., those of Andersen *et al*. [65, 66] and Lomas *et al*. [61]) have been conducted on educational games. Like HCGs, educational games must grapple with dual design goals; in educational this tension occurs between the need to meet learning outcomes and drive knowledge retention versus player (i.e., student) engagement with the game. It is unknown whether educational game design knowledge might transfer to HCGs as solving a human computation task is a very design goal (and activity) than ensuring long-term learning gains. However, the difficulty of balancing two design goals is highlighted in many of these educational games studies. For example, Andersen *et al*. [65] demonstrate how optional gameplay elements are shown to harm player retention (and thus long term learning objectives) if they sufficiently distract

from the primary elements (mechanics) of the game.

In the domain of general human computation, formal design studies have been used to successfully measure the efficacy of structuring and distributing tasks in certain ways [1]. However, when it comes to HCGs, it is rare to see games elements evaluated for efficacy beyond that of the task results. There are a few instances in which new HCGs are compared to existing, older games (e.g., the *ESP Game*) to show improved engagement metrics [69, 51], but these lack detailed analysis and insight into what particular elements of design were responsible for these improvements or how they also affected task completion. Even rarer are instances of comparing games to non-gamified human computation systems; Goh *et al*.'s investigations of image labeling applications and HCGs [70] are a notable exception. It is only recently that researchers and HCG developers (i.e., still typically the task providers) have begun to isolate and thoroughly test individual game elements for effectiveness at task completion and player experience. The work of this dissertation is among them.

3.3 A Formal Representation of Human Computation Game Mechanics

I propose a formal representation for human computation game mechanics. The functions of this representation are threefold:

1. To provide a common vocabulary and visual organization of HCG elements.

2. To enable formal comparisons of existing HCGs to better understand the space of HCG designs.

3. To facilitate the creation of controlled experiments of HCG elements in order to build further, generalizable knowledge of HCG designs.

This representation focuses on a specific kind of game element: human computation game *mechanics*—the functional rules that define player interaction with the game. While this excludes other elements of HCGs, such as aesthetic, narrative, or contextual compo-

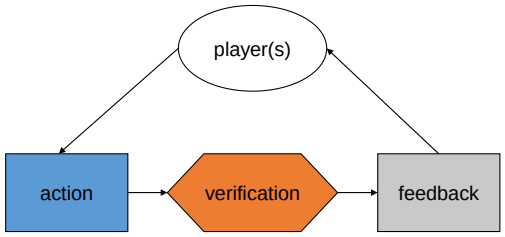

Figure 3.1: Breakdown of HCG mechanics. Players provide inputs to take actions (shown in blue), which are verified (shown in orange), and receive feedback (shown in gray) from the game. Solid lines represent transitions through the gameplay loop.

nents, this framework and its accompanying methodology may be amenable to such additions (though such work is beyond the focus of this dissertation).

The representation categorizes the mechanics of human computation games into three types: *action* mechanics, *verification* mechanics, and *feedback* mechanics. This breakdown, shown in Figure 3.1, reflects the core gameplay loop of most HCGs: a typical instance of play begins with players taking in-game *actions* relevant to the process for completing the task. These actions are typically followed by some kind of task-specific *verification* or validation of that input, which in turn is presented back to the player as some kind of (typically in-game) *feedback*.

I now define and describe these three types of mechanics in greater detail, illustrated with three successful human computation games: the original *ESP Game* [2], *Foldit* [4], and *PhotoCity* [6]. These three HCGs were chosen to match the three broad categories of HCGs previously outlined in Chapter 2. An illustrative breakdown of these images can be seen in Figure 3.2.

3.3.1 Game Mechanics

What are game *mechanics*? Game *mechanics* are functions invoked by agents (both human and artificial) that enable interaction with the state of a game. This particular definition comes from a survey by Sicart [71] which explores this exact question and proposes the above definition. As Sicart points out, the definition of game *mechanics* is often fuzzy and imprecise, due in part to conflation between the abstract description of the interaction—the

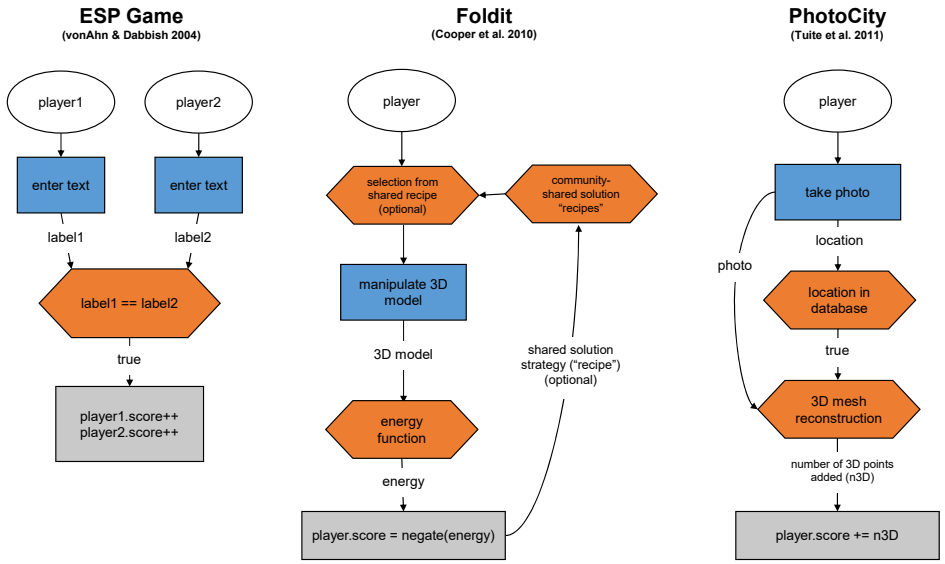

Figure 3.2: Examples of HCGs [2, 4, 6] subdivided into *action*, *verification*, and *feedback* mechanics. Arrows from feedback to players have been omitted for clarity.

rules—(e.g., "jumping" in a platformer game) and the underlying software implementation of the interaction—the systems—(e.g., a game-specific physics system and input system which jointly enable a player to "jump").

Sicart, as well as Adams and Dormans [72] (who likewise survey the definitions of game mechanics) focus primarily on the former, formalizing game mechanics as *rules* operating on game state which are described by *verbs*. This use of *rules* and *verbs* as terminology is common across game design literature. For example, in Anthropy and Clark's game design vocabulary [8], the term *mechanics* is eschewed altogether, but in favor of using both the terms *rules*—defined as relationships between game elements—and *verbs*—specific kinds of rules that enable player interaction.

By contrast, Hunicke *et al*. [52] define *mechanics* as the data representation and algorithmic components of a game, whereas *dynamics* are defined by runtime behavior of the mechanics due to player interaction.

In the framework I propose, I consider a single *mechanic* to map to a rule or verb which describes a function (action) taken as part of the human computation game loop.

The further classification of *action*, *verification*, and *feedback* mechanics is based on their function and contribution to the human computation process.

Note that a *mechanic* may consist of multiple rules or verbs. For example "jumping" may be specified as a sequence of rules wherein the user *pressing* a button, the game then *applying* an impulse to an avatar's game state (i.e., position), and then optionally, the game *applying* a gravitational force downwards. The choice and composition of these verbs may vary on a per-game basis. However, game designers still think of "'jumping" as a singular mechanic where the choice of exact rules composing it is considered a detail of the particular game implementation.

Additionally, most definitions of game mechanics refer to or focus on what are known as *core mechanics*—the rules which have the most impact on play. These are the mechanics which considered the most representative of the genre or type of game they belong to. For the subsequent examples and discussion, I will primarily on what may be considered *core mechanics*.

3.3.2 Action-Verification-Feedback

Action Mechanics

Action mechanics are the functions that enable a player to complete the human computation task. This task completion is accomplished through the in-game *actions* a player takes through what are commonly referred to as the "gameplay elements" of the game. These mechanics often require players to utilize skills necessary for solving the task as part of play. Such mechanics may be as straightforward as entering a word into a text field or as complicated as piloting a space ship through a virtual environment. These mechanics may vary wildly based on the nature of the task and may be the combination of many in-game systems intermixing in order to enable a single player action.

For example, "jumping" to avoid an obstacle (e.g., perhaps a metaphorical player action "avoiding" an incorrect label in classification task) is a mechanic that relies on interactions

between multiple possible systems: an input system responsible for processing the player input, a physics system that translates that input into the physical simulation of the game to determine whether or not the player's in-game avatar collided with the object representing the classification label, the game state management (system) that tracks which labels were avoided or not, a health system that decrements the player's health, etc. Suffice to say, there may be a composition of multiple systems behind a single "mechanic," however "jumping" does not prescribe a specific set or combination of systems. Instead, the exact number and kinds of systems that enable "jumping" as an action are an implementation detail of the specific HCG. Therefore, in the following examples, a single *action* mechanic reflects the overall single (or type of) player interaction, omitting the implementation of its underlying system from the definition.

Examples of Action Mechanics

In the *ESP Game*, players are shown an image in each round of the game. Player actions are limited to entering possible labels for that image via text entry into a text field. For a given image (in each round of the game), each player's provided labels are tracked for future comparison.

In *Foldit*, players are shown a complex three-dimensional model representing the structure of a protein. Player actions include various operations to handle and move (i.e., rotate and translate) different parts of this model in order to shape it into different configurations. These manipulations may be applied in a sequence or individually. (Note that in Figure 3.2, I denote these manipulations as a single node for clarity, but this node represents either a single or repeated application of these actions.) Updated versions of the game include additional actions, such as applying a specific sequence of moves (known as "recipes") based on the past actions and efforts of other players.

In *PhotoCity*, players are given a location on their mobile device. Player actions are a combination of physical activity (e.g., physically relocating themselves to that real-life

location) and in-game actions (e.g., taking pictures of buildings at that location using the camera on their phones). These pictures are then uploaded to a database and later used to construct a three-dimensional representation of the buildings in that location.

Discussion of Action Mechanics

Generally, the *action* mechanics in human computation games have been most closely aligned with the direct process of solving the human computation task. The three examples described above and also shown in Figure 3.2 describe *action* mechanics that directly facilitate with the process of solving a human computation task (as opposed to abstracting the problem solving process with an unrelated player action). This suggests that the mechanics of these games may have been designed first and foremost with the task in mind, as opposed to adapting the mechanics of existing, entertainment-oriented digital games.

There exist actual explorations of adapting mechanics from mainstream, entertainment-oriented digital games for human computation games. However, these instances are few. One example concerns the game *OnToGalaxy* [22], which addresses the ontology-construction task tackled by previous HCGs such as *Ontogame* [21]. In *OnToGalaxy*, the task completion process is mapped onto to a space shooter game akin to the classic arcade game *Asteroids* and the *action* mechanics consist of piloting a spaceship and shooting at the correct ontological relations. Unfortunately, it is difficult to draw any conclusions about the efficacy of this approach. *OnToGalaxy* was compared against the original *ESP Game*, and the wide variations between the game mechanics as well as the human computation tasks make direct comparison and generalization difficult. Another example (which I discuss in Chapter 4), is the game *Gwario*, an HCG that turns a classification task from prior work [73] into a platformer game resembling the classic action game *Super Mario Bros.* Accompanying the development of this game was a survey sent to current HCG developers and researchers, asking directly if mechanics from entertainment-oriented digital games should be adapted into HCGs. In a divergence from prior HCG design theories [54, 55], these experts favored

the idea of adapting mainstream game mechanics to HCGs, though not without cautioning that such mechanics ought to be selected very carefully to avoid compromising the task results.

Finally, a notable, inverse example comes in the form of *Project Discovery* [24]: an HCG which is incorporated directly into an entertainment-oriented game (*EVE Online*) as a "minigame" (i.e., a game within a game) in contrast to adopting mechanics directly from an existing game. The *action* mechanics within *Project Discovery* game consist of annotating images of proteins by highlighting portions of a given image with different types (class) of protein patterns using the mouse. However, this experience, while set in the diegetic narrative of *EVE Online*, scarcely resembles its encompassing game, which is a massively-multiplayer galactic simulation.

The longstanding question of whether or not to adapt mechanics from popular, entertainment-oriented games is driven by one of the major criticisms of human computation games. Tuite details this issue in her critique of HCGs [55]. One of her primary observations is that HCGs may often be perceived as shallow compared with entertainment-oriented digital games and therefore may fail to attract potential or sufficient players. The adaptation of familiar or successful mechanics from existing digital games might serve to address this issue.

Action mechanics are the primary candidate mechanics for this kind of adaptation or inspiration from entertainment-oriented games (compared with shortly-to-be-discussed *verification* mechanics, which are admittedly not necessary for games focusing on entertainment that do not have task results to validate). One reason is that these mechanics are the mechanics with which the player interacts with first in an HCG. Given their forward placement, these mechanics are the most likely to maintain player engagement and retain players throughout the task-solving process. Proper adoption and implementation of such mechanics could have a number of benefits. An HCG which successfully adopts *action* mechanics from a pre-existing genre or popular game may find a wider audience, particularly

among players who prefer similar games and might not have tried an HCG otherwise. Additionally, players familiar with such mechanics may also find it easier to learn and adjust to gameplay, compared with players who are less familiar with HCGs. Conversely, there are also potential pitfalls. *Action* mechanics may outright distract or impede the player if poorly implemented, raising again the question of whether or not *action* mechanics should be non-orthogonal to the task or not. Additionally, just as the *action* mechanics might attract players to the game initially, familiar mechanics may also raise expectations about that game that may not be met, especially given the constraints of solving the task alongside expected gameplay. For example, an HCG which adopts *action* mechanics from a successful game may negatively impact players' experiences (and thus, their continued participation) if the design quality of the HCG does match that of its original inspiration.

As previously stated, there is no single or correct answer on the question of incorporating game mechanics from entertainment-oriented digital games into human computation games or conversely, adapting HCGs into such entertainment-oriented games. I raise this question here in part to highlight the lack of exploration and research in this area. This absence necessitates the creation of more HCGs that attempt to incorporate successful elements of entertainment-oriented games, as well as the deliberate exploration and documentation of these attempts.

Verification Mechanics

Verification mechanics are the functions which take an input of player actions to compute an output of task-relevant outcomes. These mechanics facilitate the assessment and measurement of task-relevant metrics such as the quality, volume, diversity, and rate at which results to the completed human computation task are acquired.

Verification mechanics may exist as part of gameplay, responsible for validating these task results in real-time. Alternatively, these mechanics may function as systems or applications external game interface entirely, validating results sent from the game to a delayed

and often external mechanism (e.g., a network server). I refer to these as *online* and *offline* mechanics respectively, referring to the response time in which and the location where this verification occurs. I also note that these two categories are not mutually exclusive as some HCGs utilize a combination of both.

Examples of Verification Mechanics

For many human computation tasks, an aggregated consensus on player input often serves as verification. In the *ESP Game* (and many of the games inspired by its structure per the *output-agreement* template), task results are verified using an *online* agreement check that filters correct answers from incorrect answers by relying on agreement (i.e., receiving identical or similar answers) between players. Arguably, the *ESP Game* also uses an *offline* check as well, by further aggregating and filtering results saved following game rounds. While initially, this *offline* check did not impact the game, subsequent versions of the *ESP Game* would declare the most commonly-suggested results to be "taboo words" that could not be entered, in order to promote data diversity once sufficient consensus on particular labels was reached.

In *Foldit*, task results are verified using an *online* check: a task-based evaluation function that takes a given protein structure and computes its energy configuration. *Foldit*'s protein energy configuration function thus determines the quality of player solutions in real-time. Additionally, the game makes use social and sharing mechanics, allowing players to share solution procedures (called "recipes") through its community interfaces. While the application of recipes may be considered part of the *action* mechanics of the game (in which players utilize existing recipes uploaded by other players as starting points for solving tasks), repeated use and ratings of a recipe's utility may act as an informal validation of the encoded solving strategy.

In *Photocity*, task results are verified as part of an *offline* process. After players take photos at the desired location, the uploaded images are processed on an external server;

player feedback is determined by the resulting alterations to a constructed three-dimensional mesh of the desired location.

Discussion of Verification Mechanics

Verification requirements for a task greatly impact game design decisions such as the number of players required to play the game—singleplayer (asynchronous play) versus multiplayer (synchronous play)—and how these players might interact with each other.

The complexity of verifying the task results is often a determining factor how verification is handled. Task results that can be evaluated using an objective function or can be compared against existing data can utilize either *online* or *offline* verification processes. In *Foldit*, protein configurations can be evaluated quickly using an objective function, enabling *online* verification. Contrastingly, *PhotoCity* requires comparing an image against many existing images and computing the resulting three-dimensional mesh, a process handled *offline* due to computational requirements. However, both games are similar in that neither requires the consensus of multiple players at the same time, enabling these games to be asynchronous, singleplayer experiences.

By contrast, the *ESP Game* relies on consensus from multiple players even if the verification step is a string comparison between player inputs. This makes the *ESP Game* a synchronous, multiplayer experience, in which players are networked across the internet simultaneously. Multiplayer experiences may also rely on simulated players to compensate for instances when multiple players are required, but not necessarily available concurrently. Such fallback mechanisms (demonstrated in the *ESP Game*, for example) may also allow old solutions to be re-verified using fewer players.

Unfortunately, much like *action* mechanics, few design experiments have tested alternative *verification* mechanics in human computation games, let alone explored different designs. One notable exception is the game *KissKissBan* [69], which modified the original version of the *ESP Game*by adding a third player to "ban" commonly-used words in order

to promote data diversity. Compared with the *ESP Game's* eventual use of taboo words, *KissKissBan* enables alternative *verification* mechanics that players found to be engaging. I discuss this further in Section 3.5.

In most human computation games, particularly those which rely on synchronous consensus, players are commonly forbidden from direct communication with each other. Multiplayer *verification* mechanics are typically implemented using an anonymous pairing with no means for a player to determine the identity of the other players they interact with. This particular design paradigm was first implemented in the *ESP Game* as a way to prevent minimize collusion between players, as such behavior might induce players into providing deliberately incorrect answers while still succeeding at the game. For this reasons and others (e.g., the added complexity of implementing multiplayer mechanics), many games do not allow players to communicate or choose to remain entirely singleplayer experiences. Only a few games (e.g., *Foldit*) allow players to interact through communication channels such as game forums or community interfaces. I explore this specific design paradigm (i.e., banning direct communication to avoid avoid collusion) in Chapter 4.

Overall, *verification* mechanics are sadly, under-explored in human computation games. However, there are still many unanswered questions about how such mechanics work. How do different mechanics for verification influence task results (e.g., how do the sets of diverse labels compare between the *ESP Game's* taboo word implementation and *KissKissBan's* antagonistic third player)? Is *online* or *offline* verification preferable for some tasks over others, and how might these affect both the task results and player feedback?

Feedback Mechanics

Feedback mechanics are functions take the results of player actions—partially or fully-verified task completion–and provide players with information or digital artifacts. These mechanics commonly encompass gameplay elements such as rewards and scoring. Moreover, the feedback provided by these mechanics can possibly be mapped to evaluation

metrics for the underlying task, enabling both researchers and designers to assess player performance at both the completion of the task and progression through the in-game experience.

Examples of Feedback Mechanics

For all of the games shown in Figure 3.2, players receive feedback as a change (typically an increase, but occasionally a decrease for some HCGs) to a tracked score. In the *ESP Game*, both players receive points when they agree on a label for a given image. By contrast, *Foldit* rewards players with points based how well they can minimize the value of the energy function derived from the protein's structural configuration. Since a *lower* energy value is desirable, the visible score presented to the player is actually the negation of the energy function value, which then lets players work towards a *higher* visible score. Finally, *PhotoCity* rewards players for the number of points their provided photos add to the reconstructed three-dimensional mesh. The scale of points differs between games (i.e., one point in the *ESP Game* is not equivalent to one point in *Foldit*) as a game-specific value based on the results of the completed task. These games are similar in that the feedback "currency" is nominal—points which contribute to a numerical, increasing score—but vary in what players are rewarded for.

Discussion of Feedback Mechanics

Aspects of *feedback* mechanics are some of the most well-explored and understood design elements in human computation games. These mechanics are synonymous with in-game rewards, which some consider comparable to the monetary compensation in other crowdsourcing systems.

One design question to be asked is what should feedback be provided for: player actions in the game, player performance at the task (as evaluated by *verification* mechanics), or a combination of both? Typically, positive *feedback* in HCGs is structured to encourage

participation in the crowdsourcing process. Players may receive rewards for first completing a task. Then, if verification of the task is immediately available, they may receive additional rewards for completing it correctly or sufficiently. (For *offline verification* mechanics, in-game feedback may be delayed depending on the length of time taken to verify results.)

Typically, this kind of positive feedback is given in response to *collaborative* player behavior, which is reasonable given that human computation and crowdsourcing are aggregate, collaborative processes that often rely on consensus. However, in entertainment-oriented games, reward systems such as point-based scoring and player leaderboards are often included as avenues to afford and encourage *competitive* player behavior. In the context of HCGs, the inclusion of competitively-motivated rewards is a concern, since it may be desirable to discourage competition in order to keep players focused on completing the task correctly (i.e., players may optimize their in-game actions for the goal of outperforming their peers, rather than towards completing the task successfully). Conveniently, HCG research has actually explored the question of using *collaboration* versus *competition* as motivator in HCGs. Both the aforementioned HCG *KissKissBan* [69] and a study by Goh *et al*. [70] examined the question in the context of the *ESP Game*. Likewise, I explored this question using the games *Cabbage Quest* [73] and *Gwario* [74]. In broad summary, this research all shows that there are potential tradeoffs between emphasizing collaboration versus competition. However, for the moment, I withhold further discussion on the question here; it will be revisited in extensive detail in Chapter 4.

In addition to examining what players are rewards *for*, one must also consider what players are rewarded *with*. Multiple types of *feedback*—the many kinds of in-game rewards—exist. As players have different motivations that may cause them to respond more positively or negatively to different types of rewards, the types of rewards an HCG provides may be the difference between attracting many versus few players to the game. But what kind of players, and by extension their motivations, should be considered? Crowdsourcing research

has found that players who are intrinsically-motivated to solve tasks are often disengaged by monetary compensation [75] and that curiosity can be a strong incentive for crowdsourced work [76]. A common paradigm in HCG design is to consider standard in-game rewards (i.e., point-based systems) as a replacement for monetary compensation, but the aforementioned crowdsourcing research suggests that this substitution may not appeal to all kinds of players. Specifically, players who are dedicated to the task might disengage with the game if it emphasizes a certain kind of reward, particularly if it is one that they are uninterested in. The existence of intrinsically-motivated players is not in question; in their analysis of *Foldit*, Cooper *et al*. [9] identify a subset of such intrinsically-motivated players who are driven primarily by their participation in the scientific discovery process.

So what are the alternatives to the standard point-based systems and player leaderboards for players who may not find typical *feedback* systems compelling? In the context of HCGs, research has explored a variety of different reward systems within the context of the game. Goh *et al*. [77] report on utilizing points, badges, and non-gamified statistics in a location-based content sharing HCG. Similarly, I investigate leaderboards, avatar customization, unlockable narrative, and non-gamified statistics in the game *Café Flour Sack* [78]. Additionally, Gaston and Cooper [79] explore three-star reward systems in the context of *Foldit*. In broad summary, this research all shows that player audiences may be affected by how and what kinds of rewards are available, and that players may behave differently under different conditions. As before, however, I abstain from further discussion on the topic; it will be revisited in extensive detail in Chapter 5.

Once again, it is worth noting that *Project Discovery* is an unusual case study in the context of *feedback* mechanics among human computation games. The playable experience directly integrated in *EVE Online* rewards players with currency that can be spent within the greater game world, thus providing a potential motivation for the broader *EVE Online* player audience to participate. Direct integration of HCGs into existing (entertainment-oriented) games remains otherwise unexplored, but the potential to leverage large existing

player audiences into playing HCGs still remains (provided there are opportunities to enable it).

3.4 An Experimental Methodology for Human Computation Game Design

The formal representation described above provides a breakdown of the different kinds of mechanics in human computation games and their functionality in the human computation process. Using this division, an HCG designer or developer might be able to express or consider where exactly in the game loop they might target a specific mechanic and broadly, how it might affect the completion of the task through play. However, this representation alone is not enough. How might one explore the space of HCG designs, particularly in a way that allows the buildup of generalizable design knowledge?

To complement this mechanics representation, I propose using a methodology of controlled A/B design experiments that explore the space of human computation designs. This procedure can be summarized as the use of between-subjects (or alternatively, within-subjects) experiments on versions of HCGs with different mechanical variations while measuring their effects on both the completion of the task and the overall player experience.

These design experiments should (1) implement a task ideally with a known solution, while (2) focusing on a single element of an HCG's design.

First, testing with a known solution permits objective evaluation of task-related metrics without the conflation of simultaneously solving a (potentially novel) human computation task. Such known solutions may be the result of pre-solved human computation problems (e.g., image labeling, which will be further discussed in Section 3.5) or simple tasks (e.g., problems requiring commonsense human knowledge or which are easily verified). Data collection tasks which prioritize a large quantity of (but not necessarily quality) data, such as those for building training datasets for machine learning algorithms, are also amenable because they do not require evaluation for quality or diversity.

Second, focusing only on one particular element of an HCG's design allows us to understand exactly what kind of impact that single element may have on both the players and the task with minimal interaction effects The mechanics representation can be used to assist in understanding where and how the introduction of a varied element (which may affect one or more of the *action*, *verification*, and *feedback* mechanics) may affect the HCG game loop.

Most importantly, these experiments should simultaneously evaluate how design decisions meet the needs of *both* players and the task. Optimizing only for the player may result in a game with mechanics that do not effectively solve the human computation task if even if the game is considered engaging. Optimizing only for the task may result in a game that players do not find engaging (and thus will not play) even if the game effectively solves the task. I refer to these two axes of metrics as the *player experience* and the *task completion*.

Player experience encompasses both quantitative and qualitative metrics such as:

1. **Engagement**: how players interact with the game or rate their overall experience with it
2. **Retention**: how likely players are to continue playing or return to the game after a single play session
3. Other subjective metrics related to how players interact and perceive the game (e.g., preferences, unstructured self-reported feedback)

Task completion refers to (mostly quantitative) task-related metrics such as:

1. **Quality**: correctness or accuracy or task results
2. **Volume**: amount of completed tasks
3. **Diversity**: variation, breadth, or coverage in the domain of task results
4. **Rate of Acquisition**: speed or delivery measurement of completed task results

The exact metrics to measure and test for often depend on a variety of factors. For the *player experience*, the type of game (and its game elements) may determine or naturally emphasize what kinds of metrics are important. For example, an HCG with daily challenges may be more concerned with retention than an HCG intended to be played once (e.g., one with a single series of challenges or specific tasks to be solved). Likewise, the platform of distribution may enable or dissuade the collection of certain kinds of player-provided feedback (e.g., asking players of a mobile game to provide free form, unstructured feedback is impeded by the limitations of a touchscreen keyboard). The nature of the underlying human computation task may also affect which *player experience* metrics should be prioritized. For example, in HCGs with tasks whose solutions necessitate an extended tutorial or training period for players (e.g., the complex optimizations of *Foldit*), player *retention* may be considerably more important to measure than for HCGs where maintaining a skilled player base is not necessary or a priority (e.g., the commonsense-knowledge-powered image labeling process of the *ESP Game*).

Likewise, *task completion* metrics similarly depend on the game elements, but more so on the nature of the human computation task. For example, an HCG designed to collect a dataset for a machine-learning algorithm may prioritize *volume* of task results whereas an HCG designed to solve a specific scientific optimization problem may be more concerned with the *quality* of the task results. (This should not suggest that an HCG should not consider all of these metrics, but optimizing for all of them simultaneously may be unfeasible.) Additionally, these requirements may change over time. Task providers may find that their initial *task completion* results may not be sufficient or require additional refinement. Nowhere is this better illustrated than in the evolution of the *ESP Game* and its variants (a discussion which I will examine shortly in Section 3.5).

Altogether, this methodology can be broadly summed up as a series of steps.

1. Select a human computation task, ideally one with easily-verifiable or known results.

2. Select a specific or single game element that can be varied. The variation is the

independent variable in the experiment. I should note that the mechanics representation above is designed to help isolate such game mechanics (or potential elements/conditions) to try and avoid interaction effects as much as possible.

3. Build an HCG designed to test the specified game element with room for variation.

4. Run a between-subjects experiment (AB test)/within-subjects experiment across at least two versions of the HCG.

5. Measure and evaluate the results of all conditions, focusing on both *player experience* and *task completion* metrics.

The methodology which I enumerate here is not novel in the domain of human computation games, as similar experimental testing approaches have been previously applied (prior to the work in this dissertation). In one such early instance, Goh *et al*. [70] compared a non-gamified control application for image labeling against two versions of the *ESP Game*, one using collaborative scoring mechanisms and one using competitive scoring mechanisms. Goh *et al*.compared completion results of the image labeling task, as well as various user interaction aspects, across all three conditions: the application and the two game variations. Their experiment and results are further described in the subsequent section. In another early instance, I conducted an experiment using the HCG *Cabbage Quest* [73] to test collaborative and competitive scoring mechanisms. The HCG utilized a task with a known solution—categorizing everyday objects by their potential purchasing locations—and controlled experiment compared two in-game scoring mechanisms: one collaborative and one competitive. Task results were compared to a gold standard answer set to evaluate *task completion* metrics while player actions and survey responses were logged to evaluate *player experience* metrics. Both Goh *et al*.'s study and my study follow this proposed methodology of taking a problem with a known solution or gold-standard answer set, testing game elements by treating a set of game mechanics as independent variables, and measuring aspects of both the *player experience* and *task completion*.

While both experiments benefited from having known or evaluated solutions, this is not a strong requirement. In some cases, a preexisting solution may not be available, especially if the task of choice is a novel human computation problem. In such an instance, it may be sufficient to evaluate first for *player experience*, then follow with late evaluation of *task completion* upon verification of initial task results.

3.5 A Case Study: Comparison and Evolution of Image Labeling Games

In this section, I provide a discussion of the evolution of game design for image labeling human computation games. This examination is intended to demonstrate how one might utilize the mechanics framework to visualize and discuss the common elements (and adaptations) of these HCGs. Because the games in this example also underwent some amount of empirical evaluation of their specific game elements, I also speak to how these games partially apply the experimental methodology I proposed and use this to discuss how the variations and evolution of certain game elements might have impacted aspects of both *task completion* and *player experience*.

As previously discussed in Chapter 2, image labeling—the task of annotating or classifying an image with labels or tags describing its visual contents—is one of the most iconic and well-studied tasks in human computation games. Figure 3.3 illustrates three HCGs designed to solve the image labeling problem, represented using the mechanics framework. From left to right are the original *ESP Game* [2], followed by *KissKissBan* [69] and Goh *et al.*'s *ESP Games* [70]. Figure 3.3 also colors the mechanical structure of the original *ESP Game* in gray, thus highlighting (in white) the additions and adaptations of the two subsequent iterations.

3.5.1 *KissKissBan*

Ho *et al.*'s *KissKissBan* was motivated by need to generate a wider, more diverse set of labels for given images. One issue that emerged following the deployment of the origi-

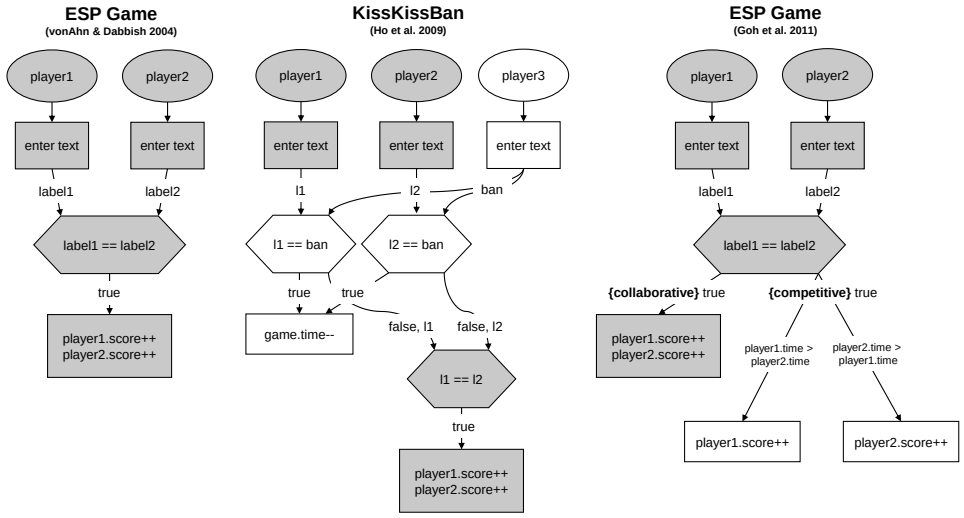

Figure 3.3: Mechanics breakdown of three image labeling HCGs [2, 69, 70]. On the left is the original *ESP Game*, followed by subsequent variations that modified elements of its original design. The mechanics of the original *ESP Game* are colored in gray; novel mechanical variations are colored in white.

nal *ESP Game* was that the initial label results for an image tended to converge to a correct, but limited set of labels, often shorter in length due to the time constraints of each round. While not shown in Figure 3.3, the original *ESP Game* eventually adopted the use of "taboo" words—a set of banned words based on repetitive solutions derived from the initial labelings—that players could not imput, forcing players to generate a more diverse range of labels. *KissKissBan* providers an alternative solution to the problem, specifically by introducing a third player. This third player, known as the "blocker" is tasked with entering words that he or she might suspect the first two players, known as the "couple," may attempt to input. If the blocker inputs a label before both members of the couple input the same label, the label is considered a "ban" and invokes a penalty on the couple. The inclusion of this third, adversarial player, therefore encourages the first and second players to avoid inputting more "obvious" words in favor of different words that the third player is less likely to immediately consider (and therefore ban).

The inclusion of the third player affects mechanics at all levels of the human computation game loop as shown in Figure 3.3. The *action* mechanics are similar to those of

the original *ESP Game*; all players input words as potential labels, including the new third player. However, *verification* mechanics change substantially. If both couple players agree on a label (forming a pair), the label is first considered as a potential solution. However, if the blocker has already suggested the label, this label is considered a banned label. Note that agreement between a label suggested by the blocker, and either or both of the couple players constitutes a potential solution to the task.

These results of verification are resolved through the game's adjusted *feedback* mechanics. The original *ESP Game* gave both players points for agreement and incremented a score upon each successful label pair entered. By contrast, the couple in *KissKissBan* needs only one such successful label pair to succeed at a round of the game. However, this label pair cannot be banned by the blocker. If the blocker has banned the label, then the couple loses five seconds from the round. Should the round time expire without a successful pairing, the blocker than automatically wins.

KissKissBan thus demonstrates one potential modification of the original *ESP Game*. Experimentally, Ho *et al*. did re-implement a version of the original *ESP Game* as a comparison point. However, the only metric considered in their evaluation was the diversity of the labels, specifically the number of distinct labels. Their brief results show that the original *ESP Game* (without the "taboo" words) produced nearly half as many distinct labels as *KissKissBan* (an average of 6.56 to 11.54). This result suggests that *KissKissBan*—and its mechanical variations—may be more effective at promoting label diversity compared with the original *ESP Game*, thus resulting in a possibly-compelling alternative. Unfortunately, very little else can be drawn from their conclusions regarding the *player experience*. While Ho *et al*. did conduct a gameplay survey on *KissKissBan*, it was not administered for their *ESP Game* reimplementation. Thus, while players (unsurprisingly) suggested that *KissKissBan* was fun, it is impossible to confirm how "fun" it was compared to the original *ESP Game*.

3.5.2 Goh *et al.*'s *ESP Games*

Goh *et al.*also demonstrate a modification of the original *ESP Game*: a competitive version of the *ESP Game* wherein the two players compete for the highest score, rather than attempting to maximize a joint score shared by both players. Traditionally, competitive elements in human computation games are downplayed and restricted to extradiegetic elements of the game, such as global leaderboards. This choice of design is driven by the concern that competitive elements might distract players from providing quality task results. However, competitive gameplay elements are considered to be some of the most engaging elements to (certain kinds of) players. Thus Goh *et al.*sought to answer the question of how competitive elements might function in the context of a traditionally-collaborative game.

The changing of the *ESP Game* from a collaborative game to a competitive game is shown in Figure 3.3. Of note is that both the collaborative and the competitive versions of the *ESP Game* remain very similar in that players take the same actions and verification of their results is identical (thus the *action* and *verification* mechanics remain unchanged). Instead, the *feedback* provided to the players is where the two versions diverge. This divergence (i.e., the split between the variations) is reflected as two branching nodes labeled with the variations (i.e., conditions of the study). In the collaborative version, both players are rewarded jointly when they agree on a label. In the competitive version, only the first player to suggest the label is rewarded when both players agree.

Goh *et al*. then conducted a study using these collaborative and competitive versions of the *ESP Game*, comparing them against a non-gamified control application. (Note: I examine the results of this particular study in greater detail in the following chapter, along with the discussion of other studies conducted on collaboration and competition in games, but summarize some relevant results here.) Importantly, Goh *et al*. examined not only how these three conditions affected the quality of tag labels (*task completion*), but also player perceptions (*player experience*). One relevant finding was that Goh *et al*. found no significant difference between the quality of labels (tags) between the collaborative and com-

petitive versions of the game. This finding might suggest that collaborative and competitive feedback mechanics might be interchangeable. Meanwhile, unstructured qualitative feedback (i.e., commentary from study players) suggested that players found the competitive version more compelling than the collaborative version. However, quantitative survey feedback on questions such as "challenge," "learnability," and "appeal" demonstrated no significant differences between the collaborative and competitive versions. Taken together, these results might suggest that the competitive version of the *ESP Game* might be a viable alternative to the collaborative version, as task quality was not compromised. Such a change might also be simple to implement, since the changes required only affect the *feedback* mechanics.

Summary

Understanding how certain mechanics are modified or extended on top of existing games can provide potential insight into how these might be applied to other human computation games. This discussion of image labeling games shows how one might take the changes utilized by *KissKissBan* or Goh *et al*.'s *ESP Games*, and apply them to games that are structurally similar to the *ESP Game*(or any other "input-agreement" inspired games for that matter). Furthermore, the mechanics framework provides a consistent way to talk and compare these games. For example, both *KissKissBan* and Goh *et al*.'s *ESP Games* add "competition" to the *ESP Game*, but these variations are very different. Furthermore, these variations address different metrics of the task results: *KissKissBan* is concerned with label diversity, whereas Goh *et al*.'s *ESP Games* are concerned with label quality (correctness).

Using an experimental methodology that considers both *player experience* and *task completion* can provide task providers with a better understanding of just how certain mechanics (or variations) are likely to affect the *player experience* and *task completion*. For example, Goh *et al*.'s competitive *ESP Game* demonstrates a change of *feedback* mechanics that may be interchangeable with the collaborative *feedback* mechanics (with respect to

quality task results and player perceptions of the game). Ideally, as more comprehensive design knowledge is built up, HCG developers might be able to treat these collections of mechanics as modular when applying them to new tasks and integrating them into new games—a transfer process which could be accomplished with confidence assisted by an understanding of the potential effects on the *player experience* and *task completion*.

3.6 Conclusions

In this chapter, I describe a framework for breaking down and illustrating the mechanics of human computation games. This formal representation describes the mechanics of HCGs based on their functionality in the human computation game loop: *action* mechanics which describe the task-relevant interactions through which the player may solve the human computation task, *verification* mechanics which enable online or offline (or both) validation of task results, and *feedback* mechanics which respond to task results by providing players (and task providers) with information and artifacts based on task results.

To complement this framework, I also describe a methodology of running design experiments that measure aspects of both *player experience* and *task completion*. These small-scale experiments focus on systematically isolating and testing specific game elements in order to understand the effect of variations on both players and task results.

To illustrate how one might use the framework and methodology to better understand the impact of design changes on human comptuation games, I present an examination of image labeling games based on the original *ESP Game* and its variants. Using the mechanics framework demonstrates how subsequent image labeling games such as *KissKissBan* and Goh *et al.*'s *ESP Games* explored mechanical variations on top of the original *ESP Game*. Using the experimental methodology as a lens demonstrates how their findings could have affected *player experience* and *task completion*.

Taken together, the combination of the mechanics framework and experimental methodology offers a way to illustrate and rigorously define how game elements (particularly game

mechanics) should be considered, composed, and evaluated in HCGs. I will utilize this framework throughout the entirety of this dissertation to identify and visualize particular game mechanics that I test and evaluate.

CHAPTER 4
REWARD FUNCTIONS — COLLABORATION, COMPETITION, AND CO-LOCATION IN HCGS

4.1 Introduction

As previously emphasized, one longstanding concern in human computation game design is ensuring that these games are entertaining enough to attract a sufficient audience of players to merit the overhead of solving a human computation task using a game as the completion interface. Developing a game is an expensive, time-consuming process; compared with designing a web form to distribute a task on an online crowdsourcing platform, game development incorporates not only technical (i.e., engineering) expertise, but also aesthetic, design, production, and even business expertise to produce a fully-functional artifact. In addition, while better development tools and more available platforms for game distribution (i.e., mobile devices, consoles, etc.) can facilitate easier HCG development and distribution (in particular for task providers who are not often professional game designers), these same benefits also have resulted in an explosion of more and more games—all of which compete for players' time and attention. So while early HCGs such as the *ESP Game* may have been able to benefit from novelty, modern HCGs can no longer afford to be experiences focused solely on solving the human computation task. Ensuring that these games are also compelling experiences to play and engage with now becomes a determining factor in whether or not the underlying human computation problem can be solved sufficiently.

As discussed in previous chapters, one longstanding question in human computation game design is whether or not elements of successful or iconic entertainment-oriented games can be adapted into HCGs. The motivation for this question is driven by the reality that human computation game developers are typically not professional or industry

game developers. Barring a collaboration with industry developers, a task provider with no game development experience might first opt to look at existing HCG literature for inspiration. Alternatively, if a task provider is actually concerned with ensuring that their game is entertaining enough to attract players, they may instead turn towards modern games and design resources for inspiration. But while game design is rapidly becoming a formalized field of study, most game design programs and resources focus primarily on entertainment-oriented games. Thus, it is unclear if simply taking game elements from popular games is enough—and more importantly, not detrimental—to the process of solving the human computation task. At worst, adding too many game elements optimized for a positive player experience or engagement may distract from the human computation task, yielding a game that fails to achieve any useful results.[1]

This problem is exacerbated by the fact that solving a human computation may be a short or mundane procedure. Therefore, game mechanics which facilitate task completion may risk ending up equally short or mundane, which is orthogonal to most player expectations of the game being engaging. Given that much of the foundation of HCG design comes out of academic research in which a game was developed to solve a novel human computation problem (wherein the game itself was argued to be the research contribution), I posit that task providers in such instances were more likely to avoid any game elements that could compromise the task completion. [2] Therefore, I reiterate that it is absolutely necessary to explore this problem of adapting game elements—specifically game mechanics—from successful entertainment-oriented games, particularly in scenarios where solving a novel human computation problem is not the priority.

I propose that one avenue of exploration is to examine what measures existing human computation games have taken to deliberately stymie "adversarial" player behavior—that

[1]For this reason, a common, risk-averse paradigm for low-budget HCGs is to copy an existing HCG (e.g., building an output-agreement clone of the *ESP Game*) and then to recruit enough college-age participants to yield a moderate solution.

[2]Not to the mention, between conference page limits and a general aversion towards publishing negative results, valuable anecdotal evidence of failed mechanics or game designs is likely omitted from these publications.

is, behavior considered orthogonal to task completion. Much of HCG design is based on anecdotal experiences from developing these games, however modern games have changed dramatically since many of these design recommendations were made. As games (both for entertainment and human computation) become more sophisticated and the audiences of players interacting with these games continue to evolve, it is entirely possible that that anecdotal HCG design tenets no longer hold true for HCGs (or may differ in utility depending on the kind of task or selected game elements).

4.1.1 Collaboration versus Competition

Crowdsourcing or human computation is an inherently *collaborative* process, in which consensus between task solvers and/or experts (not to mention systems built to emulate experts) is frequently required to validate task solutions. Collaborative consensus therefore manifests as a *verification* mechanic in many games, where the game helps to facilitate verification such as relying on player agreement in the *ESP Game* [2] or through the refining process of recipes in *Foldit* [4]. Additionally, as previously mentioned in Section 3.3.2, collaboration also influences much of the *feedback* to players. Players may be rewarded first for participating in the human computation process (which is by definition, collaborative), not to mention for any collaborative behavior they demonstrate or successful (collaborative) verification they accomplish.

However, anyone who has interacted with modern entertainment-oriented games might be familiar with just how many *competitive* elements these games contain. The presence of game rewards such as numerical scores, leaderboards, level-up systems, and customizable items for in-game avatars, facilitates comparison between players and therefore provides players with a means to compete or outperform each other. At the highest levels of play, the most popular games in e-sports belong to fiercely competitive genres such as battle royale games (e.g., Epic's *Fortnite*) and MOBAs ("multiplayer online battle arenas" such as Riot's *League of Legends*). In all of these games, players are rewarded primarily for *com-*

petitive performance, even in games where collaboration (e.g., team support in Blizzard's *Overwatch*) may be necessary.

Researchers have studied the effects of player collaboration (also described interchangeably as "cooperation") in the context of multiplayer games. Seif El-Nasr *et al*. evaluated and identified common patterns in cooperative play [46]. Other studies have since looked at how collaboration and its converse—competition—affect player experience metrics in a broad variety of game types and genres: motor performance games [80], educational math games [81], and co-located multiplayer games [82].

As previously described in Section 3.5, HCG research has also explored variations in collaboration and competition. *KissKissBan* [69] modified the original collaborative version of the image-labeling *ESP Game* [2] by introducing an adversarial third player to create competitive mechanics that yielded more diverse label results, Goh *et al*. [70] conducted their comparison of collaborative and competitive version of the *ESP Game* versus a non-gamified control application. After measuring both the task results (e.g., the number and quality of image labels) and various player experience metrics (e.g., aspects such as appeal, challenge, social interaction), they found no significant differences in task results. Players did, however, find the competitive version more compelling to engage with. Likewise, my colleagues and I [73] conducted a study comparing collaborative and competitive scoring systems in the context of a (simulated) networked multiplayer HCG called *Cabbage Quest*. We found no significant differences in completed task accuracy, however players found the competitive system more engaging. However, despite all of these investigations, HCGs tend to avoid the inclusion of competitive elements beyond the form of external leaderboards (which are typically isolated from *action* or *verification* mechanics of solving the task) in order to preemptively minimize any negative effect such elements might have on *task completion*.

Taken together, the prevailing game industry practices and the results of these many studies—both for entertainment-oriented games and more specifically for human computa-

tion games—highlight two key points. First, players do find competitive games engaging, often more so when compared to collaborative games, a result that is not entirely unexpected given the demographics (i.e., students experienced in playing games) audiences in these studies. Second, these results demonstrate that competitive game mechanics may not actually adversely affect solving the human computation task. This notion runs counterpoint to concerns that competitive game mechanics might compromise the quality of task results through distracting gameplay.

4.1.2 Collusion

One of the defining characteristics of the original *ESP Game* and other games in its lineage is that while these games facilitate and require multiplayer mechanics, these synchronous players remain anonymous from each other and unable to directly communicate through in-game means. The justification given for these limitations on player communication is that it is absolutely necessary to prevent player "collusion"—the ability for players to jointly communicate and then optimize for game objectives orthogonal to the task, such as individual players' scores [50, 1]. For example, in the *ESP Game*, a degenerate strategy employed by two players able to "collude" would be to (verbally) agree between themselves to enter the same, shortest possible answers for every label (e.g., the letters "a," "b," "c," etc.). This procedure would then ensure that together, both players could enter in and agree on as many labels as possible within the given time period, regardless of whether or not these labels would actual describe the image in question (e.g., a "cat," a "helicopter," etc.). If HCGs (accidentally) enabled such strategies, task providers would then have to grapple with a percentage of players focused solely on maximizing task-orthogonal objectives (or worse, deliberately providing incorrect solutions), which would risk tainting and ultimately compromising the quality of the final task results. This concern is justified particularly given that HCGs do not always have the luxury of an established, (and for some tasks, a trained) player base or an industry-scale marketing campaign to attract users, therefore relying on

smaller audiences where every solution may count.

However, many entertainment-oriented games which support multiplayer make no such efforts to stymie player communication. Cooperative games which support local (i.e., co-located) multiplayer will often present players with side-by-side views of play on the same screen, allowing players to see another player's progress if needed. Likewise, cooperative games played across networked connections will typically support communication through in-game mechanics (i.e., in-game voice chat); if not, external communities and tools are often readily available via websites or through (text/audio/video) chat software during play.

Research on co-location in digital games has validated that players behave differently when playing with or against other human players, compared with singleplayer experiences with or against artificial agents. For example, Webhe and Nacke [83] conducted a study of the effects of co-location on players, finding that players demonstrated higher pleasure and perceived arousal when in co-located multiplayer conditions than in singeplayer conditions. These findings echo those of Mandryk and Inkpen [84], in which players found co-located multiplayer gameplay with a friend to be more engaging—more "fun," less frustrating, and less boring—than the same (singleplayer) experience against an artificial opponent.

Additionally, studies of co-location have also been conducted in the context of educational math games [85, 81]. These studies showed that players demonstrated higher engagement in co-located multiplayer experiences compared with singleplayer experiences. Notably, no differences were found in educational outcomes. As educational math games are analogous to human computation games in their pursuit of a secondary design goal (i.e., enabling positive learning outcomes), these results suggest that while co-location might not improve secondary objectives, it might also not negatively affect such outcomes. Not to mention, for some HCGs, permitting external (offline) communication often yields the refinement and verification of task solutions, such as recipe sharing in *Foldit* [9]. These results suggest that optimistically, allowing players to sit next to each other and communicate without restrictions might be able to improve task completion metrics, outweighing any

negative results from players maliciously attempting to optimize for non-task outcomes.

4.1.3 Non-Puzzle Human Computation Games

A human computation task begins with an unknown input and asks players to provide a solution to a question about that input. A natural mapping of this procedure to game mechanics is to frame each task as a level, a challenge, or a game round within a puzz le game, or a game with puzzle-like elements. Many of these games, which containing other game elements, still involve deduction or reasoning, such manually searching for a desired protein configuration in *Foldit* or determining what the best location is to take a photo in *Photocity*. While these games may be solving different tasks and therefore present players different kinds of task input (e.g., rasterized pixels/picture data, simplified 3D models of chemical structures, etc.), these games are all similar in that the *action* mechanics of these games correspond to puzzle-solving mechanics. However, mechanically (and aesthetically) these games rarely resemble their entertainment-oriented counterparts because of this additional focus on solving the human computation task.

From humorous text-based adventures told through a terminal window to hyper-realistic flight simulators raytraced with the fastest graphical hardware, modern entertainment-oriented games range across a wide variety of genres and types. So then why is it then, that HCGs are often limited to what players might describe as puzzle games, which rarely look or play like their entertainment-oriented counterparts? One potential reason is that task providers typically do not have the development resources, the budget, or the training that industry or even independent game developers do, which often limits the aesthetic and design quality of these games. Another reason is the fear that the addition of any game mechanics that do not map directly to the task (i.e., mechanics which are "non-orthogonal" [55] or are not "isomorphic" [54]) will distract from the problem solving process, thus compromising the results. Even if a task provider wishes to look towards entertainment-oriented genres of games, there are no guarantees about how a game mechanic known for resulting a positive

player experience may affect *task completion* due to a lack of design resources.

The result is that very few efforts have accomplished making a human computation game look and play like a non-puzzle, entertainment-oriented game. Notable exceptions to these include *OnToGalaxy* [22] and the integrations of *Project Discovery* [24] and *Phylo* [30] into *EVE Online* and *Borderlands 3* respectively. Both *Project Discovery* and *Phylo* demonstrate the potential to leverage very large, existing player bases, which suggest that players sufficiently dedicated to a certain game or certain genre of game might be enticed into participating provided that the game was familiar (or looked like such). Thus, it could be entirely possible that making an HCG look like an existing game or modeling it after an existing popular game genre might be able to yield more players or optimistically, improve the quality of the task results.

4.1.4 Summary

Ultimately, the work in this chapter is motivated by three questions: First, how do *competitive* reward mechanics affect *task completion* and *player experience* compared with more traditional *collaborative* reward mechanics? Second, does *collusion* between players actually have an adverse effect on *task completion* and *player experience*? Third, can a human computation game look and play like an entertainment-oriented game; alternatively, can the mechanics from a successful entertainment-oriented game be adapted to a human computation game? (This last question is proposed as purely exploratory; I do not prescribe the answer is a contribution of this work. The implementation in this chapter demonstrates that yes, this is possible, but the full, proper comparison and evaluation is beyond the purview of this dissertation.)

In this chapter, I describe a game, *Gwario*, which was developed to explore these questions about mechanics and elements of human computation games. Specifically, *Gwario* explores *collaborative* and *competitive* scoring mechanics, as well as *singleplayer* and *co-located multiplayer* mechanics—all while looking like a popular platformer game. I then

describe a human-subjects study using pairs of players interacting with different variations/versions of *Gwario* to examine these questions. I also provide the anecdotal details of a short survey regarding the research questions, which was sent to HCG experts (i.e., researchers and developers) independently of this study. I then summarize and discuss the results of the primary study as four design implications regarding these three questions. Throughout this chapter, I use the language of the mechanics framework and follow the experimental methodology for testing game mechanics described in the previous chapter.

I wish to highlight that while this dissertation focuses primarily on *feedback* (reward) mechanics, the latter two questions capture aspects of *verification* and *action* mechanics respectively. I opt to address and discuss these questions as part of this chapter to present the most comprehensive set of findings achieved with *Gwario*.

This chapter consists of four parts:

1. A description of *Gwario*, a game developed to study the three questions proposed above.

2. A human-subjects study using *Gwario* and its results.

3. A small survey of HCG experts regarding the mechanics explored in the study.

4. Four design implications drawn from the results of the study.

For reference, the peer-reviewed version of this work (in its entirety) was published as a full conference paper at the Foundations of Digital Games Conference in 2017 [37]. A preprint of this work can also be found on arXiv [86].

4.2 The *GWARIO* Game

To test all of these hypotheses about various game mechanics: singleplayer versus multiplayer, collusion, and collaboration versus competition, my colleagues and I developed a custom, human computation game called *Gwario*. *Gwario*—a portmanteau of the terms

"GWAP" and "Mario"—is at its core, a 2D platformer game inspired by the original *Super Mario Bros.* (hereby abbreviated as "*SMB*" for this and all future chapters).

Super Mario Bros. asks players to control a player avatar—the titular Mario—as they attempt to navigate and clear a series of 2D sidescrolling levels without falling into gaps and colliding with too many enemies. As Mario, the player may move left and right, and may also jump in order to avoid gaps in the level topology or to land on top of enemies (thereby eliminating them from the level). Various, collectible "powerup" items are also hidden in the level to assist the player. In addition to the primary goal of clearing the level by reaching the end, the game presents players with a secondary objective: increasing a numerical score. This score increases as players eliminate enemies, collect powerups, and (most importantly for this discussion) collect coins scattered throughout the level. However, maximizing this score is effectively optional since it provides no explicit (i.e., mechanical) benefit to the player's ability to complete the levels or the overall game.[3]

One immediate observation is that *Gwario* does not resemble a typical human computation game. Traditionally, HCGs tend to take the form of cooperative puzzle games [50] or gamified interfaces designed specifically around the particular human computation task [4, 87]. HCGs are rarely classified into the popular genres of games designed purely for entertainment (e.g., "role-playing games" or "action-adventure games"). A singular exception to this is *OnToGalaxy* [22], an arcade shooter game whose authors (i.e., developers) emphasize the need to further explore this specific issue as part of their motivation behind the feel and play of their game.

The decision to use an existing game—specifically *SMB*—as inspiration for a human computation game was prompted by the fact that platformers in the *Mario* franchise are considered design exemplars not only for the genre of platformer games, but among all digital games. These games and their characters are both iconic and familiar to players. Additionally, these games are well-studied by video game scholars (e.g., Anthropy and

[3]Games in the *Mario* franchise also do award players for specific scores, such giving the player an extra life for every 100 coins they collect.

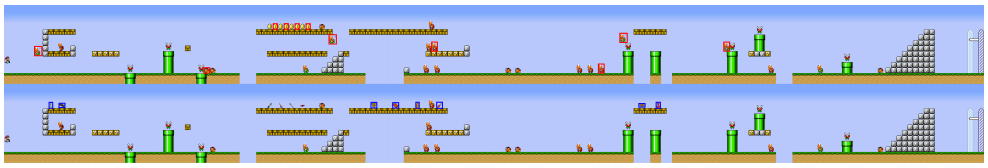

Figure 4.1: A comparison of the original level from *Super Mario Bros.* (top) and the edited level in *Gwario* (bottom), with item/enemy additions in blue and deletions in red. Coins have been replaced with sprites of everyday items.

Clark [8] provide one of many studies into the design of the iconic "Level 1-1" of the original game) and utilized as environments for testing game-related research [88].

In the context of human computation game research, choosing a game as well-studied as *SMB* is analogous to the previously-proposed experimental methodology of selecting a human computation task with a known solution to avoid the complications of solving a novel human computation problem while simultaneously attempting to answer a research hypothesis about HCG mechanics. Here, the game mechanics have been deliberately chosen to ensure that the resulting HCG's mechanics might be as entertaining as possible initially. However, HCGs are not typically designed to adapt mechanics from commercial or entertainment-oriented games. This adaptation necessitated changes to the original mechanics. To ensure that *Gwario* still functions as an HCG despite its origins as a platformer game, my colleagues and I used von Ahn and Dabbish's classical definition of HCGs [50] as our design focus: "[games] in which people, as a side effect of playing, perform tasks computers are unable to perform" with a focus on "useful output." The subsequent sections now describe *Gwario*'s adaptation and how it adheres to this definition.

4.2.1 Adapting *Super Mario Bros.* to an HCG

The original *Super Mario Bros.* does not have an implementation for, and therefore is not playable on, currently-available modern computer hardware, barring (typically hobbyist and unofficially-supported) software emulation of discontinued console hardware. Thus, *Gwario* is implemented in a modern game engine [89], which in turn was based on both the code of an older "Infinite Mario" engine [88] and personal observations of the

original game (in emulation). In order to better match the expectations of a modern audience, *Gwario* uses the equivalent visual components (i.e., 2D sprite sheets) from a later game in the franchise, *Super Mario World*, rather than using or mimicking the original visual elements from the original *SMB*.

As previously mentioned, the original *Super Mario Bros.* also does not function as a human computation game. To adapt it into an HCG, I selected a simple human computation task—per the methodology previously described in Chapter 3—with a known solution: matching everyday items to a set of purchasing locations (e.g., one might purchase a "breakfast cereal" item at the "supermarket"). This task was used in prior HCG work comparing game mechanics [73] and provided an existing gold-standard answer set. As the answer to these problems is already known, it thus becomes possible to measure the accuracy of the task results objectively, without attempting to simultaneously solve a novel human computation problem. The answer set contains around 60 items, each of which can be assigned to one or more of three categories: "supermarket," "department store," and "hardware store."[4]

To incorporate solving this task in *Gwario*, my colleagues and I altered *Super Mario Bros.'s* existing secondary objective by first, replacing collectible coins with images (i.e., 2D sprites) of purchasable items, as seen in Figure 4.1. Next, we altered the game to assign players a purchasing location (i.e., a category) at the start of each level. Players are then asked to explicitly collect only those items which can be bought at their given location. Visibly, this location is displayed at the top of the level screen as a reminder to the player.

Each playable game level has twelve items, four of which respond to one of three purchasing locations ("supermarket," department store," and "hardware store"). This selection ensures that there is always a consistent set of correct answers in each level. Players, however, are not informed of this distribution in order to avoid compromising task results.

Just as players in *Super Mario Bros.* receive points towards their total score for col-

[4]The original answer set used in prior work contained a fourth category—"pharmacy"—but it was omitted from this set as only corresponded to two purchasable items.

lecting coins, *Gwario* players receive points for collecting items. However, the scoring function for item collection is changed to reflect the process of solving the task (as opposed to simply bestowing players a point for collecting a coin or powerup item). First, for collecting *any* item, a player receives some initial "base" points, which reward the player for participating in the human computation process (i.e., players are going out of their way to collect items, even if it does not assist level completion, and this necessitates positive feedback). Additionally, if a collected item correctly corresponds to the player's assigned purchasing location, the player then receives additional points, which reward the player for providing the correct answer. While players are told initially that they receive more points for correctly collecting items, they are importantly *not* informed how many points they receive for collecting an item when they collect it. Furthermore, the player's total score, unlike that in *SMB* are, is not displayed to the player until the completion of the level. While this lack of immediate feedback might be perceived as player-hostile, withholding the player's score until the end of the level is deliberate and helps to ensure that players to not attempt to infer item correctness or incorrectness by reasoning over which or how many items they have already seen.

Just as in *Super Mario Bros.*, a level ends when the player either reaches the endpoint of the level (i.e., the rightmost edge of the level content) or fails to complete the level (by dying) after three "lives" (attempts). Failure to complete the level results in a game over screen at the end of the level, as opposed to a screen which displays the player's score. Note that failure does not necessarily render the task result unusable. At worst, the task results are incomplete for any items that the player could have collected (or likewise avoided) in portions of the level that the player was unable to reach. Any results collected prior to the player's failure are still available for verification.

Gwario contains one additional, but significant design depature from the original *Super Mario Bros.*: the addition of a simultaneous, two-player (multiplayer) game mode. The original *Super Mario Bros.* contained a "multiplayer" mode for the game, but the two

players do not interact with the game simultaneously. [5] Instead, only one player avatar is present on the screen at one time, and players manually switch control of the game when the avatars change. While *Gwario's* multiplayer mode is a significant depature from that of original game, it was based on later multiplayer mode implementations in the franchise, such as that of the game *New Super Mario Bros.*

In *Gwario's* multiplayer mode, the first player controls the "Mario" avatar and the second player controls a "Luigi" (Mario's brother) avatar. Both player avatars are present on the same screen at once and the game's camera view of the level centers on the first player's Mario avatar. This paradigm of using a single camera view, as opposed to games that show two screens side-by-side (i.e., splitscreen) is used by modern games in the Mario franchise.[6]

Regarding the human computation task, each player in multiplayer *Gwario* is assigned a different purchasing location. This purchasing location is set manually (*i.e.*, by the experimenter) for each player at the beginning of the level. In the context of the experiment described below, this purchasing location is made to be the same location assigned to that player for the singleplayer mode of the game (i.e., Player 1 will be assigned "supermarket" both in their singleplayer and multiplayer playthroughs of the game).

Moreover, *Gwario's* multiplayer mode contains two versions of scoring mechanics: *collaborative* and *competitive*, which correspond to two experimental conditions elaborated on below. The differences between these two versions lie in the scoring function and the visual presentation of the players' scores. In the *collaborative* version, both players' individual scores are combined together at the each of each level. In the *competitive* version, each player's score is tracked separately. At the end of each level, players are then presented their scores. In the *collaborative* version, the end level screen displays the single, combined score both players worked to accomplish, whereas in the *competitive* version, the

[5] The original *Mario Bros.* game, however, did include a two-player game mode with both player avatars on the screen at the same time.

[6] While this design choice arguably biases the game in favor of the first player, I chose to remain faithful to the decision made by the Mario franchise game designers.

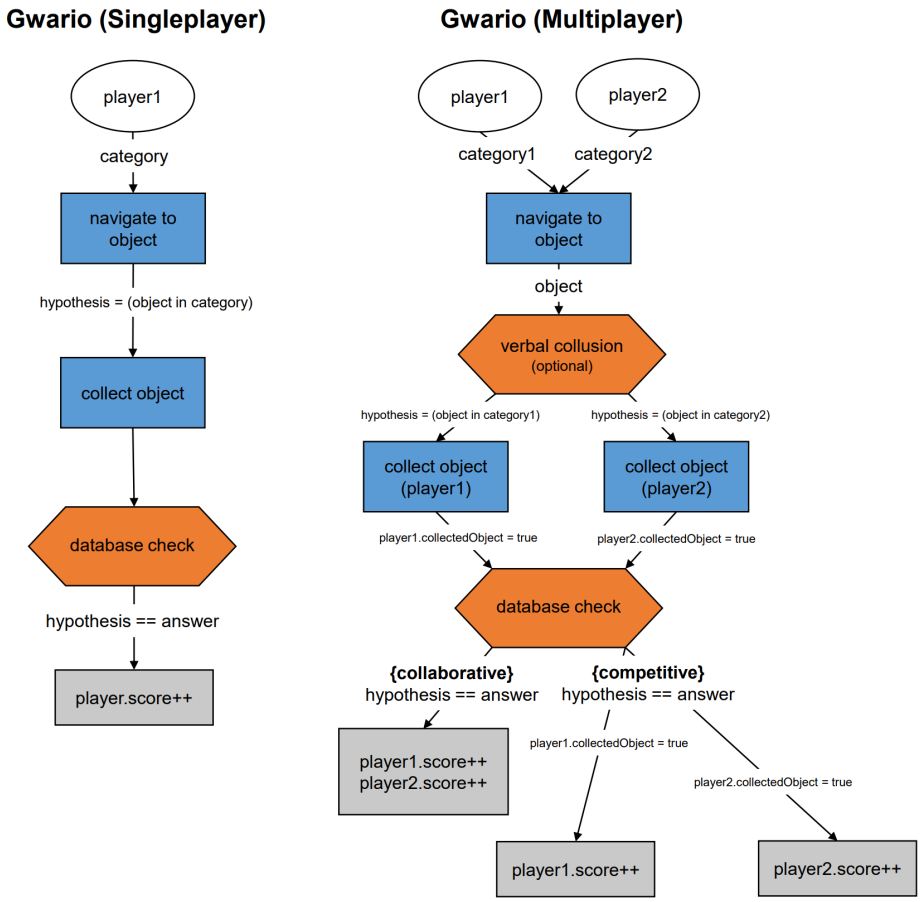

Figure 4.2: The breakdown of *Gwario's* mechanics. The singleplayer and multiplayer versions of the game are split for clarity, while the two versions of multiplayer (collaborative versus competitive scoring) are noted using boldfaced braces.

end level screen displays each player's score on a separate line to allow for visual comparison. The *collaborative* version, with its single joint score, thereby encourages both players to work together to maximize their score. Meanwhile, the *competitive* version, with its separate, comparable scores, encourages both players to compete for the higher score.

Figure 4.2 shows the full comparative breakdown of the mechanics between the singleplayer and multiplayer versions of *Gwario* using the mechanics framework from Chapter 3. The diagram demonstrates that while the basic *action* mechanics are the same between the two versions, the multiplayer version adds not only another player, but also tracks additional *verification* mechanics—the ability for players to verbally communicate as a conse-

quence of co-located play. Likewise, the *feedback* mechanics in the multiplayer version have been extended to allow for another player and a choice of either a collaborative or competitive scoring function.

4.2.2 Game Levels

The levels of the original *Super Mario Bros.* are among some of the most played and iconic platformer environments in gaming history. However, using these levels directly in *Gwario* presents a potential problem, as experimental participants who may have previously played *Super Mario Bros.* or any similar game in the Mario franchise would than have an advantage over participants who might not. Conversely, given the expectations of quality game design associated with games in the Mario franchise, it is likewise unrealistic to assume that researchers (i.e., experts in game research, but not necessarily professional game design) might be able to design new game levels of equivalent or expected quality. Thus, *Gwario* uses four levels from *Super Mario Bros: The Lost Levels*, a Japan-exclusive sequel to the original *SMB*, which would be less familiar to a Western audience.

These four levels underwent some transformations in order to implement them in *Gwario*. First, the existing coins in each level were changed to sprites of unique purchasing items. Next, items that were spaced too closely together were either removed or respaced, as every item needed enough space to be avoidable by a player (i.e., by jumping over or taking an alternative route through the level). Then, in order to ensure uniformity between the levels, additional items were added to levels that contained less than twelve items. These new items were added in locations similar to those that existed in the level initially, as demonstrated in Figure 4.1. At the end of this process, each level contained an equal share of coins-changed-to-items from the three purchasing locations ("supermarket," "department store," and "hardware store").

Super Mario Bros.: The Lost Levels is somewhat notorious among players for its intense difficulty. Based on preliminary personal and colleagues' playthroughs of the game, it was

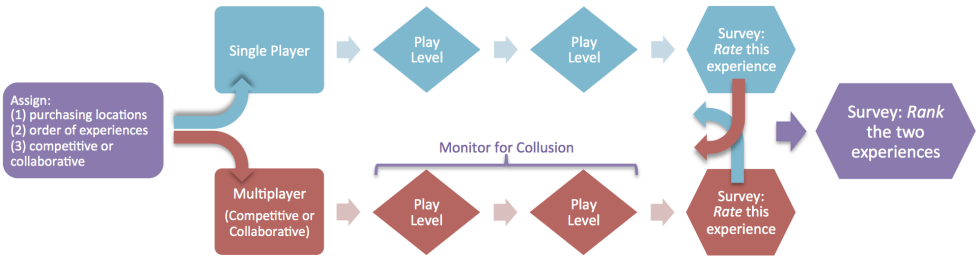

Figure 4.3: A flowchart of the methodology used in the *Gwario* study.

anticipated that this difficulty would negatively impact any study results. To scale down the difficulty, a serial of initial changes were implemented: removing aerial enemies from the level entirely and replacing especially difficult jumps by additional additional platform or floor blocks.

To address any concerns about the appropriateness of these design changes, I conducted a pilot study with ten subjects in five pairs in which subjects played through the levels and reported on their difficulty. Based on both reported and observational results of the pilot—which indeed verified that these levels were still considered very challenging—the levels were then adjusted even further. First, the maximum jump distance between particularly wide gaps was decreased by adding even more blocks around gaps. Next, powerup items were added to the beginning of each level to give players additional assistance. Finally, the density of enemy groups was lowered as not to overwhelm players with large numbers of enemies. An example of one of the final levels in comparison to its original is shown in Figure 4.1. One should note that this level already contained powerups (i.e., hidden within the question mark blocks) at the beginning, so no additional powerups were added for this particular level.

4.3 Methodology

Equipped with *Gwario's* singleplayer and (two) multiplayer modes, my colleagues and I conducted a within-subjects study with pairs of participants. The study consisted of two round of gameplay, followed by surveys after each round as well as a final survey following

both rounds. Each round of gameplay used a different game version, either singleplayer or multiplayer, and within the multiplayer condition, either the collaborative or the competitive version. Figure 4.3 shows a visualization of the overall study flow.

A primary consideration in the study was to emulate a casual play experience as much as possible. This factor guided many of the experimental design choices, such as tracking collusion by hand instead of relying solely on microphones, utilizing post-play, external surveys instead of building these intrusive measures into the game, and tracking whether or not pair of participants were *natural* or *artificial*. Regarding this last item, participants were able to sign up in pairs or they could sign up individually to be paired by the experimenters. Pairs of participants who signed up together are distinguished as self-selected or *natural* pairings, as the participants were more likely to have been acquainted prior to the experiment. Pairs of participants who were scheduled together by the experimenters are referred to as *artificial*, as such participants were unlikely to have paired up to play a co-located game together in a natural setting.

Upon arrival for the study, pairs of participants were randomly assigned to play either the singleplayer or the multiplayer round of the game first. For the multiplayer round, the pair was also randomly assigned either the collaborative or the competitive version. For each round, participants played two of the four levels (the previously-described, adjusted game levels from *Super Mario Bros.: The Lost Levels*). To reduce the effects of ordering and difficulty, these levels were randomly assigned across the conditions upon arrival: two levels for the singleplayer round (in a random order) and the remaining two levels for the multiplayer round (also in a random order). Each participant played the same two levels, in the same order, for the singleplayer round; both players played the same two levels together during the multiplayer round. This random ordering of game conditions and levels, as well as the choice of collaborative or competitive multiplayer, was generated using a simple computer script to avoid bias.

For the singleplayer round, participants were seated at separate computers and each

played through the two singleplayer levels in isolation. For the multiplayer round, the participants were seated at the same computer and presented with a single screen for gameplay (i.e., no splitscreen). During the multiplayer round, participants were explicitly told the victory condition—either collaborative or competitive scoring—and were also told that they could communicate if they so desired.

During the multiplayer round, the individual conducting the study would tag whether or not collusion (i.e., any discussion of the task) occurred between the participants, as well as any relevant quote(s). These quotes were later verified as instances of collusion by a second individual, with disagreements resolved via further discussion. A verified example of collusion would be players discussing what category a particular item belonged to (e.g., "I believe that's your category, not mine.") whereas communication unrelated to the task was not classified as collusion (e.g., "Oops, I fell into a pit.") and excluded.

There were no time limits imposed on play, however as previously described, participants were given three "lives" or chances to restart upon death for each level. In the multiplayer version, if one participant exhausted all of their lives, the remaining participant was still allowed to progress through the remainder of the level alone (while their life-exhausted partner watched). As previously mentioned, even if a participant exhausted all of their lives, their task results remained usable and level completion was not tied to anything but player experience and satisfaction.

Finally, after each round, participants were asked to answer several survey questions about their experience with the just-completed round. After both rounds, participants were given a final, longer survey to establish demographic information and to compare their experience across both rounds.

4.3.1 Evaluation Metrics

To understand the efficacy of *Gwario's* game mechanic variations, the evaluation focuses on metrics concerning both the *player experience* and the *task completion*. This section

outlines precisely what data and information were gathered.

For evaluating the *player experience*, I report on data from the post-round and the post-game surveys, as well as logged gameplay events and observations of collusion. Regarding the surveys, participants were asked Likert-like questions on a scale of 1-5 about their perceived fun/engagement, challenge, and frustration with the experience after each round (of singleplayer or multiplayer). After both rounds, players then provided comparative rankings between the two conditions (i.e., between singleplayer and multiplayer) for perceived fun, challenge, frustration, and finally overall preference.

For evaluating the *task completion*, I report on additional results from telemetry logging (i.e., data recorded from user interaction events in the game). Task answers were logged throughout gameplay and compared against our gold-standard answer set to determine correctness. A player's overall *accuracy* at the task was then calculated as the percentage of their correctly-collected items over all collected items (e.g., if a player collected six of ten items correctly, their overall accuracy would be 60%). In addition to this information, the game logged the number of tasks completed overall, as well as the times (in seconds) it took players to answer tasks and to complete the levels.

4.4 Results

My colleagues and I collected results from 64 individuals in 32 pairs over a two week period. We advertised for participants in two undergraduate computer science courses (and further via word of mouth). 18 of these participants self-identified as female; 46 self-identified as male. Nearly all participants reported themselves between the ages of 18 and 24, with three participants reporting age 25 or older. 73% percent of participants reported that they played games regularly; all but 10 of these participants reported playing a platformer game before. While this distribution is admittedly a somewhat homogenous mixture of participants, I contend that the majority of participants having played games before is an acceptable stand-in for a population of players who would play a game that

looks and feels more like a conventional, entertainment-oriented game than a traditional HCG. Additionally, seventeen participants reported having tried or played an HCG before, which represents a significant number of experienced, or at the very least well-informed, HCG players.

The study focused on co-located pairs of subjects. 15 of the 32 pairs were reported as *natural*. As previously described, a *natural* pairing meant that both individuals signed up to take the study together purposely. The remaining seventeen pairs were *artifical*, consisting of participants who signed up to complete the study without a predetermined partner. These singular participants were randomly assigned to an available partner based on their available time slots.

For the multiplayer rounds, 8 *natural* pairs were randomly assigned to the *collaborative* version and the remaining 7 to the *competitive* version. Similarly, for the *artificial* pairs, 8 pairs were randomly asssigned to the *collaborative* version and the remaining 9 to the *competitive* version.

While there was concern that differences (i.e., prior acquaintanceship with a partner) might impact the results, no significant differences were found between *natural* and *artificial* pairs of participants across any of the subjective (i.e., *player experience*) or objective (i.e., *task completion*) metrics described below. Additionally, no significant differences were found when accounting for participant gender.

4.4.1 Subjective Metrics—Player Experience

Surveys were used to collect participant responses on subjective experience for self-reported fun (i.e., engagement), frustration, and challenge with five-point Likert-like ratings (1 being the least and 5 being the highest). Overall experience preference was collected as a comparative ranking between singeplayer and multiplayer (e.g., of the form “singleplayer” was “more challenging” than “multiplayer”).

No significant differences were found for subjective items with Likert-like ratings (us-

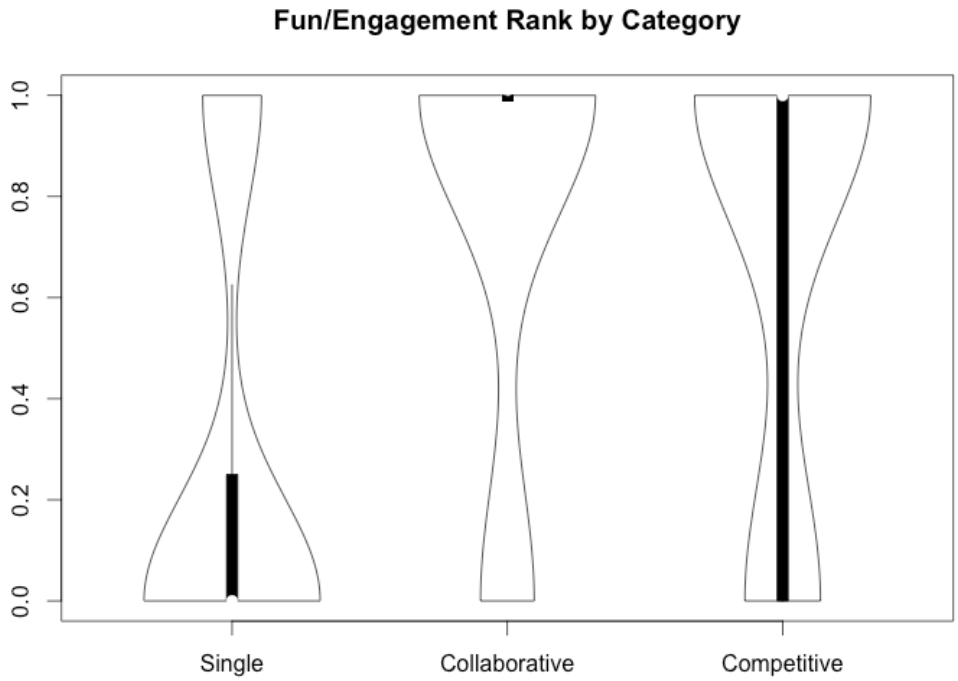

Figure 4.4: A violin plot of fun/engagement rankings across the *Gwario* gameplay variations from 1.0 (most) to 0.0 (least) fun/engaging. The diameter across the length of the violin indicates the number of results of that value—showing subjects preferred multiplayer to singleplayer. The dark bar in each violin runs between the first and third quartiles.

ing the paired Wilcoxon Mann-Whitney U test), except in one case. Specifically, participants in the *competitive* multiplayer condition rated the multiplayer round to be significantly more challenging than the singleplayer round ($p < 0.01, U = 365.5$).

The ranking data were much more discriminatory. As shown in Figure 4.4, participants across both multiplayer conditions ranked multiplayer as more fun/engaging than singleplayer. Additionally, participants overall preferred multiplayer to singleplayer (using the paired Wilcoxon Mann-Whitney U test, ($p < 0.05, U = 170$)). Finally, participants in the *competitive* multiplayer condition ranked multiplayer as more challenging than singleplayer ($p < 0.01, U = 17.5$), which reinforces the previous Likert-like challenge rating results from participant surveys.

4.4.2 Objective Metrics—Task Completion

Several objective metrics were used to measure participant performance at the human computation task: the per-task accuracy (i.e., the percentage of correct task assignments to total tasks in a given round), the average number of tasks completed, and the time it took a participant or pair of participants to complete a level (successfully or not). Across all study conditions, participants had an average 81.7% per-task accuracy. Additionally, participants completed an average of 4.4 tasks per attempt, where an "attempt" corresponded to a player's life (from the moment they appeared in the game level till the moment they either completed the level or died due to environmental hazards). As participants were given three lives for each level, these numbers taken together with other metrics and factors (e.g., the fact that each level contained twelve items—possible tasks) suggest that the participants were able to perform fairly well at the task, despite having to also simultaneously play through the game level.

Table 4.1: A summary of the objective results of the *Gwario* study across the three game conditions.

	Singleplayer	Collaborative	Competitive
Accuracy Avg.	82%	86%	77%
Time(s) Avg.	212	612	505
# of Tasks Avg.	22	16	14
# of Deaths Avg.	4.9	4.7	5.1

Table 4.1 summarizes these objective metrics by study condition. From these averages, some distinctions between the two multiplayer and the singleplayer conditions emerge. Focusing first on task accuracy, collaborative multiplayer has the highest average, however there is no significant difference between any of these distributions. Within the two multiplayer conditions, however, there is a significant difference between the *collaborative* and *competitive* conditions: players in the *collaborative* condition had significantly higher task accuracy than in the *competitive* condition (using the Wilcoxon Signed Rank test ($p < 0.05$)).

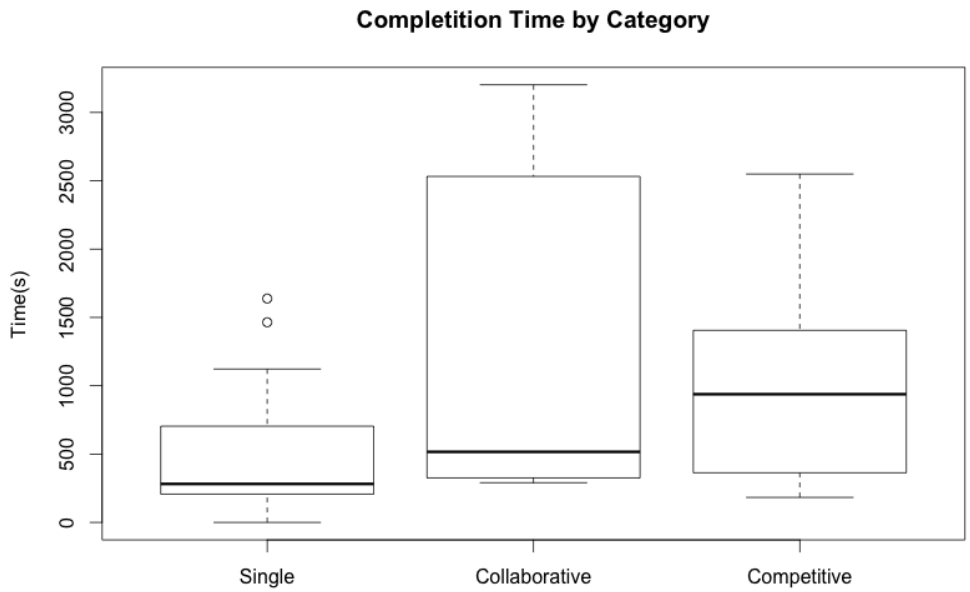

Figure 4.5: The level completion times across the singleplayer and both multiplayer conditions in the *Gwario* study.

Focusing next on average times, the results of Table 4.1 might suggest that collaborative playthroughs of the levels took much longer than either singleplayer or competitive multiplayer playthroughs, but this is not true for most cases. Figure 4.5 demonstrates that collaborative playthroughs had a far wider distribution of times and a high upper bound, but that the median collaborative time is below the median time.

The third row of Table 4.1 also shows the average number of tasks completed across both levels within the condition (i.e., not the average number of tasks per attempt). Here, participants accomplished the most tasks in the singleplayer mode, with *collaborative* and *competitive* multiplayer second and third respectively. Finally, the fourth row shows the average number of deaths per participant across levels. Despite participants consistently rating and ranking *competitive* multiplayer as more challenging than singleplayer, no significant difference in the number of deaths was found between the two conditions.

Additional analysis was conducted to better understand the variance in accuracy between the *collaborative* and *competitive* conditions. Due to the non-ordinal nature of the data, an ANOVA with permutations was used to consider various factors: multiplayer con-

dition, whether or not the pair was *natural* or *artificial*, all demographic information, and collusion—whether or not the participants spoke about the task during gameplay. Of these factors, collusion was the only significant predictor of accuracy ($r = 12.24, p < 0.01$). This result may also suggest why the collaborative condition had the greatest variance in time, as participants would have taken varying amounts of time to discuss the tasks between each other.

4.5 Expert Opinions

On top of these results and to better contextualize these findings in the space of human computation game design, my colleagues and I also pursued expert opinions on the mechanical variations that were tested in *Gwario*.

My colleagues and I identified experts who were responsible for developing nine recent human computation games (all of which were described in Chapter 2). An expert was considered to be someone who had developed an HCG and (with one exception) had authored peer-reviewed publications on the findings or design of the game. None of the experts we queried were made aware of the *Gwario* study or its results prior to taking the survey.

These experts were sent an online survey which consisted of four topics. The first three topics covered different variations in human computation game mechanics: single-player versus multiplayer, collaborative versus competitive scoring, and prohibiting versus permitting direct player communication (i.e., disabling or enabling the opportunity for collision). For each of these topics, experts were asked how they thought a hypothetical game employing the second condition would compare to the first when considering two metrics: task accuracy and player engagement.[7] For each metric, experts then chose from three multiple choice answers: "increased (accuracy/engagement)," "no difference (in accuracy/engagement)," and "decreased (accuracy/engagement)." Experts were then

[7] I use the terms "task accuracy" and "player engagement" here (as well as in the survey) instead of *task completion* and *player experience* as the prior two are more common terminology (i.e., the two primary metrics that most HCGs are concerned with measuring), even if I prefer the latter terms for generality.

able to provide optional short-answer descriptions about their choices. The last topic covered a longstanding HCG design question—should mechanics from successful digital (i.e., entertainment-oriented) games be incoporated into HCGs? Experts provided an answer of "yes/no/maybe" as well as an optional short-answer elaboration.

Table 4.2: The results of the expert survey asking about expected accuracy and engagement for the variations in HCG mechanics tested in the *Gwario* study.

	Multiplayer would be	**Competition** would be	**Direct communication** would be
Expert 1	no more accurate more engaging	more accurate no more engaging	more accurate more engaging
Expert 2	more accurate more engaging	no more accurate more engaging	more accurate more engaging
Expert 3	more accurate more engaging	less accurate no more engaging	more accurate more engaging
	than **singleplayer**	than **collaboration**	than **no communication**

Three experts responded to the survey. Due to the small number of respondents, I acknowledge that these results are limited and at best, anecdotal, but I have opted to include this information to provide a broader context for the results of the *Gwario* study. [8] The results of these questions are summarized in Table 4.2.

When looking at singleplayer versus co-located multiplayer, experts agreed that co-located multiplayer would increase (or have no difference in accuracy and player engagement). This suggests that co-located multiplayer might be perceived as more effective and beneficial than singleplayer. Notably, Expert 3 compared the benefits to pair-programming, but expressed that "local, as opposed to remote, co-operative games are harder to coordinate."

When looking at collaborative versus competitive scoring, experts did not agree on how it would affect task accuracy. Expert 1 noted that "competition could increase accuracy for certain tasks because it gives players a way to measure themselves and their contributions."

[8] Additionally, it is still the case that there is very little work in surveying HCG developers on game design wisdom.

By contrast, Expert 3 noted that "competitive players will always find ways to win which do not advance scientific goals if they are more expedient," suggesting that adding competition could encourage competitive players to provide less useful results at the expense of non-competitive players. Regarding player engagement, experts were also mixed, but all agreed that competition would not decrease player engagement. In particular, Expert 2 cited examples of HCGs where competition was shown to have positive benefits on player engagement.

When looking at prohibiting versus permitting direct communication, experts all agreed that direct player communication would lead to both increased task accuracy and player engagement. However, Expert 1 did caution that allowing direct communication would not work "if the mechanics are directly related to players coming up with ideas (e.g., *ESP Game*)," but noted that not all HCGs follow the same format.

Finally, all experts agreed that mechanics from successful digital games should possibly be incorporated into human computation games, with Expert 1 stating "maybe" and Experts 2 & 3 stating "yes." When asked why, experts focused on player familiarity with game mechanics, but noted possible concerns with incorporating these mechanics in HCGs. In particular, Expert 1 was concerned that mechanics which did not complement the task might compromise the task results (i.e., mechanics should be "isomorphic" per Jamieson et al. [54] or "non-orthogonal" per Tuite [55]). Expert 3 remarked that determining an effective mapping of mechanics from successful games (where players have different motivations for play) to HCGs remains an open question.

4.6 Discussion

In this section, I summarize the major results of the *Gwario* study (with the occasional anecdote from the expert surveys), comparing and contrasting these results against relevant prior work. Given the focus on human computation game design, this discussion section

is organized into a set of four design implications[9] and how these designs relate to both creating an engaging player experience and effectively solving the human computation task.

These design implications are as follows:

1. Mapping item collection mechanics to human computation classification tasks
2. Allowing direct communication between players
3. Supporting synchronous competitive multiplayer
4. Supporting synchronous collaborative multiplayer

4.6.1 Mapping Item Collection Mechanics to Human Computation Classification

There exist a myriad of games where exploration for collectible items (e.g., coins, rings, notes, power-ups,) is a major, but not necessarily primary, gameplay mechanic (e.g., consider classic game franchises such as *Mario*, *Sonic the Hedgehog*, and *Banjo Kazooie*). In the *Mario* franchise specifically, this mechanic is considered *secondary*; players may collect both coins and powerups, but engaging in this collection is optional (although not necessarily orthogonal) for completing the level.

In *Gwario*, the collectible items are converted from coins into purchasable items that require classification. Thus, the player's choice to collect an item or not maps to the process of answering the human computation task, making such mechanics most appropriate for classification or categorization tasks, ideally where the classes or categories are known a priori. Outside of *Gwario*, *OnToGalaxy* [22] is an HCG that explores a similar adaptation: an archetypal "space shooter" transformed into an HCG by altering collectable objects into task answers. *OnToGalaxy* differs in that its categorization tasks are much fuzzier (e.g., collect [items] that could be labeled as "touchable objects").

[9] I am obviously not suggesting that all HCGs implement these, but discuss their viability and potential implementation in the context of similar tasks or games.

Experts suggest that incorporating mechanics from successful digital games should be considered for HCGs, but caution that the mechanics should be appropriate for the underlying human computation task. *Gwario's* implementation maps *collection* mechanics to the task of *categorization*, and appears to successfully address both dual design goals of a positive *player experience* and effective *task completion*. Across conditions, we found an average accuracy of more than eighty percent and a consistent median Likert-like rating of "4" for fun/engagement on a five-point Likert-like scale. These above average values suggest strong evidence that this design retains much of the fun and familiarity associated with the original *Super Mario Bros.*, while providing an effective interface for completing the item-purchasing task. However, based on the adjustments before and after the pilot study, one might expect the impact of using this "item collection as task categorization" design to vary wildly depending on how challenging the base game (or inspiration) is and may require additional changes to the underlying mechanics or game content.

Ultimately, further comparison on a larger participant population against another HCG with the same human computation task (but different mechanics) would be required to truly determine just how much more compelling (of an experience) and effective (as a task-completion interface) *Gwario* is be compared to other HCGs.[10]

4.6.2 Allowing Direct Communication in Multiplayer HCGs

In games, direct communication between players can take on many forms and interfaces, ranging from (but not limited to) in-person conversation when co-located, audio or video chat, a text-only interface, or some discretized set of allowable messages. Communication may also also occur extradiegetically (i.e., outside of the game's interface) through applications that enable the above methods of communication, or even through game forums, online wikis, and help websites. Human computation games, however, typically

[10]Such a study would still not necessarily be able to tease out precisely which mechanics might make *Gwario* more or less compelling/effective, a problem which can be seen in results of Krause *et al.* [22] which outright compares two completely different games and tasks.

dissuade player communication altogether and implement mechanics (e.g., anonymity) to prevent players from communicating extradiegetically in real-time. This is due to preexisting paradigms which suggest that collusion between players could hurt task accuracy [50, 90].

In *Gwario*, co-located multiplayer enables direct communication by forcing players to sit within verbal proximity of each other (and then places no restrictions on verbal communication). To the best of my knowledge, *Gwario* is still the only current example of an HCG wherein co-located, direct communication between players is permitted *during the process of completing the task*. However, asynchronous and indirect communication have been previously explored in HCGs. For example, *Foldit* [4] permits asynchronous communication by supporting online forums and allowing players to share partial solution strategies ("recipes") in game.

Direct communication has been shown to benefit games in other dual-purpose domains, such as educational math games [81] in which side-by-side student play has been shown to increase learning motivation and engagement. In a domain outside of games entirely, this design manifests in the computer science practice of *pair-programming* in which two individuals solve a programming problem at the same time, typically while seated at the same device—a paradigm that has shown impressive positive results particularly in CS education [91]. This analogy was cited by one of the experts in our surveys (Expert 3).

Furthermore, in contrast to preexisting concerns about collusion, our expert surveys suggest that direct communication between players is perceived to have benefits for both player engagement and task accuracy. The *Gwario* study found direct evidence of co-located player communication's impact task accuracy, as player collusion was a significant predictor of increased task accuracy. While we did not find similar evidence linking communication and engagement, the significantly-higher ranking of multiplayer fun/engagement in comparison to singleplayer suggests a potential connection. These findings from both the study and expert surveys question the assumptions made by prior HCG research which

suggest that players may try to "game" the game to gain more rewards (e.g., in *Gwario*, players might presume such a strategy to be to collect all of the coins regardless of correctness to improve their end score). However, it is worth pointing out that there are strategies to remedy and often neutralize potential data-tainting, such as cross-validating the results across many dyads (i.e., more players) or identifying player reliability based on performance on preexisting task solutions (which can still be implemented as part of a game permitting direct communication).

As with all human computation game designs, incorporating mechanics that permit direct communication may still be sensitive to particular task and desired gameplay mechanics. *Gwario's* multiplayer requires implementation of both multiplayer mechanics *and* for players to gather at the same physical location for play—an onus that may not be amenable for all tasks. Additionally, while the most predictive factor of accuracy was our boolean measure of collusion, the most predictive factor of collusion was the *collaborative* multiplayer condition. In other words, players instructed to collaborate were more likely to positively collude. One could therefore argue that this *suggesting of collaboration* might have more of a similar effect to permitting direct communication. Thus, further verification (not to mention whether or not similar results could be achieved over networked multiplayer) is required to fully validate the design against preexisting negative hypotheses about collusion. Nonetheless, this and other prior research extolling benefits of direct communication could prove potentially useful in other serious game applications outside of HCGs.

4.6.3 Synchronous *Competitive* Multiplayer HCGs

In the context of human computation games, *synchronous* competitive multiplayer is facilitated through the inclusion of game mechanics which permit two or more players to make real-time, simultaneous decisions in an effort to outperform one another. "Outperforming" in this context typically refers to a player achieving a higher score (or other kind

of feedback/reward) than that of the other players. Many HCGs contain competitive gameplay elements in the form of leaderboards, but synchronous competition between players is less common. Examples of HCGs include many of those previously discussed in Chapter Chapter 3. In *KissKissBan* [69], the third player synchronously competes with the two other players who are working together to collaborate; all players are still contributing answers to the underlying human computation task. In both Goh *et al*.'s *ESP Games* [70] and *Cabbage Quest* [73], players (either collaborate or) compete to tag items as quickly as possible. However, *Cabbage Quest* differs from these examples in that players do not compete in person, but against an artificial player that they are led to believe is another human.

Gwario's results strongly suggest that competitive multiplayer is viewed as significantly more challenging and more fun/engaging than singleplayer in *Gwario*. These results are in line with the results from both *KissKissBan* and *Cabbage Quest*, which suggests that adding competitive elements could promote a more positive *player experience*, especially with a simple (and potentially mundane) task (as in *Cabbage Quest* and *Gwario*). Meanwhile, Goh *et al*. found no difference in player engagement between the competitive and collaborative versions of their *ESP Game*, suggesting that competitive gameplay elements were no worse than their collaborative counterparts. Altogether, these results echo the expert opinions, which suggest that competition will not negatively impact player engagement.

Meanwhile, the effects of competitive game mechanics on *task completion* are less straightforward and potentially negative. One design concern with using competitive mechanics in HCGs is a possible negative impact on task accuracy due to distraction (i.e., orthogonality) from the underlying task. Likewise, the expert surveys contain differing differing opinions on exactly how competition might impact accuracy. One expert-suggested benefit of competition is the potential feedback for players to measure themselves and their contributions, but one expert-suggest detriment is that competition might encourage expedient (but not necessarily correct) solutions. In the *Gwario* study, competitive players were significantly less accurate than collaborative players. By comparison, both Goh *et*

al.'s *ESP Games* and *Cabbage Quest* found no such significant difference. This discrepancy could be due to any number of differences between *Gwario* and these other games. In both Goh *et al*.'s *ESP Games* and *Cabbage Quest*, players were not co-located, could not communicate, and had fewer mechanics available to antagonize the other player. In competitive *Gwario*, players could antagonize each other by stealing power-ups from one another, attempting to hurt/kill each other with shells, and jumping around to distract (and thus impede) their partner.[11] While some players seemed to enjoy these affordances, the existence of these mechanic may have led to the poorer accuracy compared to that observed in *Cabbage Quest*, which utilized the same human computation task.

4.6.4 Synchronous *Collaborative* Multiplayer

In the context of human computation games, *synchronous* collaborative multiplayer is facilitated through the inclusion of mechanics that require two or more players to work together in real-time to improve the same in-game reward (i.e., score) or end result. Nearly all synchronous, multiplayer HCGS, dating back to the original *ESP Game* [2] reward players for this kind of collaborative interaction and play. Collaborative mechanics often map neatly to the structure of the human computation process, in which verification of the result may be accomplished through aggregated agreement. Feedback can often be provided quickly, as player agreement may be enough to validate the (initial) task results, yielding synchronous, real-time play. The collaborative version of *Gwario* is similar to existing collaborative design paradigms, but differs slightly from many historical examples in that both players are separately categorizing tasks using a more discrete pool of possible categories (but this still ensures that omitting an object from one category is verification that it may belong to another).

The results from the *Gwario* study match previous expectations given the historical

[11] Another issue brought up previously is that the camera tracks (i.e., focuses on) the first player, meaning that the first player can run far enough ahead/back so that the second player can no longer be seen on the screen, which is a known tactic for antagonizing players in games with similar camera setups.

use of collaboration in human computation games. Participants found collaborative multiplayer more fun/engaging and overall preferred it to singleplayer. In comparison to competitive mechanics, prior work has suggested that competitive mechanics are more engaging, but that certain aspects of the *player experience* may be higher for collaborative player (e.g., player empathy in the collaborative version of *Loadstone* [82]). Meanwhile, the surveyed experts were neutral (with one exception) on the idea that collaborative multiplayer was more engaging than competitive multiplayer. As the multiplayer conditions in the *Gwario* study were conducted between subjects, direct comparison between the collaborative and competitive versions was not possible (as participants only played and ranked one multiplayer version). Participants in the collaborative multiplayer condition did rank fun/engagment higher than their counterparts in the competitive version (on a 1-5 Likert-like scale), however this difference was not significant.

Combining collaborative mechanics and the affordance of direct communication during multiplayer play resulted in a significant increase in accuracy in comparison to the competitive version. These results are different from those of Goh *et al.*'s *ESP Games* and *Cabbage Quest*, which found no significant difference in accuracy between their competitive and collaborative versions. However, the difference here may be due to our choice of additional design features, such as permitting direct communication (i.e., whereupon collusion was shown to be a predictor of task accuracy). As previously mentioned, *Gwario's* variation of collaborative multiplayer asks players to consider two distinct tagging tasks simultaneously as opposed to working towards agreement on a single task. This setup was shown to be successful in terms of accuracy, time, and participants' self-reported fun/engagement, which suggests that *Gwario's* variation of collaborative mechanics could be used to reduce the size of the player base required to solve an HCG in the context of a similar task.

4.7 Conclusions

In this chapter, I describe a human computation game, *Gwario*, and a study comparing different mechanics variations using *Gwario*. The study compared singleplayer and co-located multiplayer versions of *Gwario*, and within the multiplayer condition, two versions of scoring: *collaborative* and *competitive*. I also describe the limited results of a survey sent out to several HCG experts about the mechanics variations tested with *Gwario*. From the combination of study results, survey results, and the context of the existing HCG design research, I discuss four design recommendations around these different HCG mechanics.

This chapter began with three questions. The first question was: how do *competitive* reward mechanics affect *task completion* and *player experience* compared with more traditional *collaborative* reward mechanics? For prior HCGs investigating similar questions, *competitive* reward mechanics have shown no difference in *task completion* compared with *collaborative* mechanics, while demonstrating more positive results for *player experience* metrics. However *Gwario* demonstrates a scenario where this is not the case. Participants in the *competitive* version were less accurate at the task than *collaborative* players, which demonstrates that the concerns that competitive game elements might distract from task solutions are not entirely unfounded. However, participants did rank the *competitive* version of the game as more fun and challenging than the *collaborative* (not to mention the *singleplayer* versions). Ultimately, some degree of competitive mechanics may help to improve participants' experiences with the game, but depending on the context, too much competition may conversely hurt *task completion* metrics.

The second question was: does *collusion* between players actually have an adverse effect on *task completion* and *player experience*? Longstanding HCG wisdom suggests that *collusion* is anathema to accurate *task completion* results, since malicious players could provide incorrect results by choosing to optimize for non-task objectives (e.g., game score). However, this study demonstrated the opposite: *collusion* was actually a predictor of task

accuracy. Furthermore, a small sample of HCC experts independently suggested this conclusion, suggesting that collusion may actually be more beneficial to human computation games than previously assumed. Adversarially-minded players will always be a concern for designers in all games, both entertainment-oriented and human computation. Given the benefits of collusion demonstrated in this study, perhaps HCG developers and researchers ought to look towards entertainment-oriented games and how these games deal with negative player behavior, as opposed to imposing design restrictions that ban collusion entirely.

Finally, *Gwario* provides an example of an HCG that can be made to look like an existing platformer. The *task completion* results from the *Gwario* study demonstrate that it can in fact successfully solve a human computation task with a known solution. I acknowledge that these results, along with the design we selected (wherein the process of task completion is adapted into the *action* mechanics facilitating the game's secondary objective), are at best anecdotal evidence that HCGs can look like platformer games. Whether or not a platformer (or platformer-like mechanics) is the best choice for solving a human computation task would require an entirely different experimental setup where the same human computation task is implemented in multiple games with different game mechanics (e.g, a "puzzle" HCG version of the purchasing task compared with a "platformer" HCG with the same task).

CHAPTER 5
REWARD SYSTEMS — PLAYER AUDIENCES AND CHOOSING REWARDS IN HCGS

5.1 Introduction

For some players, it is the rush of adrenaline upon claiming victory in a fierce round of competition. For others, it is the promise—and subsequent fulfillment—of exploring an unknown and open world. For yet others, it might be the emotional roller coaster of the story, its compelling characters, or the plot twists enabled through play. There are nearly limitless reasons—motivations—for why players choose the games they play.

Playing games may once have been marketed at specific kinds of players, as a niche hobby requiring niche hardware. However, as the barrier for entry to digital games has been lowered (i.e., through everyday platforms such as mobile devices and web browsers), games have become more mature and accessible to demographics of players who might not have initially considered play. Broader player audiences bring new motivations for play, making traditional game incentives (i.e., those implemented as *feedback* or reward mechanics) potentially less compelling to new players. In response, game design for modern, entertainment-oriented games have explored many kinds of reward systems for providing both explicit and intrinsic rewards.

By contrast human computation games have been slow to adopt these systems, relying instead on point-based numerical systems, without truly taking in the vast potential audience of players into account. I reiterate that this is a concern for two reasons. First, serious games are complicated by the addition of a secondary goal, one which may be orthogonal or unrelated to entertainment. Second, research has shown that there do exist players who are intrinsically motivated to solve the task [9] while other research shows that ex-

trinsic rewards undermine the effects of intrinsic motivation in crowdsourced workers [75]. This might suggest that focusing only on extrinsic rewards, such as point-based numerical systems, may come at the cost of disincentivizing potential players who would otherwise participate in the task solving process.

As previously emphasized in Chapter 4, there is a need to understand how one might adapt game mechanics from successful, entertainment-oriented games. The previous chapter explored, among other game mechanics, how collaborative and competitive reward systems affected *task completion* and the *player experience*. In this chapter, I now propose examining other aspects of reward systems, focusing on what the effects of having multiple reward systems might be on *task completion* and the *player experience*. Related to this, I wish to examine how different player audiences respond to multiple reward systems, and whether or not player audience has any effect on *task completion* and the *player experience*.

5.1.1 Multiple Rewards in Human Computation Games

Traditionally, most human computation games have adopted simple reward systems with mechanics that align with the collaborative and social nature of the human computation process. Point-based scoring systems are generally the most common form of feedback to players. Such systems are typically straightforward to implement (as they typically do not require additional aesthetic or artist-driven assets) and also provide a form of direct, easily-quantifiable feedback to players. Additionally, point-based scores that map cleanly to task completion metrics (e.g., scenarios where task output can be measured using some kind of objective or optimization function) may also give task providers ways to monitor and evaluate performance of task results in real-time. Point-based scoring systems also dominate the recent survey work in HCGs [13, 15], which examine rewards merely as forms of incentives available to the player.

As previously discussed in Section 3.3.2, human computation game research has explored reward types beyond point-based scoring systems. Goh *et al*. [77] compare utilizing

points, badges, and non-gamified statistics in a location-based, content sharing HCG. Gaston and Cooper [79] utilize three-star reward systems in the context of *Foldit*. The work in this chapter is similar to that of Goh *et al*. in that it investigates multiple types of rewards, but differs in certain types of rewards while also investigating the effect of different player audiences.

By contrast, many modern games designed for entertainment provide players with multiple kinds of rewards or feedback for interaction. These include genres such as massively-multiplayer-online games (MMOs) or roleplaying games (RPGs), which are designed to accommodate the diverse motivations of their monolithic player bases [92]. Within these games are myriad kinds of digital artifacts for players to collect—equipment, upgrades, customization options, badges, etc.—or explicit feedback—leaderboards, rankings, social recognition, etc.—which players might then share or compare amongst others. Tangentially, other games have eschewed these traditional, extrinsic (i.e., "gamified") rewards altogether. Such games are designed to appeal to players who are not interested in digital artifacts, but instead are often motivated intrinsically by the desire to interact with a specific experience. Examples include narrative games (which reward players with a compelling narrative or fictional outcomes) and walking simulators (which typically reward players with an emotional digital environment to explore).

Beyond the myriad of practical implementations of these many reward systems, both game design and game research have examined reward systems. Common approaches in general game design for understanding and designing effective rewards in games are driven by theories based around player motivations and incentives for play. Early approaches for designing and implementing rewards in games sought to understand how player motivations mapped to game mechanics. These investigations occurred in the context of game genres with complex game mechanics and diverse player bases, such as Bartle's player types for multi-user dungeon games (MUDs) [93] and Robin's Laws for tabletop roleplaying games [94]. While these results are based primarily on anecdotal studies of players

within these game contexts, they continue to be used as rule-of-thumb tools for digital game designers even today.

More recent research has been driven by empirical evaluation of player data and self-reported player preferences. These results include Choi's study of online games [95] and Yee's seminal work on motivational components based on a study of players in massively-multiplayer online games (MMOs) [92]. Other efforts have focused on creating player typologies based on player preferences. Brain Hex is a neurobiologically-inspired typology that classifies players into one of seven different types [96]. Similarly, the Hexad framework utilizes an empirically-evaluated survey to classify players into one of six user types derived from player motivations [97]. While these efforts are backed by data from actual players (as opposed to just anecdotes from game development), the development of these frameworks and typologies was conducted in the context of entertainment-oriented games, which do not capture motivational aspects around the secondary goals seen in serious games (e.g., motivation to learn in educational games, motivation to participate in the crowdsourcing process for human computation games).

Related research has explored models for player motivation and engagement, which incorporate psychological theories, such as self-determination theory [98]. A comprehensive overview of motivational theory as it applies to gamification and serious games can be found in the work of Richter *et al*. [99]. Richter *et al*. 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*. [100] explore motivations for participating in mobile content-sharing using a model of player gratification. Similarly, in their analysis of *Foldit* [4], Cooper *et al*. report on the results of a survey asking a subset of users

about motivations for playing the game. Their responses were categorized based on Yee's motivational components [92], amended with an additional "purpose" category to capture intrinsic motivations for participation (i.e., assisting with scientific discovery). More recently, Lessel *et al*. [10] ran a study looking at the ability to turn off gamified elements in an image-tagging task. They found that players who were given the ability to do so (and who did turn off those elements, as a possible indicator that they were not motivated by them) were more likely to do more tasks. Finally, similar investigations have been undertaken in analogous serious game domains such as educational games [101], which also make adjustments to existing theories to accommodate for intrinsic motivations beyond those driven by gameplay.

More broadly, crowdsourcing research, specifically in the context of paid crowdsourcing platforms, has also examined the effects of motivation on worker performance at tasks. In these scenarios, extrinsic motivation is captured by financial compensation for the work, whereas intrinsic motivation workers may have for the task is captured by workers' self-motivation or self-satisfaction. Existing research demonstrates that monetary reward may undermine the effects of intrinsic motivation in crowdsourced workers [75] and that while increasing the amount of financial compensation may yield more results, these results are not necessarily of a higher quality [11].

Studies have also examined the interchangeability between paid crowdsourcing platforms and human computation games [102, 103]. The results suggest that the quality of the completed work between the two interfaces is interchangeable, however Sabou *et al*. [103] remark that maintaining player motivation in HCGs may be more difficult than that of financially-compensated crowdsourcing platforms. This suggests that motivational findings in the context of financially-compensated crowdsourcing platforms may not translate directly to HCGs (i.e., when such results are examined over a longer period of time). As a result, it is unclear whether how, if so, and to what extent, rewards in HCGs compare directly with or map to financial compensation.

Altogether, there is a breadth of work looking at rewards in games, how these rewards correspond to player motivations and types, and finally how motivation affects crowdsourcing in general (and how this crowdsourcing research may generalize from paid crowdsourcing platforms to human computation games). Beyond the few explorations of different reward systems, it is not clear how applicable any of this preexisting work is when translated to HCGs. A first step would be to explore novel or different reward systems, which comes with the promise of broadening the potential audience of players (and would permit studying motivations and player types), but also comes with potential tradeoffs in the time and effort necessary to implement these systems. But before one might even start to pose the question of how many or which reward systems to put into an HCG, it is necessary to address the question of whether or not having multiple reward systems will even matter.

5.1.2 Player Audiences

Traditionally, human computation games have paid little attention to the audiences of players they recruit. Research efforts in human computation games generally rely on word-of-mouth marketing or traditional human subject study pools to find players to interact and play their games. These practices can be fairly reliable, as many human computation tasks require commonsense or intrinsic knowledge that an average game player can be expected to possess. Alternatively, the task can be broken down into smaller sub-problems that support many kinds of players specialized in or able to be trained on various parts of the task [4, 9].

By contrast, entertainment-oriented games are no strangers to understanding their players. Players are drawn to particular games based on game elements which appeal to their motivations for play. Some games may choose to target niche audiences, whereas other games may desire to appeal to the largest audience possible. In the latter case, these games often implement myriad kinds of rewards and feedback systems in order to support diverse player audiences, wherein there are players who may not always be motivated the same

kinds of rewards, yet can play the same game by interacting with the only systems they prefer. Unsurprisingly, much of the foundation of motivational research in games is built on player experiences from these games (e.g., MMORPGs [92]).

Player audience is yet another factor that could potentially affect *task completion* and *player experience* within the same HCG. However, very little understanding or study has been dedicated to player audience in the context of HCGs. However, the importance of understanding player audience cannot be understated. In order to remain effective at solving tasks, human computation games must be able to retain prior players and attract new players, often in competition with the ever-growing list of games that continue to be released year after year. Moreover, assumptions and findings about successful game elements may not hold if HCGs wish to attract new or specific audiences. For example, an HCG asking for language translations may wish to attract a player audience from particular locales; in order to be truly effective, its developers need to know what kinds of game elements players in that locale prefer. Ideally, they might even know how these elements affect the *task completion* and *player experience*. But with next to little knowledge on player audiences for HCGs, such questions cannot be answered.

5.1.3 Summary

The work in this chapter is motivated by the following two questions: First, how does *randomly* distributing a reward versus giving players a *choice* of reward affect *task completion* and *player experience*? Second, do different player audiences have noticeable differences on *task completion* and the *player experience*?

In this chapter, I describe a game, *Café Flour Sack*, which was developed to explore hypotheses around multiple reward systems in human computation games. *Café Flour Sack* provides players with four different reward systems to interact with: leaderboards, customizable avatars, unlockable narratives, and (non-gamified) progress tracking. I then describe a human-subjects study using *Café Flour Sack* that compares two different types

of reward distribution across two different player audiences. I then summarize the results of this study. I also describe a followup study, again using *Café Flour Sack*, that combines a play session with semi-structured interviews to better understand sentiment around rewards and motivations in the context of HCGs. I then discuss several design considerations around reward systems, reward distribution, and player audience. Throughout this chapter, I continue to use the language of the mechanics framework and follow the experimental methodology for testing game mechanics described in Chapter 3.

This chapter consists of four parts:

1. A description of *Café Flour Sack*, a game developed to examine hypotheses around multiple reward systems.

2. A human-subjects study using *Café Flour Sack* and its results.

3. A summary of eleven semi-structured interviews discussing multiple reward systems in the context of *Café Flour Sack*.

4. Three design implications drawn from the results of the study.

For reference, the peer-reviewed version of (most of) this work was published as a full conference paper at the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY) in 2016 [78].

5.2 Expanding on Rewards in Human Computation Games

Beyond point systems and leaderboards, we know very little about how other types of reward systems behave in human computation games. However, it is clear—based on the myriad of different player motivations and types—that not all players are necessarily motivated by point systems and leaderboards. These players are motivated by other extrinsic reward types or more immersive reasons which are not always captured by the commonly-used numerical point systems (and the occasional, accompanying online leaderboards).

The diversity of reward and feedback systems in modern, entertainment-oriented games, demonstrates many attractive and potential alternatives, thus resulting in the followup question of how can these systems (such as customizable avatars or game narrative) be utilized in HCGs?

5.2.1 A Game with Multiple Reward Systems

To explore this question, I built a human computation game with multiple reward systems. This game—called *Café Flour Sack*—is a cooking-themed HCG that assigns players the culinary-commonsense-knowledge task of pairing food ingredients (e.g., "eggs," "sugar," and "cabbage") to recipes which are likely to contain those ingredients (e.g., "cake," "ice cream," or "stew"). *Café Flour Sack* contains four different reward systems (also referred to as reward *categories*) for players to interact with: global *leaderboards*, customizable virtual *avatars*, unlockable *narratives* and a global *progress tracker*.

Café Flour Sack's cooking task is an artificial task with a known solution set, which—per the methodology described in Chapter 3—enables evaluating the efficacy of the reward mechanics without the complications of needing to simultaneously solve a novel human computation problem. I chose ingredient-recipe classification as described above due to its similarity to other classification and commonsense-knowledge problems, not to mention its relative simplicity (i.e., players do not need actual culinary training, but merely knowledge of what ingredients could be used in classes of recipes). For this experiment, I 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 also belong to multiple recipes (e.g., "sugar" is an ingredient in both "cake" and "ice cream" but not in "meat stew"[1]).

Players complete tasks by accepting rounds of gameplay; each round consists of five

[1] Arguably, there exist meat stews in which sugar is an ingredient. Similarly, "trout" can in fact be made into "ice cream." I note that due to the binary nature of the classification problem, this answer set—given that it was not constructed by molecular gastronomists or modern chefs—arguably reflects whether or not an ingredient is *likely* to be used in a given recipe, as opposed to whether or not it is possible that an ingredient *can* be used in a given recipe. I would also argue that this is appropriate, given that the audiences involved in this study are very unlikely to make dishes such as trout ice cream.

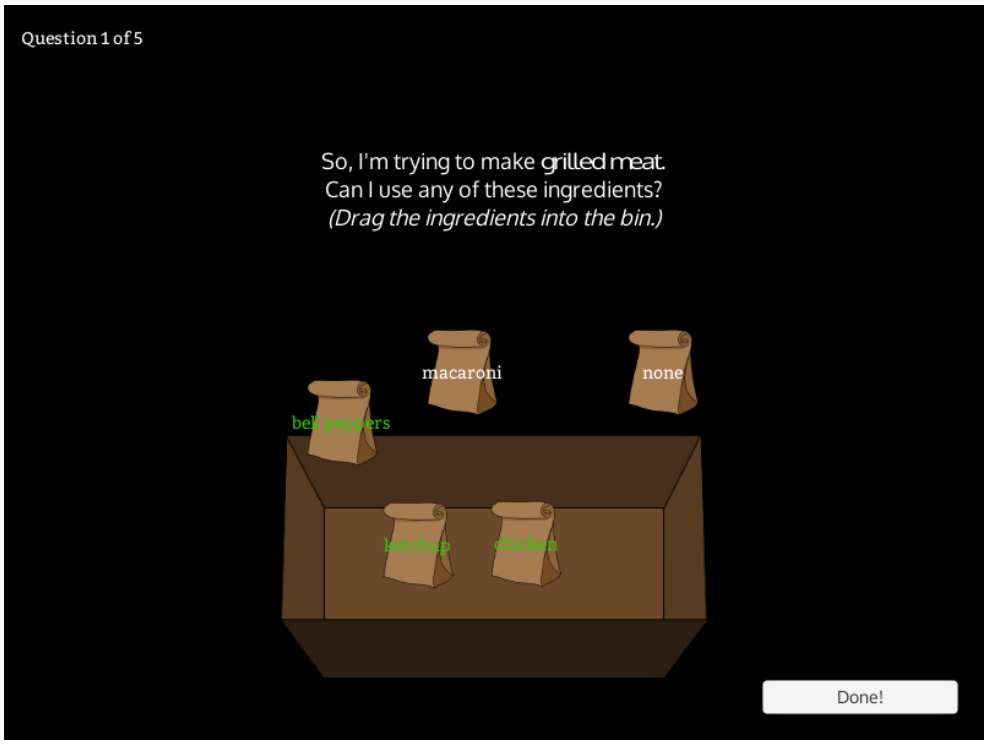

Figure 5.1: An example minigame from *Café Flour Sack*. Here, the player drags all ingredients that can be used in a corresponding recipe ("grilled meat") into a bin in the center of the screen.

small minigames. Each minigame presents the player with a recipe and four possible ingredients to select from (which either belong to the recipe or do not). Figure 5.1 shows an example of one such minigame, in which a player is presented with a recipe ("grilled meat"), four ingredients ("bell peppers," "macaroni," "ketchup," and "chicken"), and a fifth ingredient labeled "none" which the player can use if they believe none of the items belong to the recipe. [2] The player must then drag the correct ingredients into a bin in the center of the screen, then may hit the "Done!" button to proceed to the next minigame. At the end of the round, a player receives a reward from one of three (of four) possible reward systems. The amount reward is based on the particular system it is associated with, and consists of a base amount for completing the task and an additional amount based on how many minigames they completed successfully.

[2]The impetus for this fifth ingredient is to ensure that players actually attempt the task seriously by manually forcing them to commit to an answer if none of the provided ingredients are correct instead of allowing them to proceed without clicking on anything.

Café Flour Sack's four reward systems were selected to appeal to a broad audience of players and to thereby cover a variety of different motivations for play (e.g., such as those described by Yee [92]). *Leaderboards* are the most commonly-utilized system in most HCGs (popularized by the *ESP Game* [2] and customizable *avatar* systems are similar to other collectible item systems (e.g., badges [77]). Unlockable *narratives* were included as a novel reward system that would both address alternative player motivations and would not interfere with the other reward systems. Finally, the global *progress tracker* was added to accommodate players uninterested in extrinsic digital game rewards, but who still might be intrinsically by participating in the learning or crowdsourcing process (i.e., those akin to intrinsically-motivated players identified in *Foldit* [4]).

Each reward system, barring the global progress tracker, has its own "currency" which corresponds to the nature of the reward available in that system. Completing a task (i.e., a series of minigames further discussed in detail below) awards players currency for one of the reward systems. The exact amount of currency reward and its utility within that particular reward system again depend on the nature of the system. One commonality between all systems is that the currency units are referred to as "points" (as opposed to other synonyms for monetary units), but are prepended with the respective system name (e.g., points in the unlockable narrative are referred to as "narrative points").

The user interfaces for the four reward systems correspond to four screens in the game and are shown in Figure 5.2. I now describe these systems in detail as follows:

Global Leaderboards

In the global leaderboards, "leaderboard" currency is automatically applied upon acquisition to the player's total score on the leaderboard. In addition to a numerical score (i.e., the number of leaderboard currency points), a player also is given a "rank" which corresponds to successfully reaching a score threshold. There are five "ranks" a player can attain on the leaderboards, with "base" being the lowest and "platinum" being the highest.

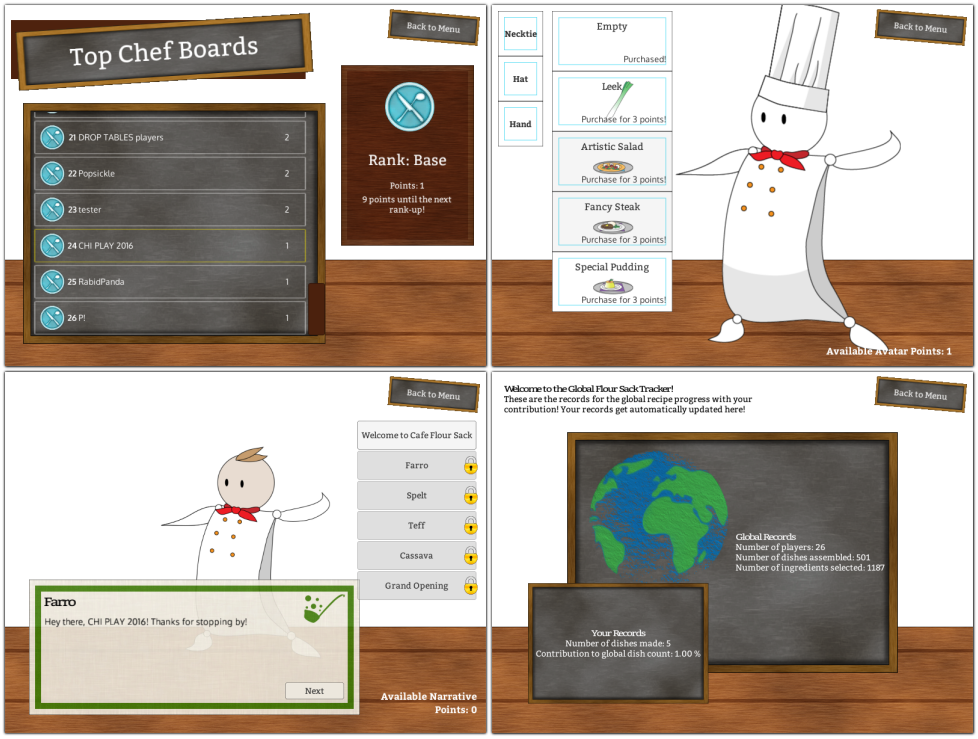

Figure 5.2: The four reward systems in *Café Flour Sack*. Starting clockwise from the upper-left: the global leaderboards, the customizable avatar, the progress tracker, and the unlockable narratives.

Figure 5.2 shows the leaderboard screen in the upper left corner. The user interface consists of a sorted list of all player entries on the leaderboard. The list is sorted based on player score, from highest to lowest. Each entry of the leaderboard shows—in order from left to right—a player's rank (denoted with a medal consisting of a chef's knife and ladle in a rank-appropriate color), followed by their numerical position relative to other players, the player's name, and finally the player's numerical leaderboard score. For example, in Figure 5.2, the player named "CHI PLAY 2016" (an homage to the conference in which this work was published) is at the "base" rank and in the 24th position relative to other players with a numerical score of 1. Finally, the panel on the right side of the screen clarifies and communicates the player's rank, as well as the number of points the player must acquire before moving up to the next rank.

All players are added to the leaderboards by default, however players who do not re-

ceive leaderboard currency (or choose not to) remain at the default score of 1 (and therefore at "base" rank).

Customizable Avatars

In the customizable avatar system, "avatar" currency is given to the player in the form of points that can be spent on customizing a digital avatar representing the player. Possible customization options for the player's avatar (a chef made out of a flour sack) include various chef-themed clothing items and culinary objects.

Figure 5.2 shows the avatar screen in the upper right corner The user interface consists of a digital avatar on the right and an expanding menu of customizable items on the left. The menu of items contains buttons for each item category; within each category is a list of available items (again as buttons). Each item button indicates its name, visual appearance, and cost. Clicking on an item (while having sufficient points) will allow the player to purchase that item; if a player already owns that item, clicking the button will instead equip (i.e., visually place) the item on the character. Finally, the number of available points to spend is listed in the bottom corner. As an example, in Figure 5.2, the player is currently viewing the "Hand" category of items wherein items such as "Leek" and "Special Pudding" are available for purchase; currently the player has only 1 point available, rendering them unable to purchase any of the 3-point items currently shown.

All players are equipped with a default chef's hat at the game's start, but may purchase more items if they choose to pursue rewards in this system.

Unlockable Narratives

In the unlockable narrative system, "narrative" currency is used to unlock short stories set in the culinary universe of the game. These stories are presented in a linear progression; the player is required to unlock the preceding stories before a subsequent story may be unlocked. The content of the stories plays out as conversation scenes between the player

and the in-universe characters of a restaurant (the titular "Café Flour Sack"); this interface is similar to that seen in narrative game genres such as visual novels.

Figure 5.2 shows the narrative screen in the bottom left corner. The list of available stories is shown on the right; stories annotated with lock icons indicate these have not yet been unlocked. Players may unlock a story by clicking on the respective story button, provided they have enough narrative points (as indicated by the label in the bottom right). Additionally, players may also replay through previous stories by simply clicking the respective story button again. On the left, a text box displays dialogue with the given character image (rendered behind the text box). In Figure 5.2, the player is in the middle of a conversation with a character named Farro as part of the "Welcome to Cafe Flour Sack" story.

All players are given enough narrative points at the beginning of the game in order to unlock the first story.

Global Progress Tracker

In the global progress tracker, currency is unavailable because the reward system, by design, does not give the player any form of virtual reward for completing tasks. Similarly, the global tracker cannot be selected as a reward option for completing tasks, as it acts as an intrinsic reward for players who may not derive enjoyment or motivation from the other extrinsic, in-game rewards. Instead, the tracker allows players to view statistics showing their overall contribution to the tasks being completed by all players in the game.

Figure 5.2 shows the global tracker screen in the bottom right corner. The panels on the screen show both the player's individual records as well as the global totals of all player records. In Figure 5.2, the player has completed 5 tasks (which resolves to the 1% of the current number of tasks completed). The global records on the adjacent panel reveal that 26 players have completed 501 tasks, utilizing 1187 ingredients.

5.2.2 An Experiment with Multiple Reward Systems

Combined with the methodology proposed in Chapter 3, *Café Flour Sack* and its multiple reward systems provides an environment to explore questions around reward systems in human computation games. A starting question for a game with multiple reward systems might be to ask what the most preferred reward system is. This particular question was previously explored by Goh *et al*. [77], who tested three versions of an HCG each with a different reward systems: leaderboards, collectible badges, and a non-gamified control. The authors found the leaderboard and badge versions yielded the highest (player-perceived) task metrics (accuracy and completeness) and enjoyment metrics, with both significantly outperforming the gamified control. Between the leaderboard and badge versions, no significant differences in metrics were detected (save for greater cognitive enjoyment [3] in the badge version). Overall, these results suggest that having any kind of reward is more beneficial than no rewards, and that these two different reward systems perform similarly with respect to *task completion* and *player experience* metrics. However (and as Goh *et al*. point out in their conclusions and limitations), it is insufficient to draw conclusions from the results of one game, especially when rewards and their content are specific to the particular game—and thus the particular task—in which these systems are implemented. For example, leaderboards in a singleplayer game are not necessarily comparable to leaderboards in a multiplayer game, despite similar presentations and player interactions. Beyond testing multiple reward systems across multiple HCGs, it is difficult to draw conclusions about a single reward system when so much of its content is contextualized by the task and the HCG solving it.

I instead propose testing other aspects of reward systems in the context of having multiple reward systems. For example, does having multiple reward systems have an impact on *task completion* and *player experience*? On the one hand, multiple reward systems—and

[3] By Goh *et al*.'s definition, "cognitive enjoyment" refers to measures "to which the user perceives favorable thoughts and beliefs about HCGs such as being worthy, effective, and interesting." [77]

the ability to let players interact with them—might not only improve the *player experience* but also potentially yield better *task completion* metrics. On the other hand, implementing such systems is costly (in both time and effort) when it is unclear they may have no such effect, or worse, prove to be distractingly orthogonal to the completion of the task.

This question of choice—the ability to let players choose whether or not they wish to interact with rewards—is the condition I propose testing. To minimize the differences between game versions for such an experiment (i.e., avoiding a complicated permutation of different reward systems across many conditions), I propose that both versions of the game utilize the same set of reward systems and that the player's ability to choose which rewards they want is the experimental condition. I posit that randomly-assigned rewards (i.e., the control condition) is a proxy for how current HCGs behave. Specifically, most HCGs make no effort to tailor, let alone understand, what the most effective reward system is for a given player or player audience. If the game knows nothing about what rewards players prefer *and* supports multiple reward systems, it can, at best, assign these randomly. Given any player, a randomly-assigned reward may align with their motivations or preferences, whereas other times, it will not. Meanwhile, players given the ability to choose their rewards are more likely to interact primarily with the systems they find compelling.

Additionally, if having multiple reward systems is intended to broaden the potential audience of players, do different player audiences even have an effect on *task completion* and *player experience*? This question is particularly pertinent in the context of HCG research, which often uses either university student pools or outreach to professional crowdsourcing workers to test research hypotheses.

Finally, as a side effect of these particular questions and experimental setup, I also chose to investigate preferred reward systems using *Café Flour Sack*, although I caution that, in addition to the concerns identified above, these results are not fully representative of games with multiple reward systems. Games which contain multiple kinds of extrinsic and intrinsic feedback often intertwine these systems. For example, level-up mechanics (e.g., those

which reward the player with new abilities) require player engagement should the player wish to progress further in the game, thus making enjoyment of other progression systems, such as that of the game's narrative, dependent on player engagement with leveling. *Café Flour Sack's* reward systems are deliberately kept mechanically isolated from each other to avoid this exact scenario, as forcing a player to engage with a system they may or may not prefer in order to engage with another would complicate measuring player experience metrics between these reward systems (e.g., if what customizable items players could unlock in the *customizable avatars* somehow depended on one's position in the *global leaderboards*, players would be forced to participate in the leaderboards and might report greater dissatisfaction with the customizable avatar if they did not prefer leaderboards). [4]

Figure 5.3 shows the full comparative breakdown of the mechanics between the *random* and *choice* versions of the game, using the mechanics framework from Chapter 3. The diagram illustrates how following the *action* and the *verification* mechanics, the difference between the two versions is in the *feedback* mechanics of the game. In the *random* condition, the game automatically selects one of the three rewards for the player, whereas in the *choice* condition, the player makes a choice of which of the three rewards they wish to accept. Furthermore, the diagram shows the two additional conditions for this experiment: the two audiences of players receiving the game.

5.3 Methodology

Café Flour Sack was released as an online game.[5] Upon starting the game, participants were randomly assigned to one of two conditions—*random* or *choice*—which reflect the control and experimental conditions respectively. The choice of condition was determined by a call to the game's backend servers (ensuring that the experiment was conducted from

[4] Anecdotal feedback from a test pilot used to validate the Amazon Mechanical Turk platform, followup feedback, and the later interviews suggested players were, in fact, convinced they were playing against an actual player base and not an artificially-simulated one.

[5] While a version of *Café Flour Sack* remains online as of the writing of this dissertation, the particular versions utilized in this experiment are no longer available.

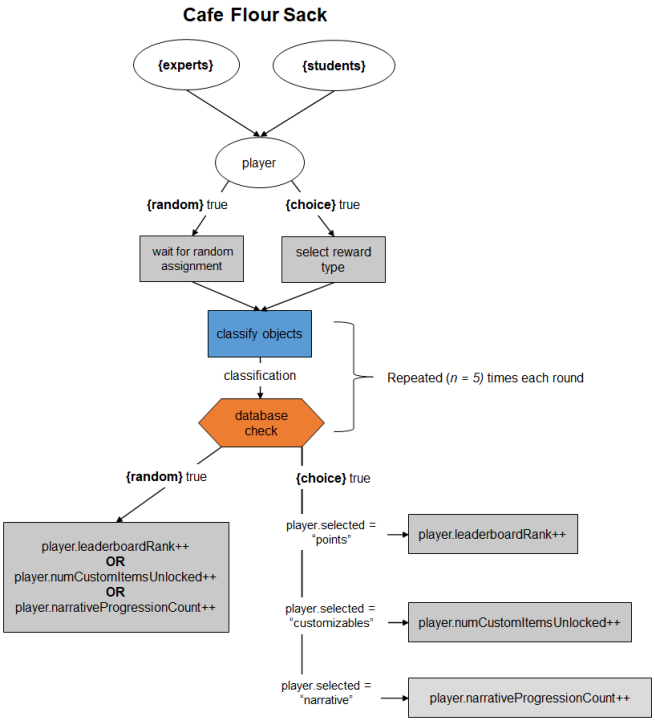

Figure 5.3: The breakdown of *Café Flour Sack's* mechanics and experimental conditions. At the top, two different player audiences constitute two conditions, while the ability to choose a reward category versus having it randomly assigned constitutes another two conditions. Experimental conditions are noted using boldfaced braces.

the same random source).

The condition changed how participants were assigned rewards from the three available reward categories: the global leaderboards, the customizable avatar, and the unlockable narratives. (The global progress tracker was excluded from this selection since it deliberately does not reward players for anything but participation, which is tracked automatically.) In the *random* condition, participants were automatically, randomly assigned a reward from one of the three available reward categories upon starting a round (i.e., series of five minigames). In the *choice* conditions, participants were allowed to manually selected one of the three reward 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 5.4. Participants in the *random* condition were shown a high-

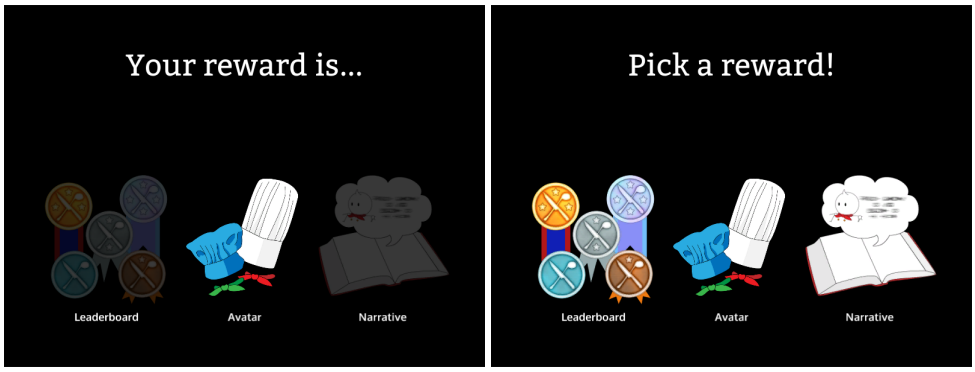

Figure 5.4: Screenshots of the reward selection screen between the two versions of *Café Flour Sack*. On the left, the *random* version selects a reward category (in this case, the avatar category) automatically. On the right, the *choice* version allows the participant to click on their preferred category.

lighted icon of the reward selected for them, whereas participants in the *choice* condition could click on one of the three icons to select their reward of choice.

Gameplay began with a short tutorial round of five minigames, after which participants were given currency in all three possible reward categories (minus the global tracker, which has no currency). Participants were then instructed to view each of the reward menus in order to view or spend those currencies before further progressing in the game, which both informed participants of and instructed them how to spend currency for each respective reward system.

Participants were then asked to complete as many tasks as they desired for the remaining duration of the experiment. They were also informed that they could interact with as many or as few of the reward systems as they desired. From the start of the tutorial round through the end of the experiment, participants were given a total of twenty minutes (but were not explicitly informed how long this duration was, although participants were told in the consent form that the entire experience would take "around half an hour"). Finally, participants were asked to fill out a post-game survey at the end of the experiment. Throughout gameplay, the game continuously logged telemetry (event) data on player completion of tasks and interactions with the various reward systems.

I recruited participants from two populations. The first population consisted of crowd-

sourcing professionals (workers) through Amazon Mechanical Turk. Previous work has successfully explored the use of paid crowdsourcing platforms—Amazon Mechanical Turk among them—and this was taken into consideration when setting up this experiment. *Café Flour Sack* was made available as a task (HIT) on the Mechanical Turk platform, where workers were shown an external link to the game and then returned with code retrieved at the end of the post-game survey. Inputting this code into the Mechanical Turk task would then compensate workers once the code was validated against the telemetry data.[6] The second population consisted of university students recruited through an undergraduate computer science class. These students were compensated with course credit for writing a report on their experience participating in the experiment.[7]

One long-term goal of human computation games is to broaden their accessibility, so I deliberately chose to evaluate this work not only across two different experimental conditions, but also across two different audiences—a first for HCGs. On the one hand, the Amazon Mechanical Turk workers represent a more demographically diverse group of players who are highly-skilled experts at crowdsourcing work, but who typically perform such work through a monetarily-compensated online interface. These workers are therefore not necessarily experts playing at human computation games, nor can it be expected that they might be familiar with games at all. On the other hand, the student population (particularly one in an engineering field of study) represents an audience likely to be familiar with digital games, but not necessarily crowdsourcing work.

Because this experiment looks at understanding player experience and engagement in the context of rewards, I took some additional steps to account for the fact that players might have extrinsic motivations for completing the task quickly. First, as previously men-

[6]Workers were paid $7 (USD) for completing the experiment; I validated with other colleagues that this was appropriate considering the then-commonly-utilized $10 (USD) minimum wage for an hour's work in the United States).

[7]We received feedback post-proceedings publication that a more considerate approach would be to compensate students both monetarily for participation and then with (lesser amounts of) course credit for an external report. I wish to acknowledge this concern as well as to reiterate that future experiments implemented this better compensation model or utilized monetary compensation only.

tioned, participants were required to play the game for at least twenty minutes, during which they were allowed to freely allocate their time between completing tasks (thus yielding additional reward currency as a byproduct of task completion) and interacting with the reward systems (thus viewing or spending that currency on the various reward systems). This forced duration of play was implemented to ensure that participants would not be incentivized to rush through the experiment as quickly as possible. Without this, completing the experiment in the shortest (optimal, particularly for Amazon Mechanical Turk workers) time would be to avoid interaction with the reward systems at all. Similarly, participants were not required to complete a certain number of tasks.

Second, I introduced a button in the game's menu, which I refer to as the "boredom" button. Participants 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 this button was optional and did not have any impact on whether or not (Amazon Mechanical Turk) participants were compensated. The inclusion of this button was designed to act as a proxy of retention. Measuring retention in this experiment would otherwise be untenable, given that a fixed play duration was enforced and that without it, players would optimally strive to complete the experiment as quickly as possible.

Finally, I wished to ensure that participants who completed the study later would not be biased by the presence and progression of players whose results were already a part of the reward systems with visible social elements—namely the leaderboards and the global progress tracker. In order to preserve the social elements of the game while maintaining consistency across all participants, real-time adjustments to both the leaderboards and progress tracker were simulated using a fake set of players and results. After each round of the game, these fake players were updated with artificially-simulated progress, which also included the occasional addition of new fake players (e.g., as new scores that would appear on the leaderboards, etc.) to further create the perception that other players were

simultaneously playing the game.

5.4 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. I report on results from 78 participants who took part in the study. When divided by condition, 40 participants were placed in the *random* condition and 38 were placed in the *choice* condition. When divided by audience, 39 participants were from Amazon Mechanical Turk (randomly selected from a larger population of 59 workers) and 39 participants were students.

When considering population demographics, 24 participants self-reported as female and 54 participants self-reported as male. Most players reported themselves as 18-40 years old. Additionally, around 80% of participants reported prior experience playing games; however only around 20% of participants reported any prior experience playing human computation games.

The evaluation focuses on both the results of the *player experience* and the *task completion*. Investigation focuses on differences between the two conditions of *random* (the control) and *choice* (the experimental). I also investigate differences between the two populations of the player audience—Amazon Mechanical Turk workers and university students—while accounting for interaction effects with the experimental conditions. The majority of the dependent variables had nonparametric distributions. To measure the differences and interactions between the conditions, I used two-way ANOVAs with aligned rank transforms [104] to account for the nonparametric nature of the data unless otherwise stated. The subsequent sections report on the results; the discussion of their implications follows in Section 5.6.

Table 5.1: Counts of participants' favorite rewards across both experimental condition and participant audience in the *Café Flour Sack* study.

		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

5.4.1 Subjective Metrics—Player Experience

The data contributing to the evaluation of the *player experience* consist of telemetry events detailing player interactions with the game, combined with player responses to questions on the post-game survey. Of interest is understanding how participants engaged with the reward systems, as well as why they may have become disengaged with these systems. I first report on participants' survey responses regarding their favorite and least favorite reward systems in *Café Flour Sack*, an question of whether or not players perceived they had a choice of reward systems. Next, I report on participants' interaction time within the each of the reward systems. Finally, I report on participants' interaction with the boredom button in order to understand why they would have disengaged with the game—and if our reward systems were responsible for that disengagement.

Reward Preference

By design, *Café Flour Sack* provides players with four different reward systems—a natural investigation is to determine how participants responded to each of the different reward systems made available to them. In the post-game survey, participants were asked to provide their favorite and least reward systems in *Café Flour Sack*. For participants' favorite reward systems, 39 participants selected the leaderboards, 19 participants selected the unlockable narratives, 18 participants selected the customizable avatar, and 2 participants selected the progress tracker. Table 5.1 shows the exact breakdown of participants' favorite rewards across both the experimental condition and participant audiences.

Table 5.2: Counts of participants' least favorite rewards across both experimental condition and participant audience in the *Café Flour Sack* study.

		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

Meanwhile, when it came to participants' least favorite reward systems, 35 participants selected the unlockable narratives, 18 participants selected the customizable avatar, 14 participants selected the progress tracker, and 11 participants selected the leaderboards. Table 5.2 shows the exact breakdown of participants' least favorite rewards across both the experimental condition and participant audiences.

There were no differences or effects on task performance based on participants' favorite or least favorite reward systems.

Perception of Choice

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

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

Table 5.3: Mean duration (in seconds) of time spent in a single view for all four reward systems across both participant audience type and experimental condition in the *Café Flour Sack* study.

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

Duration of Play

As previously described, interaction within the game was limited to 20 minutes. It was assumed that Amazon Mechanical Turk participants were already incentivized to participate for financial reasons (and would therefore complete tasks as quickly, but adequately as possible). Furthermore, Amazon Mechanical Turk also imposes a time limit for submitting task results, so these participants would have been highly unlikely to continue playing under this additional time. Under these limitations, it is insufficient to look at total duration of play as an indication of engagement or retention.

Instead, I examine where and how players spent their time during those 20 minutes of play. Specifically, the metrics of interest are how long players spent in each of the different reward systems. Each reward system had its own dedicated screen and the game recorded how long players spent in these screens. Direct comparison between these screens is immediately useful as some screens, such as the leaderboards and the progress tracker, show very short durations. In these cases, short durations are expected as participant interaction was limited to viewing information such as leaderboard rank or task progress, rather than engaging with the reward interfaces. By comparison, the unlockable narratives required

participants to read and actively click through character dialogues. Table 5.3 shows the mean time spent in each view of the corresponding reward menu, broken down by experimental condition and participant audience.

In the leaderboards, both experimental condition and player audience has a significant main effect on the duration of interaction. Participants in the *random* condition spent longer in the leaderboards than participants in the *choice* condition ($F = 7.319, p < 0.01$) with a mean time of 11.904 seconds versus 8.868 seconds. No interaction effects were observed.

No significant differences in duration of interaction were observed between experimental condition and participant audience for the three remaining reward systems: the customizable avatar, the unlockable narratives, and the global progress tracker.

Boredom

62 of the 78 participants in the study pressed boredom button. Of these participants, 32 were in the *random* condition (resulting in an 80% press rate) and 30 were in the *choice* condition (resulting in a 79% press rate). 34 of these participants were from Amazon Mechanical Turk whereas 28 were student participants. When observing the times from the start of the game to the point at which the boredom button was pressed, no significant differences were detected between the experimental conditions and the participant audience.

As part of the post-game survey, participants were asked to clarify why they had pressed the boredom button (if they had chosen to do so). These answers consisted of free-form sentences; participants were not limited to a single reason. (As a result, the following counts are not exclusive of each other.) Overall, 26 participants (42% of participants) described their primary reason for pressing the boredom button as due to the repetitive nature of the task (i.e., lack of variety in the tasks or tasks that were too similar). 10 participants described their primary 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 participants), general confusion or unfamiliarity with certain ingredients (4 participants), a lack of challenge (3

participants), and a lack of purpose and/or learning (3 participants).

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

5.4.2 Objective Metrics—Task Completion

To evaluate the *task completion*, I highlight and focus on three metrics: the correctness of the task answers, the number of tasks completed, and the timing of task completion. These metrics reflect the typical considerations of task providers. For an actual human computation task, different metrics might be prioritized over others depending on the task requirements; here however, I present 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 a 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 ratio of 0.75) or more of its ingredients belonged the given recipe. For example, the "ice cream" recipe would be considered correct if "milk," "eggs," and "strawberries" were correctly selected (and if "onions" was not).

The results show that both experimental condition and participant audience had significant effects on answer correctness. Participants in the *choice* condition had higher mean scores than players in the *random* condition: 0.724 versus 0.696 ($F = 9.474, p < 0.01$).

Table 5.4: Mean task scores split by experimental condition, first broken down into separate player audiences and then shown in total, in the *Café Flour Sack* study.

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

Amazon Mechanical Turk participants had higher mean scores than student players: 0.724 versus 0.692 ($F = 9.072, p < 0.01$).

The experimental condition $\times$ participant audience interaction was significant ($F = 28.648, p < 0.001$). Table 5.4 shows the mean task scores split across experimental condition and participant audience. Amazon Mechanical Turk participants in the *random* condition demonstrate the highest mean scores (0.7254) with student players in the *choice* condition performing closely behind (0.7245). Meanwhile, student players in the *random* condition demonstrate the lowest mean scores (0.670).

Number of Completed Tasks

I also looked at the number of tasks completed per participant across both experimental condition and player audience. These observations are further broken down into three categories: the *total* number of tasks completed, the number of *correct* tasks completed, and the number of *incorrect* tasks completed. On average, Amazon Mechanical Turk participants provided significantly more total answers (82.308 answers) compared to student participants (70.128 answers) ($F = 5.083, p < 0.05$). Additionally, when looking only at correct answers, Amazon Mechanical Turk participants provided significantly more correct answers (57.410 answers) compared to student participants (44.590 answers) ($F = 5.083, p < 0.05$). No other significant effects were observed across experimental condition and participant audience.

Table 5.5: 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, in the *Café Flour Sack* study.

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

Timing of Completed Tasks

The final task completion metric is the time (in seconds) it took players to complete tasks. Similarly to the observations of number of tasks completed, these results are broken down by the timing of *total* tasks, *correct* tasks, and *incorrect* tasks.

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

There were also interaction effects. When accounting for experiment condition $\times$ participant audience interaction across all tasks, there was a significant effect ($F = 40.875, p < 0.001$). Table 5.5 shows the mean task completion times for all tasks split across experimental condition and participant audience. Overall, Amazon Mechanical Turk participants in the *random* condition demonstrated the fastest mean times (8.382 seconds) and are slightly slower in the *choice* condition (9.128 seconds). Conversely, these results are flipped across conditions for student participants, who demonstrated faster mean times in the *choice* condition (12.753) compared with the overall slowest mean times in in the *random* condition (14.229 seconds).

Next, when looking only at the mean times it took participants to complete tasks cor-

rectly, it was once again the case that both experimental condition and participant audience had significant effects (though no interaction effects were observed). Participants in the *choice* condition were faster at completing tasks correctly than participants in the *random* condition, 9.773 seconds versus 10.820 seconds ($F = 5.809, p < 0.05$). Meanwhile, Amazon Mechanical Turk participants were faster at completing tasks correctly than student participants, 8.348 seconds versus 12.780 seconds ($F = 190.930, p < 0.001$).

Similarly, when looking only at the mean times it took participants to complete tasks incorrectly, both experimental condition and participant audience had significant effects. Participants in the *choice* condition were slightly faster completing tasks incorrectly than players in the *random* condition, 12.262 seconds versus 12.726 ($F = 10.22, p < 0.01$). Again Amazon Mechanical Turk participants were faster at completing tasks (incorrectly) compared to student participants, 9.802 seconds versus 15.174 seconds ($F = 21.868, p < 0.001$). Significant effects for experimental condition $\times$ participant audience interaction were also observed ($F = 43.596, p < 0.001$). Amazon Mechanical Turk participants were the fastest overall at 9.371 and 10.167 seconds in seconds in the *random* and the *choice* conditions respectively. Student participants were slower overall at 14.991 and 15.531 seconds in the *random* and *choice* conditions respectively.

In summary, participants in the *choice* condition had faster times for overall (all) task completion than participants in the *random* condition. Additionally, Amazon Mechanical Turk participants were significantly faster at completing tasks than student participants. This particular finding was observed not just across all tasks, but also for tasks answered correctly and tasks answered incorrectly. When simultaneously considering both experimental condition and participant audience, Amazon Mechanical Turk participants in the *random* condition were the fastest at completing total tasks, while student participants in the *random* condition were the slowest. Similarly for incorrectly-answered tasks, Amazon Mechanical Turk players in the *random* condition were the fastest, while student participants in the *choice* condition were the slowest.

5.5 Interviews

The study with *Café Flour Sack* demonstrated how giving participants the choice of reward (or not) and participant audiences affected *task completion* and *player experience*.

However, *Café Flour Sack* was designed specifically to test multiple reward systems in a human computation game as part of an experimental setting. To better understand how the results of that study might compare or generalize to other HCGs, not to mention reward systems in other games, I wanted to untangle *why* and *how* participants might have responded as they did. After all, there still remain many outstanding questions about how reward systems function in HCGs. For example, how do rewards in HCGs compare to other rewards in entertainment-oriented games? Additionally, what might motivate players to play HCGs and would particular kinds of rewards able to motivate players who might normally not be interested? Furthermore, what kinds of rewards in addition to the rewards tested in *Café Flour Sack* would players even want to see in HCGs? These—and many other questions—remained unanswered from both the quantitative results of prior work and the results described above from the *Café Flour Sack*study.

So to dig deeper into some of these questions, I ran an exploratory, two-part followup study using *Café Flour Sack*. The first part of the study consisted of a play session, in which participants played through *Café Flour Sack*, going through a similar flow as the prior study. The second part of the study consisted of an optional, semi-structured interview following the play session. This interview covered questions spanning four topic areas: rewards in games, HCGs, rewards in *Café Flour Sack*, and future rewards in HCGs. The full interview script can be found in Appendix A.

Participants were recruited via word of mouth and email, and were recruited from a similar student population as that of the student audience from the previous experiment. Given the delay between the previous experiment and this study, none of the original participants from that audience were available to participate (nor would an interview based

on a recollection of an experiment taken several months prior be likely to yield clear recollections). Study participation was forcibly limited to the student audience as Amazon Mechanical Turk's Terms of Service prohibit direct contact between task workers and task providers.[8]

The followup study was conducted in the lounge area of a research lab located on a university campus. For the first part of the study, participants were given a computer (or allowed to use their own if preferred) and given the link to the online version of *Café Flour Sack*. The version of *Café Flour Sack* used for this followup study was the *choice* condition from the previous study, in order to allow participants to explore the different reward systems as freely as possible. This version also differed by providing a new pre-game survey (not present in the prior study) and a modified post-game survey following gameplay.

Participants were asked to play for fifteen minutes; free-form feedback from multiple participants in the previous experiment noted that the original study duration felt too long. Participants were not observed during this portion of the study in order to provide a more natural play setting. After around twenty total minutes of interaction (the fifteen minutes of play in combination with pre-game and post-game survey completion), the game would direct participants to conclude.

For the second part of the study, participants were informed that the interview was optional. If they chose to be interviewed, the interview took place in the lounge area of the lab and were recorded via a hand-held audio recording device. The interview consisted of a minimum of ten questions (and as many as sixteen depending on their responses to certain questions). After the interview, participants were allowed provide additional feedback or ask questions about the study.

Regarding compensation, all participants were compensated with gift cards based on

[8]I am aware that researchers and task providers do solicit audio recordings from Amazon Mechanical Turk workers. However, an interview is well beyond the purview of this (arguably unscrupulous) kind of data collection (i.e., which is intended for training data acquisition rather than qualitative analysis).

whether or not they participated in both parts of the study. Participants would receive $5 of gift credit for taking the first part of the study and could receive an additional $5 of gift for the interview. All participants chose to be interviewed and thus received a total of $10 of gift credit for study participation.

5.5.1 Results

The study was conducted over the course of several weeks, during which this particular version of *Café Flour Sack* was made available online (although participants would only receive the link during the study). I report on results from 11 participants who responded and who took part in this followup study. Participant names have been anonymized and abbreviated (e.g., "Participant 1" will referred to as "P1" and so on). Demographic and initial game experience data was taken from the post-game survey in the first part of the study[9]. All other results came from transcriptions of the interviews.

Demographics and Game Experience

As previously mentioned, all participants came from a student population that was similar to the previous university student population. 10 of 11 participants reported as male; the remaining participant reported as female. All participants were between the ages of 18-30. 7 of 11 participants reported that they played games regularly (although all but one participant listed their favorite games, indicating that 10 of 11 participants had any experience playing games). Only one participant listed any experience with HCGs. Finally, participants reported their top-three favorite genres of games. Participants listed 15 separate genres[10] Of these, genres favorited by at least 3 participants included role-playing games—RPGs—(5 participants) and shooters/first-person shooters (3 participants).

[9]The previously-mentioned pre-game survey was ultimately not used.

[10]The full list of genres included: action, adventure, fighting, massively-multiplayer online first-person shooters (MMOFPS), massively-multiplayer online role-playing games (MMOs/MMORPGs), platformers, rhythm action games, real-time strategy (RTS), role-playing games (RPGs), shooters/first-person shooters (FPS), simulation games, sports, strategy, and turn-based strategy.

These participants represent a limited subset of the audiences from the prior experiment, particularly with regards to participant demographic. I acknowledge that this demographic distribution—primarily self-reported male and regular players of games—is far from ideal, particularly in comparison to the demographics from the previous study. While I do discuss the results of this work as part of the discussion in this chapter, I emphasize that at best, these results should be considered exploratory, anecdotal, and/or prompts for avenues for future work on reward systems in HCGs. Further followup on this kind of study would be compelling, in particular since at the time of writing, qualitative research of HCGs remains scarce and no qualitative studies beyond these interviews have been conducted on HCGs players in the context of reward systems.

5.5.2 Analysis

Participant interviews were recorded using handheld audio equipment during the study and then manually transcribed into text prior to analysis. From initial readings of the transcripts and several of the items captured in the survey data described above, my colleagues and I developed a set of 17 codes. These codes capture motivations (or inversely, inhibitions) for play, and range from abstract concepts (e.g., "challenge") to specific genres of games (e.g."RPGs") wherein some desired experience is reflected. For convenience, I classify these codes into three categories: concepts, rewards, and game genres.

The *concepts* codes describe broad aspects of game mechanics and/or the overall play experience. As the name suggests, they are not necessarily tied to specific game elements or mechanics, but instead reflect interactions that players desire to have (or to avoid) within games.

- Compensation

  Reward of some real or digital currency in response for completing in-game actions (e.g., quests in RPGs, tasks in HCGs, etc.). Often contextualized by fairness or value.

- Progression

The presence of changing interactions of play, often in the form of the addition of new or the escalation of existing game elements/experiences.

- Aesthetics

  Emphasis on a game's visual elements.

- Repetition

  Use of repeated elements in the game. Typically a source of frustration for players.

- System Interaction

  Whether or not certain mechanics interact with other systems in the game; in this study's context, primarily whether or not feedback mechanics interact with action mechanics both in HCGs and other games.

- Completionism

  Whether or not players feel compelled to experience all game content. Some players respond positively to this; others do not.

- Challenge

  The difficulty of interaction involved with the game's mechanics. Some players respond positively to this, while others do not.

The *rewards* codes describe explicit kinds of reward or feedback systems present in *Café Flour Sack* as well as other games. Like the previous codes, these codes—specifically the interactions players have with these systems—are cited as motivations for play. The first four codes correspond to the four reward systems available in *Café Flour Sack*.

- Leaderboards/Points/Ranking Up

  Systems or mechanics that give players a nominal ranking or comparison metric

against other players or in-game entities. [11]

- Customization

  Systems or mechanics that allow the player to personalize some aspect of their game, typically related to their in-game representation.

- Narrative

  Systems or mechanics that provide players with story elements and/or background information about the diegetic context of the game.

- Purpose

  Systems or mechanics that convey the purpose or alternative goal of the game (e.g., the task for HCGs, learning outcomes for education games, etc.).

- Collection

  Systems of mechanics that allow players to aggregate in-game currency or objects.

- Social

  Systems or mechanics that encourage the player to and facilitate with other players.

The *game genre* codes describe various genres brought up by participants in the study. Participants often referred to these genres to provide (mostly positive) comparisons or examples in response to interview questions. Originally, this set of codes contained all genres from the previous post-game survey demographics, but those which were not explicitly mentioned were removed.

- RPG (Role-Playing Games)

  Story-driven games in which players typically assume the role of a virtual charac-

[11] This item does not include "leveling up" or "level systems" which are covered under the "Progression" item in the previous category.

ter. Often contain diverse sets of game mechanics such as combat, navigation, and strategy.

- Shooter/FPS (First-Person Shooters)

  Action games, which often place the player in the role of a combatant and typically emphasize skilled hand-eye-coordination. Often presented from a first-person camera view.

- Strategy Games

  Games designed around problem-solving and logical reasoning through the simulation of realistic or real-world scenarios.

- MMO(RPG) (Massively-Multiplayer Online Role-Playing Games)

  A specific subset of RPGs in which the game environment simultaneously hosts many players connected via a network. Tends to encourage social or multiplayer interaction (more so than traditional RPGs).

Rewards in Games

First, participants were asked about their experiences and motivations with games in general. Participants were asked to describe which rewards they liked in games they played, followed by what motivated them to interact with these rewards. While least one participant touched on nearly every conceptual theme as a motivation for play, motivations for play were clustered around specific reward systems. Multiple participants (P3, P4, P5, P7, P8) cited leaderboards or ranking up as reward systems that they resonated with in games. Collection (P2, P4, P5, P10) and narrative (P6, P8, P9, P11) were also cited as motivators for play, although there was variation in the kinds of collectible objects and narrative elements. As examples of variation in collectable objects, P2 cited collectible "things" (using the *Pokémon* series' collectible monsters as a further example) whereas P5 described collectible badges. As examples of variation in narrative elements, P6 emphasized "lore

(...) story or background worldbuilding" whereas P9 emphasized narrative "themes" (using the game *Bioshock* as a specific example). Notably, many participants drew comparisons to their preferred genres of games such as role-playing games (P1, P6), shooters (P4, P5, P8, P11), and strategy games (P9), and often provided specific examples of games in these genres. [12]

Inversely, participants were then asked if certain rewards made them feel less motivated to play a game. Several common themes that participants highlighted negatively included challenge (P9, P10), customization (P4, P7), and social (P1, P11). Regarding challenge, P9 emphasized beating an easy game was not sufficient motivation for playing it, whereas P10 expressed dislike for games where there was too much challenge. Regarding customization, both P4 and P7 described customization systems as not providing enough value to play the game. Additionally, several participants (P2, P6, P11) cited that specific (to each participant) elements of MMORPGs made them less motivated to play those games.

Human Computation Games

To better understand the findings from the post-game survey (in which only one player reported any experience with HCGs), participants were then asked if they had previously played human computation games. All participants were given the clarification that "human computation games" was synonymous with other terms such as "citizen science games" or "crowdsourcing games" (to name a few).

Four participants (P1, P5, P6, P11) definitively stated that they were familiar with or had heard of human computation games, or similar games. Of these, three (P5, P6, P11) described having played games that would meet the definition of a human computation game (per the definition from Chapter 2). Participants did not express wide interest in playing these games nor did they choose to clarify why. [13] Only two participants (P1, P11)

[12]The diversity of games named in these examples was quite broad. Across all interviews, no two (or more) participants used the same game as an example, with the exception of *Call of Duty* (twice).

[13]I did not press this question further during the interviews to avoid coming across as confrontational to interviewees.

stated outright that they would be interested in playing an HCG in the future (with another two (P8, P9) suggesting "maybe"). Only one participant (P11) gave a detailed answer as to why they might be interested to try an HCG: to "try to see what issues people face and trying to solve this problem and like see how hard it is to do it and why don't you take a stab it yourself and in this process learn something."

Rewards in the Context of Café Flour Sack

Participants were then asked four questions related to the rewards in *Café Flour Sack*. First, they were asked how they thought the rewards in *Café Flour Sack* compared to rewards in other games. While some participants emphasized that the rewards were comparable to those seen in existing digital games, three participants (P1, P2, P6) expressed that the rewards in *Café Flour Sack* did not feel like rewards in other games because of a lack of system interaction (i.e., the reward systems were isolated and did not interact with the other mechanics in the game). These three participants expressed some desire to see their rewards feedback into the other parts of the game; for example, P2 mentioned (in the context of the customizable avatar) that "you couldn't ever see them outside of the avatar editor" and as a result, "(...) it was like sort of like not real, like I couldn't really use the things I was getting so that kind of distanced me from it."

Next, participants were asked what their favorite and least favorite reward systems in the game were. Participants expressed favorite and least favorite reward systems across all reward systems, with the exception of the global tracker. [14] Below, I detail the responses for each reward system.

Global Leaderboards Five participants cited the leaderboards as their favorite reward system in *Café Flour Sack*. Reasons for preference in the leaderboards included an interest in ranking up or seeing one's position relative to others, preferably on the top (P3, P4, P7,

[14]Regarding participants' favorite and least favorite reward systems, no participants considered the global tracker as their favorite or least favorite reward system, with the exception of one participant (P10) who listed the global tracker as their least favorite reward system because they "don't pay much attention" to it.

P10). Other participants highlighted the social aspects of the leaderboards (e.g., P7 stated "I'm thinking mostly the media and peers playing the game as well").

Conversely, one participant cited the leaderboard as their least favorite reward system in *Café Flour Sack*. This participant (P2) cited their disinterest in competition as the reason for not interacting with the leaderboards: "I didn't do anything with the leaderboards because I just don't really like competition that much."

Customizable Avatars Two participants cited the customizable avatar as their favorite reward system in *Café Flour Sack*. One participant (P2) cited their enjoyment of collection and character customization, while the other participant (P5) described the customizable items as the most rewarding because "the items were something physical and kind of fun in the game."

Conversely, seven participants cited the customizable avatar as their least favorite reward system in *Café Flour Sack*. Multiple participants (P1, P4, P6 P11) highlighted that the lack of interaction with the rest of the game turned them off from the customization options. For example, P4 stated "cause for me that doesn't really help me in any way in achieving any other goal in the game." Two participants (P3, P7) added that the value of the rewards wasn't sufficient: either too expensive (P3) or "not really sufficient" (P7). Regarding the items, P7 also added that "I wasn't really interested in keeping them since it was like a one-time game." Finally, one participant (P9) stated that it was their least favorite simply by process of elimination (i.e., they preferred both the leaderboards and the narratives more).

Unlockable Narratives Four participants cited the unlockable narratives as their favorite reward system in *Café Flour Sack*. Two participants (P1, P6) described their interest in the narratives (and "worldbuilding") of other games they had played, including RPGs (P6), whereas the other two participants (P9, P11) stated that they opted to interact with the narrative when the other reward systems did not meet their expectations. [15]

[15] Of note with regards to the game development process, P9 pointed out that "at least the narrative somebody like put work into this," among their reasons for interacting with the narrative over other reward systems.

Conversely, two participants cited the unlockable narratives as their least favorite reward system in *Café Flour Sack*. One participant (P5) commented that the narratives "were kind of long" while the other participant (P8) stated they expected the quality of the narratives to be better.

Reward System Rankings

Participants were asked to rank the rewards from most preferred to least preferred. While nearly all participants did not consider the global tracker to be a reward system for the purposes of the previous question, participants were asked to include it as part of their rankings for this question. Furthermore, participants were allowed to rank systems as equally preferred (e.g., a participant could rank both the global tracker and the leaderboards as their "least preferred" reward system).

Overall, participants described a total of eight different rankings. Three of these rankings were shared between two different participants; no ranking was shared between more than two participants, suggesting a diversity of preference rankings.

No participants ranked multiple reward systems equally as their most preferred, suggesting that all participants had an explicitly-preferred choice of reward. Meanwhile, three participants ranked their second and third preferences as equal while three other participants ranked their third and fourth (i.e., their least) preferred reward systems as equal.

Future Rewards for HCGs

Finally, participants were asked what rewards they would have liked to have seen in human computations games (i.e., what reward systems did they think were missing from a game such as *Café Flour Sack*).

While some participants (particularly P11) did remark that *Café Flour Sack* covered sufficiently many reward systems, participants still provided suggestions for game elements they thought might improve the game. One common theme that emerged was a desire for

progression to the rewards, be it through different tasks (or "levels") (P1, P3, P4, P5), bonuses (i.e., "combos") across multiple tasks (P3), or through an player-centric, leveling-up systems (like those seen in RPGs) (P1, P6). For example, P1 went so far as to describe "a roadmap [for] all of the (...) all of the possible achievements or to say rewards you can get" and how it would aid completionists. Similarly, P8 described a concrete example of their ideal interaction between the reward systems: "So like let's say if I have like this special scarf that the chef like wears or like special like knife or whatever, then you get twice as many uh points on the leadership board when you answer a question correctly or something like that."

Similarly, this theme of progression was commonly tied to desires for better system interaction (between the rewards and the action mechanics of the game). Two participants (P8, P10) separately described "power-up" systems that would reward players with abilities that could adjust the challenge of tasks.

Finally, other suggestions included using of badges/trophies (P5) as seen in other games or adding more social elements to existing game elements such as the leaderboards (P9).

5.6 Discussion

In this section, I summarize the major results of the *Café Flour Sack* study and its followup interview study, comparing and contrasting these results against relevant prior work. Given the focus on human computation game design, this discussion section is organized into a set of three design considerations, and how these topics relate to both creating an engaging player experience and effectively solving the human computation task.

These design considerations are as follows:

1. Supporting Multiple Reward Systems in HCGs
2. Offering Reward Choices to Players
3. Adjusting Reward Mechanics for Specific Player Audiences

5.6.1 Supporting Multiple Reward Systems in HCGs

Before examining the considerations around reward choice and player audiences, I first begin by discussing the considerations about supporting multiple reward systems in human computation games.

As previously reiterated, most human computation games utilize point-based reward systems as their primary in-game feedback mechanics for players. Common game elements include the use of an increasing, nominal score and often a leaderboard presentation or side-by-side listing where players can compare themselves against each other. With the exception of the work conducted by Goh *et al*. [77], most human computation games have not explored different types of reward systems.

In their work, Goh *et al*. examined the task of providing location-based content (i.e., annotating map locations with user-generated comments). They developed three mobile games, each with a different reward system, then conducted a within-subjects study which compared *task completion* and *player experience* metrics across these three game versions. The "Track" version awarded players points, which were then displayed on a global leaderboard. The "Badge" version of the game awarded players digital badges for completing certain actions within the game (e.g., contributing content, rating content, etc.) which could then be displayed on the players' profiles. The "Share" version of the game acted as a control and provided no rewards beyond a summary of player actions performed. These three reward systems may be considered comparable to the global leaderboards, the customizable avatar, and the global progress tracker in *Café Flour Sack*.

Overall, Goh *et al*. found the "Track" and "Badge" versions of the game performed better than the "Share" version of the game, both with respect to *task completion* metrics such as perceived content accuracy and *player experience* metrics such as perceived enjoyment. However, no differences were found between "Track" and "Badge" versions of the game, save for increased cognitive enjoyment in the "Badge" version of the game. These results suggest that having some kind of reward system in such HCGs would be ultimately better

than no explicit reward system (with respect to more effective *task completion* and a more positive *player experience*), but do not prescribe *which* specific reward system would be the most appropriate.

Café Flour Sack is a very different game from the three mobile games tested by Goh *et al*., but there are some similarities in the reward systems utilized by both sets of studies. *Café Flour Sack's* global leaderboards and the "Track" version both use nominal point comparison between players as feedback for task completion. Likewise, *Café Flour Sack's* global progress tracker displays statistics on player task completion comparable to those displayed in the "Share" version. Finally, *Café Flour Sack's* customizable avatars share some similarities with the "Badge" version; both use a personalizable avatar that can be decorated with digital items, however the process of acquiring and the type of said items differs greatly.

Unfortunately, the results of the studies and interviews I conducted with *Café Flour Sack* are not directly comparable to those of Goh *et al*., as *Café Flour Sack* does not specifically treat different reward systems as experimental conditions (since it provides all of them). There are some similarities; much like how Goh *et al*. found nearly no differences between their two "Track" and "Badge" reward systems (besides cognitive enjoyment), I similarly found no differences in *task completion* metrics based on the *player experience* results of participants' favorite and least favorite rewards. These similarities could possibly reinforce the conclusion that having some kind of reward system at all has more of an effect on *task completion* and *player experience* than the actual type or kind of reward system provided. However, drawing general conclusions about what kind of reward system is most appropriate for a given HCG would require further investigation and remains to difficult to test, since many reward systems are sensitive to the context of the task and aesthetic of the game.

Ultimately, the absence of a clear, most effective or overwhelmingly favored reward system suggests that there may be no detriment to considering different (i.e., not just leader-

boards) or multiple reward systems in human computation games. I now present some of the observations gleaned from both the *Café Flour Sack* study and the subsequent interviews. The numbers from the *Café Flour Sack* study are reported in aggregate across both conditions for generality (as no difference in preferences was observed between conditions). To be clear, I do not prescribe that any one reward system should be utilized over others (beyond a humble plea that it might be worth exploring other systems beyond leaderboards), but now outline some design considerations should one wish to implement similar systems in HCGs.

Leaderboards and point-based systems are the most widely-used options for reward systems in human computation games. The implementation of leaderboards in *Café Flour Sack* is very similar to leaderboards in other HCGs, wherein players are given points for completing tasks and receive more points if tasks are completed correctly. In the *Café Flour Sack* study, 39 participants (50% of participants) selected the leaderboards as their favorite reward systems while only 11 participants selected it as their least favorite. When interviewees (around 50% of whom also selected the leaderboards as their favorite reward system in *Café Flour Sack*) were asked to further explain why they felt this way, interviewees highlighted their interests in ranking up and comparing themselves against other players, as well as the social elements of leaderboards. Conversely, the one interviewee who did not enjoy leaderboards highlighted their dislike of competition as a disincentive to engage with the system.

The customizable avatar is an example of a reward system utilizing collectible, in-game objects as rewards and allowing for player personalization. The implementation of the customizable avatar in *Café Flour Sack* mimics other personalization systems by providing objects that may be purchased using the currency resulting from task completion. In the *Café Flour Sack* study, 18 participants (around 23% of participants) selected the customizable avatar as their favorite reward system, while inversely, 18 participants selected it as their least favorite. Interviewees who considered the customizable avatar (2 partic-

ipants, or around 18%) to be their favorite system noted their enjoyment of collection, customization, and the physicality of having items in the game. Conversely, interviewees who considered it their least favorite highlighted its lack of interaction with other elements of the game and that it was a poor investment (given that *Café Flour Sack* was a one-time game). The lack of interaction with other game elements is an unfortunate limitation of how *Café Flour Sack* is set up to run experiments on, but this setup is not required (and not necessarily recommended) for potential customizable avatar implementations. Furthermore, there exist other examples of systems which may appeal to players who enjoy collection or customization mechanics. The "Badge" version of Goh *et al.*'s location-based content generation HCG presents one such alternative, which is ever so slightly better at cognitive enjoyment than leaderboards and otherwise no differently at other *task completion* and *player experience* metrics. One interviewee even highlighted badges as a system that *Café Flour Sack* could support. Finally, the interviewee feedback regarding investment highlights the importance of ensuring rewards are considered meaningful to players. This is particularly relevant for content which may be interpreted subjectively [16], as players may be actively motivated or de-motivated to interact with these systems based on their preferences and experiences with the content.

Unlockable narratives in *Café Flour Sack* are an example of a novel reward system that has not been utilized or tested in human computation games before. The implementation of narrative content in *Café Flour Sack* is visually and interactively similar to dialogue scenes in narrative games such as visual novels and interactive fiction, and consists of stories that are unlocked as players complete tasks to unlock them. In the *Café Flour Sack* study, 19 participants (around 23% of participants) selected it as their favorite reward system, while 35 participants (around 44% of participants) selected it as their least favorite. Interviewees who considered the unlockable narratives to be their favorite reward system (around 36%)

[16]Points, for example, may be considered objective rewards—all players receive the same points. There is nothing to like more or less about the points, beyond liking or disliking the point systems themselves. By comparison, a hat for a virtual avatar may be considered a subjective reward, as some players may like the hat (i.e., aesthetic qualities, utility, etc.) whereas other may dislike it for the opposite reasons.

described their interest in narratives and worldbuilding in other games such as RPGs or cited it as an alternative to other systems they did not enjoy. Conversely, interviewees (2 participants, or around 18%) who considered it their least favorite stated the narratives were too long or of insufficient quality. Indeed, participants across all combinations of condition and audience spent anywhere from four to eight times as long in the narrative rewards than they did in other reward systems. This increased time investment may bode as a positive indicator that players could become more invested in the game (and by extension, task completion) provided that they enjoy the narrative content. At the same time, there is the concern that given a limited amount of time, players spending more time interacting with the narrative may solve fewer tasks (though this could be solved by aggressively gating such content behind task completion, albeit to the detriment of the player experience). The most effective presentation of narrative content is still an open question; *Café Flour Sack* provides only one possible implementation. Another example of an HCG exploring how to incorporate narrative content (albeit not as a reward) is *Project Discovery* [24], wherein real-world scientists have been incorporated into the game universe to facilitate the activities of solving the human computation tasks, thus utilizing the broader game universe to contextualize the task and engage players. Ultimately, narrative content—from story arcs to worldbuilding aspects such as the "lore" described by interviewees—can be treated as a reward for players and represents an aspect of HCGs meriting further investigation.

Finally, the global progress tracker in *Café Flour Sack* was as evaluated reward system despite provided no explicit rewards as it still provides non-gamified feedback, similar to the implementation in Goh *et al.*'s "Share" game, which acted as a control in the absence of in-game rewards. In the *Café Flour Sack* study, 2 participants selected it as their favorite while 14 participants selected it as their least favorite; no interviewees in the followup study and interviews listed it as either, save one who reported that it was their least favorite simply by virtue of the fact that they didn't pay much attention to it. The fact that any participants even selected it as their favorite reward system reinforces that there is in fact

a (small) subset of players who do not prefer any rewards (echoing the observations by Cooper *et al*. [9]). Fortunately, given that most HCGs are already tracking the summarized information in these systems, exposing this information would benefit such intrinsically-motivated players while only requiring the implementation of an additional user interface (e.g., screen). However, per the results of Goh *et al*.'s work, providing only a global tracker or only exposing this information in lieu of other reward systems may be less effective (with respect to *task completion* and *player experience* metrics) than using it alongside other reward systems.

Beyond these four reward systems, multiple interviewees described wishing to see reward systems with some kind of progression or escalation to the distribution of rewards, as well as integration between reward systems (e.g., a player could buy a powerup, which could then be used to assist with task completion, which then in turn would help them earn more points for the leaderboards). As previously discussed, *Café Flour Sack* isolates its reward systems—by design—to avoid these specific interaction effects, but could support such interaction were such restrictions not in place (e.g., allowing players' customizable avatars to show up on the leaderboards or as part of the narrative content).

5.6.2 Offering Reward Choices to Players

As part of the loop of any game, players receive rewards as feedback for their actions in game. As part of the design and development process of a human computation game, HCG developers must decide what kinds of feedback (rewards) they wish to provide players. But beyond simply picking how many and which kinds of reward systems to implement, there are many aspects and nuances of how these rewards should behave that must also be answered Assuming multiple reward systems are available or that players are given the choice to interact with multiple reward systems at all, should players be given the choice to pick which rewards they get for completing tasks?

Affording players the agency to choose their rewards may seem like the obvious choice

for a positive *player experience*, but like all decisions regarding game mechanics in human computation games, one must balance for both *task completion* and the *player experience*. For example, if rewards are assigned randomly, players are not guaranteed to always end up with the rewards they want and might end up completing more tasks to attain those rewards, albeit potentially at the expense of a positive *player experience*. [17] Similarly, if offering players a choice of reward allows them to engage with the rewards they care about, it is possible that players may spend more time engaging with the rewards than solving tasks, especially if the time spent interacting reward mechanics is balanced poorly with respect to time spent interacting with action (task) mechanics.

Therefore, in the *Café Flour Sack* study, one of the conditions being tested was whether or not participants were allowed to choose which reward systems they wanted to interact with. Overall, players in the *choice* condition demonstrated higher task correctness and were faster at completing tasks. Additionally, players in the *choice* condition perceived that they did in fact have more choice of rewards. This, however, did not appear to significantly affect interaction with the reward systems as there were no differences found in the duration of interaction between conditions, suggesting that the lengths of player experience were similar. The only exception was that players in the *random* condition spent longer in the leaderboards, but despite these (significant) differences, the duration was only on the order of several seconds. From these results, I conclude that that offering players the choice of reward benefits both *task completion* and the *player experience*.

Finally, these results also align with the work explored by Lessel *et al*. [10], who showed that players who could (and did) turn off gamified elements in the image-tagging application completed more tasks. Despite looking at different game elements, both their findings and the results from *Café Flour Sack* suggest that providing players the choice to interact with HCGs in a way that might more closely reflect players' motivations (rather

[17] These kind of randomized mechanics are central to genres such "roguelike" games, which rely on mechanics such as randomizing item drops and thus work to retain players who are motivated by the chance of getting better items and progressing further in replays of the game.

than forcing them into an unchangeable, interactive game experience) may have benefits to *task completion* metrics, not just those related to the *player experience*.

5.6.3 Choosing Reward Mechanics for Different Player Audiences

Beyond simply looking at whether or not to give players a choice of reward, I also examined the effect of asking different player audiences to complete the task. In the *Café Flour Sack* study, the two study populations consisted of participants recruited through Amazon Mechanical Turk, a professional crowdsourcing platform where workers complete tasks for both full and part-time work, and participants from a university study population, specifically students from engineering backgrounds who were likely to be familiar with games, but not crowdsourcing work (or HCGs).

Overall, Amazon Mechanical Turk participants performed significantly better than student participants at all task completion metrics: task correctness, number of tasks completed, and rate of task completion. These results are unsurprising, given that Amazon Mechanical Turk participants are considered crowdsourcing experts. [18]

As previously mentioned, Amazon Mechanical Turk participants in the *random* condition were the most effective players overall, significantly so when it came to both task correctness and the rate of task completion. However, these differences in *task completion* metrics, compared to the next most effective population, are significant but small. When separating student participants by experimental condition, student participants in the *choice* condition have *task completion* metrics nearly comparable to those of Amazon Mechanical Turk participants. This is not the case with the *random* condition, where the gap in metrics is much larger. So while the two participant audiences performed very differently on *task completion* in one experimental condition (student participants significantly lower

[18] Arguably the AMT participants were also motivated by monetary compensation; whether or not such compensation is comparable or greater than to student participants' additional motivation by course credit is unclear. Given that that the overall play duration was a set time limit (in which there were no other requirements beyond "play"), neither audience would have been more motivated than the other to do more tasks or to complete them faster, which suggests that expertise at task solving may have played a large role.

than Amazon Mechanical Turk participants in *random*), they were comparable in the other condition *choice*. The findings are limited because the task was selected for its simplicity, relying primarily on commonsense knowledge without additional training. However, for more complicated tasks, such improvements could be very valuable. Combined with the previous consideration, these results suggest that design decisions—such as offering players multiple rewards—have the potential to greatly improve *task completion* without negatively affecting the *player experience*.

Indeed, a design concern unique to human computation game design is determining which gameplay elements have the most significant effects on both *task completion* and the *player experience*. In the *Café Flour Sack* study, the visible difference between the *random* and the *choice* conditions was a single screen that either showed the upcoming reward (*random*) or allowed the player to choose their reward before the upcoming round of gameplay (*choice*). [19] This subtle variation suggests even very small changes in design of reward mechanics could have potentially large impacts on *task completion* and the *player experience*. Furthermore, the interaction effects between how participants were rewarded and the participant audiences involved demonstrates the importance of paying attention to the intended player audience, and if possible, even tailoring small aspects of the game's design to optimize for both the task and players. Taken altogether, the results of the *Café Flour Sack* study help to reaffirm the importance of reward mechanics to both *task completion* and the *player experience*.

5.7 Conclusions

In this chapter, I describe a human computation game, *Café Flour Sack* , and a study using the game which compares reward distribution mechanisms across two different player audiences. There are two reward distribution mechanisms treated as experimental conditions: one in which players are given *random* reward versus one in which players are given

[19] This is, of course, accompanied by the very simple game logic to either randomly select a reward or to turn on the buttons allowing players to select the reward instead.

a *choice* of reward. Additionally, the study population is deliberately spread across two separate player audiences: Amazon Mechanical Turk workers and university students, reflecting expert versus amateur crowdsourcers. I also describe the setup, as well as some anecdotal results of a followup study conducted with student participants, which consists of a play session using *Café Flour Sack* and semi-structured interviews afterward, designed to better understand how participants respond to different kinds of rewards in HCGs.

This chapter began with two questions. The first question was how does *randomly* distributing a reward versus giving players a *choice* of reward affect *task completion* and *player experience*? Overall, the results demonstrate *choice* of reward is beneficial, especially for *task completion* metrics. Participants in the *choice* condition were both faster and more correct at solving tasks. The results for the *player experience* metrics are more nuanced. Participants in the *choice* condition did perceive that they had more choice of rewards, however there was no difference in the duration of interaction with the reward systems, suggesting that participants' interactions with the rewards between the two systems were similar. There was one exception: participants in the *random* condition spent significantly longer interacting with leaderboards (albeit on the order of a couple of seconds). Overall, this clear benefit to the *task completion* metrics (in combination with the few, but positive effect on *player experience* metrics) suggests that offering players the choice of reward (assuming multiple reward systems are available) is the superior option compared to randomly assigning players rewards.

The second question was do different player audiences have noticeable differences on *task completion* and the *player experience*? Here, two different player audiences—professional crowdsourcing workers (recruited through Amazon Mechanical Turk) and amateur crowdsourcers (recruited through a university student population)—were used for the study. Perhaps unsurprisingly, Amazon Mechanical Turk participants outperformed the student participants at all *task completion* metrics: task correctness, number of tasks completed, and rate of task completion. While this result may seem predictable, professional

crowdsourcing workers are also not guaranteed to be expert game players and the task in question—pairing food items to potential recipes—was a commonsense knowledge task that does not require expertise that professionals are guaranteed to have over amateurs.

However, the interaction effects between the two experimental conditions: *random* versus *choice* of rewards, combined with the two participant audiences, yielded some unexpected results. Notably, student participants in the *choice* condition performed almost comparably to Amazon Mechanical Turk participants overall in *task completion* metrics such as task correctness and the rate of task completion (whereas student participants in *random* condition yielded much poorer results). This suggests that subtle changes in rewards mechanics could potentially induce amateur crowdsourcers to perform comparably to experts. While the impact on a commonsense problem with a knowledge solution is small, such differences could potentially impact (and hopefully benefit) more complex human computation tasks. Additionally, these results reaffirm that even small changes in game mechanics, like a single screen difference related reward distribution, can have potentially significant effects on *task completion* metrics. This observation, as well as the other interaction effects described above, demonstrate that reward mechanics appear to be sensitive to player audience and that intended player audience should be taken into account when considering their design.

Finally, *Café Flour Sack*, its initial study, and the subsequent interviews demonstrate potential implementations of various reward systems in games, as well as providing quantitative data and qualitative feedback of their utility. While leaderboards remain the most preferred system in *Café Flour Sack*, reward systems such as the customizable avatar and the unlockable narratives resonated with substantial percentages of other players for different reasons. Importantly, participant preference for reward system did not impact *task completion* and *player experience* metrics, echoing similar results from studies of different reward systems [77]. This suggests that reward systems beyond leaderboards may be viable alternatives, particularly if certain demographics or audiences of players with known

preferences and motivations for play are being targeted. [20] The results of the interviews also scratch the surface of why players prefer certain kinds of rewards; further work is needed to understand just how these different reward systems map to different motivations for play. I consider all of these results promising for HCGs, as diversifying the kinds of reward systems utilized by HCGs will serve wider player audiences, thus helping to keep HCGs relevant and useful in the ever-expanding ocean of digital games.

[20] Though obviously, the effort required to implement different systems are not equal. For example, the unlockable narratives require written content, which could be consider substantially more work than implementing an automatically-increasing leaderboard.

CHAPTER 6

PERSONALIZED REWARD SYSTEMS IN HCGS

6.1 Introduction

Imagine a world where a robust and rigorous body of design knowledge for building human computation games has been developed. Task providers and small-scale game developers can look up any game element and easily understand how its inclusion into an HCG for a given task will affect both the *player experience* and *task completion*. As a result, HCGs are both easier to make and more effective to use, resulting in these games becoming broadly adapted interface for solving crowdsourcing tasks.

However, even in an unrealistically idealized scenario such as this, human computation games are not immune to changing task requirements and the whims of players preferences. What if an HCG developer wishes to reuse their currently-successful HCG for a similar, yet novel task? What if the HCG must target a new demographic or audience of players whose preferred game mechanics are the opposite of those currently implemented in the game? What happens as developments in game hardware and game design enable entirely unpredictable game experiences? Such unforeseen changes may necessitate changing, updating, or even re-implementing the entire HCG.

Even with the best possible design knowledge available at hand, making changes to a game may be expensive even for the most well-funded task providers (not to mention industry-scale game developers). Updating and changing games may be nearly as time consuming as making a new game and as repeatedly emphasized, task providers and small-scale HCG developers typically do not have the unlimited resources able to facilitate this process. As shown in the previous chapter, even small changes to an HCG may have significant effects on *task completion* and the *player experience*. So when addressing an

unknown task, a brand new player audience, a new hardware platform—any one of a myriad of unknowns for which design knowledge may ultimately not exist—HCG developers will continue to face challenges keeping their games updated, efficient, and relevant.

While I have advocated for a robust, rigorous compendium of human computation game design knowledge that addresses both task and player needs throughout this dissertation, I also acknowledge that this is just a first step to making HCGs effective and successful. Every year, more and more games become available to players. Every year, player tastes change as different genres and types of games wax and wane in popularity. To be as accessible and effective as possible, HCG design and the games themselves need to be able to adapt to these external changes, but in a way that respects the resource limitations of their task providers and game developers.

One potential means of easing the burden of HCG development on task providers is to consider systems that automatically adapt or tune elements of the game to changing circumstances. In games, algorithmically adapting or adjusting gameplay mechanics or elements is driven in response to player interactions taken in the game is commonly known as *dynamic-difficulty-adjustment* (DDA) [105] or *challenge-tailoring* [106]. These techniques observe player interaction with the game, utilize or infer a model of player performance or preferences, and then automatically select or parametrize game elements to create a more customized play experience. The potential of adaptive techniques has been demonstrated across a wide variety of game genres, including platformer games [107], action roleplaying games [108], first-person shooters [105], puzzle games [109], and interactive narrative systems [110, 111]. More recently, high-profile games such as *Left 4 Dead* [112] and *Middle-earth: Shadows of Mordor* [113] have implemented adaptive systems to commercial and critical success. Beyond specific game genres, adaptive techniques have been used to parametrize and tune more general categories of game elements, such as the aesthetics of in-game objects (e.g., ship sprites) [114] and movement controls based on player sensitivities [115]. Beyond gameplay, adaptive techniques have been applied to tools in the

game design process as a way to use *designer preferences* to explore the design space for domains such as game maps [116], rhetorically-inspired minigames [117], and aesthetic content retrieval [118, 119].

Most personalization and adaptive systems rely on the ability to understand players from in-game behavior, in order to infer player performance and preferences. Modeling player behavior has been thoroughly studied across a wide variety of game domains [120, 121], including serious games [122]. When it comes to modeling player preferences, popular approaches include utilizing existing player typologies, such as Yee's motivational components [92], the neurobiologically-inspired Brain Hex [96], and the gamification-focused Hexad framework [97], or alternatively, learning mappings of in-game behaviors (e.g., those aggregated through in-game telemetry data) to modifiable game elements/parameters directly [123].

Beyond general games, personalization and adaptive techniques have also been applied in other serious game domains. Personalization systems have been shown to be effective for serious games [124] and gamification [125]. In education, adaptive techniques are commonly used to select, generate, or personalize educational or curricular content based on student needs' or preferences. Adaptive techniques power intelligent tutoring systems [126] and other serious games [127, 101], adjusting student material to match perceived knowledge models or learning mindsets. Adaptive techniques have also been utilized in training and simulation games [110], where in-game scenarios are altered based on the needs of and in-game choices made by players. However, it is not well understood how effectively these techniques and models might transfer and apply between domains (i.e., to HCGs). For example, not all typologies designed for general games successfully transfer to serious game domains such as education [128].

Meanwhile, personalization and adaptive game elements are a relatively-unaddressed area of research and design for human computation games. At best, HCGs may update their mechanics or in-game tasks to accommodate changes to the task or its requirements.

Seminal examples of this include the addition of taboo words in the *ESP Game*, which were added after the game's task providers desired a more diverse list of labels for their images, and the updates of new protein folding challenges in *Foldit*, which came about as the result of task providers choosing to use the game as a platform to solve subsequent protein folding optimizations, not just the original set of tasks. In these examples, all of these changes were reactive and required the intervention of the games' developers to update the mechanics or content of the games. However, most recently, Lessel *et al.'s* work using a gamified image-tagging application [10] shows the potential for improving *task completion* metrics when players are given the choice to turn gamified elements on and off (i.e., allowing players the ability to "personalize" their ideal task environment and experience). So while other serious game domains have explored personalized and adaptive game elements, HCGs have only taken the first steps for investigating this kind of work.

Beyond the benefits to engagement highlighted in other domains, why could human computation games benefit from the use of personalized game elements? One reason is that, as alluded to above, limited design and developer resources make systems that handle the iterative, but timing-consuming nature of game mechanic parameter-tuning easier. Another reason is that adaptive systems rely on heuristics and player models to detect when to adjust game mechanics. HCGs already have these heuristics available—often implemented as part an HCG's *verification mechanics*—thus making optimizing for objective task quality a possibility. Thus HCGs are a good candidate for exploring personalization systems, even for "fuzzier" (but more commonly-explored) metrics such as player challenge (e.g., providing easier or different tasks if a player repeatedly fails to provide results of sufficient quality) or player engagement (e.g., enticing players with different/personalized content if they appear to be bored with existing tasks or rewards).

6.1.1 Personalizing Rewards

What would it mean to personalize *rewards*? Personalization systems typically operate by observing some input—player actions, inferred metrics, player models, etc.—and then adjusting some aspect or parameter of the game's mechanics as output in response.

For human computation games, inputs such as player performance or player preferences are similar to those seen in entertainment-oriented games, albeit with a focus on solving tasks. In the context of HCGs specifically, performance at solving tasks—how well, many, or varied the results of a task are—may be observed as players interact with *action* mechanics and validated via *verification* mechanics. For example, HCG players who have a low percentage of correctly-solved tasks might be analogous to players who repeated fail an in-game challenge in an entertainment-oriented game. In both cases, the solution might be to provide different or easier tasks/in-game challenges. Other player-centric metrics such as player preferences for certain in-game elements can be utilized. For example, players in a role-playing game may prefer certain types of quests (e.g., those which give a specific kind of reward), which in turn may induce an personalized system to offer those players more quests of that type (with those particular rewards). For HCGs, a similar example might involve players who prefer solving certain kinds of tasks (e.g., labeling only pictures of cute animals), in which a personalization system might surface these kinds of tasks more frequently to such players to keep them more engaged with the game.

Given these inputs, what specific aspects or parameters of HCG rewards could be modified as a result of player performance or preferences? Both the previous chapters and prior work have explored and discussed variable aspects of reward systems such as reward *function* (Chapter 4), reward *distribution* (Chapter 5), and reward type (Chapter 5, as well as prior work [79, 77]).

Before proceeding, I wish to highlight that personalizing or adapting aspects of in-game rewards is not as ethically straightforward as adjusting other game elements. *Feedback* and reward mechanics are typically the primary form of compensation to players, providing

intangible, digital, and emotional rewards often in lieu of what might be monetary compensation on a non-gamified crowdsourcing platform. At their very worst, any personalized or adaptive reward systems may not compensate players sufficiently for their time and work, and/or fail to disclose to players about how their rewards are being adjusted. In particular, while the *amount* of reward given to players may appear to be an easy parameter to adjust in an adaptive system, inappropriate implementations risk turning the HCG into an opaque, automated arbiter which distributes player compensation in unintended ways or induces players to "game" the system to maximize compensation at the expense of task results. However, at their best, personalized reward systems could help task providers to build a healthy relationship with their players. For example, by providing players with their preferred rewards (or extra rewards for going above and beyond the expected task work), players may be more inclined to participate in a crowdsourcing process that they feel is fair, and/or respectful of their time and interests than one which is impersonal or outright antagonistic. It is for this reason that I now propose focusing on personalized reward systems that rely on player preferences as input, specifically player reward preference. I strongly reiterate that optimizing rewards purely for the benefit of task providers at the expense of players not only risks unfair compensation, but also may result in ineffective HCGs that no one will wish to play and risk affecting the perception of HCGs as a whole.

6.1.2 Summary

The work in this chapter is motivated by the following question: how do *personalized reward systems* affect *task completion* and *player experience*? Specifically, how do reward systems which use the player's preferred reward to give players more of that reward for longer tasks compare against those which do not?

In this chapter, I describe the modifications to the *Café Flour Sack* game. These modifications include the inclusion of new reward mechanics, such as allowing players to select from multiple tasks of different lengths with longer tasks providing more rewards, and the

basic implementation of a personalized system that will automatically select the player's preferred reward when distributing these longer tasks. I then describe a study using *Café Flour Sack* that compares two versions: one with a non-personalized reward system against one with the personalized reward system. I then summarize the *task completion* and *player experience* metrics measured during the study and follow with a discussion of their potential design implications. This chapter continues to use the mechanics framework and experimental methodology described in Chapter 3.

This chapter consists of three parts:

1. A description of a modifications to an HCG, *Café Flour Sack*, made to explore personalization in HCGs.

2. A human-subjects study using personalized *Café Flour Sack* and its results.

3. Discussion of the results of the study.

6.2 Extending *Café Flour Sack*

Applying personalization to games comes with the potential to provide a more customized experience for each player that in turn may yield a more enjoyable experience with the game. In the case of human computation games, personalized game mechanics may work to serve both the dual goals of both player experience and completion of the human computation task. Personalized game mechanics may ensure that HCG players are given a play experience that aligns with their interests, ideally one which respects their particular extrinsic or intrinsic motivations for play. Simultaneously, personalization may optimize task completion metrics by inducing HCG players to complete tasks more correctly or quickly by changing in response to their performance. Like all elements of HCGs, optimizing for only one of these may come at the expense of the other; an ideal personalization system should optimize for both tasks and players.

Most existing work in personalization and adaptive techniques would apply to what may be considered the *action* mechanics of human computation games. However, this dissertation work focuses on *feedback* and reward mechanics. As previously mentioned, I propose investigating personalized reward systems that rely on player preference as input, as opposed to systems that look at player performance. In this section, I argue that adjusting the specific reward type of certain tasks is an appropriate scenario for an personalized reward system and describe an implementation as an extension to an existing HCG.

6.2.1 An Experiment in Personalized Reward Distribution

So what aspect of reward systems could be personalized? I propose if that multiple rewards are available (to accommodate different types of players and motivations) that players should also be rewarded more favorably for tasks that are more complex or time consuming. Not all (sub)tasks for a given human computation problem are created equally. An image in a data set may be perniciously blurry or contain a multitude of ambiguous objects. A particularly gnarly protein structure may have an optimal energy configuration that is near-impossible to manipulate in a 3D rendering interface. A real-life building requiring pictures for 3D model generation may be in a particularly isolated location that the average person could not reach. All of these scenarios are instances wherein a task requires more effort (or even a specialized solution) from the player compared to the average task. All of these are scenarios that task providers could potentially (and in some cases, have) encounter(ed). I refer to these kinds of tasks informally as *super tasks*—human computation tasks that require above-average time or effort to solve.

If reward systems are the primary mechanism for player feedback, compensation, and gratification, then it goes without saying that players should be rewarded more for completing such tasks. But personalized reward systems have the potential to take this further. Personalizing rewards may allow HCGs send the message to players that task providers appreciate the extra effort involved for solving such *super tasks*. I therefore propose exploring

this idea of giving players their favorite rewards for *super tasks*. For the player, receiving more rewards of their favorite category is a further acknowledgment of the additional work or effort required to complete the *super task*. For the task provider or HCG developer, a preferred reward may act as motivator for the player attempting the task or an impetus to provide a better task solution. Thus, personalizing to a player's favorite reward in the context of *super tasks* could benefit both the *player experience* and the *task completion*.

6.2.2 Updating *Café Flour Sack* for Personalization

To explore this question, I chose to re-utilize *Café Flour Sack*, the game developed previously for the experiments in Chapter 5. Using *Café Flour Sack* leverages its its multiple reward systems, its known task solution set, and the infrastructure enabled by the previous experiments.

As a brief recapitulation, *Café Flour Sack* is a cooking-themed HCG that assigns players the culinary-commonsense-knowledge task of pairing food ingredients to recipes that likely contain those ingredients. This commonsense-knowledge task is an artificial task with a known solution set that allows measuring *task completion* metrics objectively without needing to simultaneously solve a novel human computation problem. Additionally, the game contains four separate reward systems or categories: global *leaderboards*, customizable virtual *avatars*, unlockable *narrative* stories, and a non-gamified global *progress tracker*. For more details on the reward systems, please refer back to Section 5.2.1); beyond some light parameter tuning (e.g., changing the costs of certain items in *avatar* and *narrative* reward systems), these rewards remain otherwise unchanged from their original implementation.

In the original versions of *Café Flour Sack* from the previous studies, players would click a button to start each round. Based on the condition of the study, players would either be randomly assigned one of the three available reward types [1] (the *random* condition) or

[1] The available reward types are the *leaderboards*, the customizable *avatar*, and the unlockable *narratives*. The global progress tracker is not available as a reward type because it has no explicit reward currency

Figure 6.1: The new start screen for *Café Flour Sack*. For each round, the player selects one of three available options; this round has a *super task* available for the avatar category.

be allowed to choose a reward type at the start of the round (the *choice* condition). Each round would then contain five minigames (tasks). Each task would consist of a recipe for a food dish (e.g., cake) and a random selection of ingredients, from which the player selected those which could (or could not) be used to make the assigned food dish (e.g., "flour" may be used to make cake, but "anchovies" may not). At the end of the round, the player would be informed how many tasks they completed correctly and this would correspond to the amount of reward they would receive for the given reward system.

Selecting Multiple Round Options and the Super Task

As *choice* of reward was determined to be the better option, I modified the implementation slightly to make multiple options (i.e., sets of minigames) available to the player. Instead of pressing a button get a single option of five randomly-generated minigames, players

attached to it.

are now shown a list of three round *options*. Figure 6.1 shows a screenshot of this new implementation, with a list of three options. Clicking on an option will start a round where players will be given rewards in the system whose icon is on the option.

While this implementation may appear mechanically identical to the use of the reward selection screen from the *choice* condition, exposing separate round options per reward type comes with a benefit. The original *Café Flour Sack* has no notion of anything like a *super task*, since a round was defined by exactly five tasks, each with a randomly-selected recipe and randomly-selected ingredients. In the updated *Café Flour Sack*, a *super task* is a round of the game which increases the number of minigames (tasks) from five to eight (i.e., multiplying the amount of work by 1.5). [2] I chose to make the *super task* a longer round of minigames rather than a more "difficult" series of minigames, since a more "difficult" task is not particularly easy to define for a commonsense-knowledge problem with a known solution (particularly one which involves a basic understanding of what ingredients go into common food items).

To compensate players for completing the *super task*, the amount of reward they are given at the end of the round is doubled. For example, if a player earns four points of leaderboard score by solving four of five tasks in a round correctly, these would be doubled for a final score of eight points were this the *super task*. This particular multiplier makes selecting the *super task* an attractive option for players looking to maximize their rewards, as players can do one and a half times the amount of work for twice the rewards. *Because of the scaling factor, I had to adjust some of the costs for items in the customizable avatar and the unlockable narratives, as the original costs would have allowed players to unlock them too quickly.*

However, since the rewards for accepting the *super task* are (deliberately) skewed in favor of the player,[3] the *super task* appears once every three rounds of gameplay (i.e.,

[2]The increase from five to eight minigames was chosen after some amount of tuning for this particular study; doubling the minigames, in particular, was found to be overly long.

[3]In reality, an HCG would probably be much more selective about balancing parameters like multipliers and such.

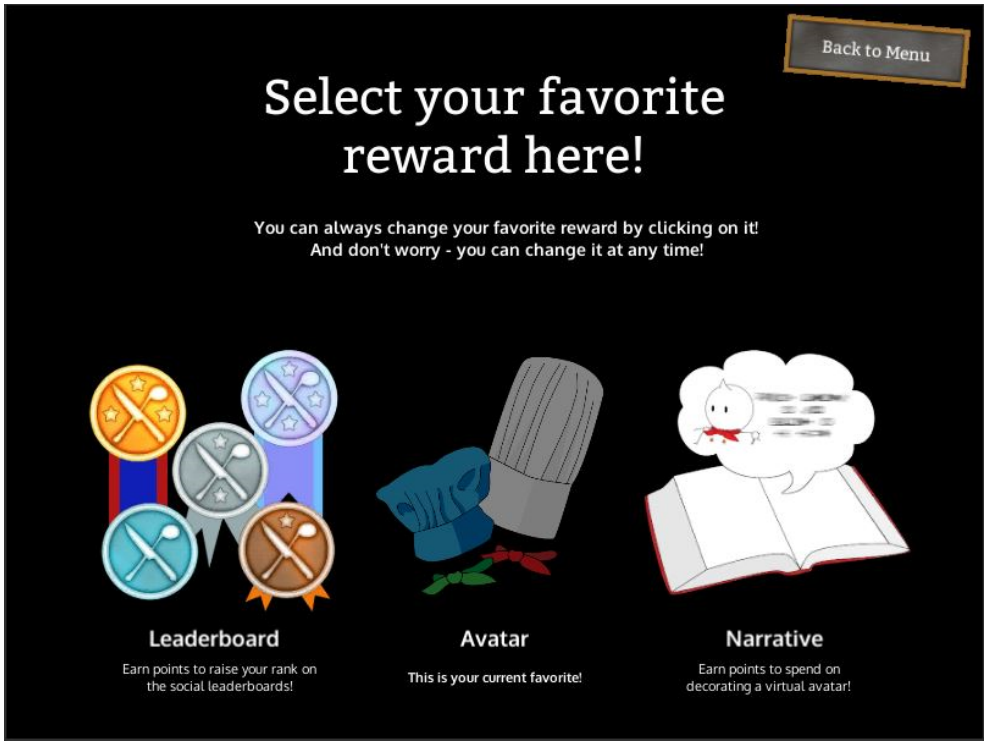

Figure 6.2: The new reward selection screen for *Café Flour Sack*. The player can click on an icon to select a reward; currently the player has selected the customizable avatar as their favorite reward type.

every third set of options), and is assigned to only one of the three options available for that round. Figure 6.1 shows such a set of options where a *super task* is available for the *avatar*; the corresponding option shows that it has eight minigames (labeled as "dishes" in the in-game text) with a "2x" or double multiplier. With these adjustments, the decision of *which* reward type is assigned to the *super task* can now be random...or in the case of this experiment, delegated to a personalization system that looks at player preferences to determine the appropriate reward type.

Identifying Player Preferences

In order to make a personalized reward system work, *Café Flour Sack* needs a way to infer player preferences—specifically which reward type at any given point in time, is the player's favorite reward. One method to determine this information is to train a player modeling system to infer a player's preferences based on their in-game actions, where the

benefit of automatic inference is that there is little to no direct player input. However, given that *Café Flour Sack* is designed to be run in an experimental setting for a much shorter duration of gameplay than most games, it is difficult to automatically infer a player's preferred reward with such limited information (which would also require training a model on players in the target audience). Another way is to simply ask players, either directly or through indirect means such as a pre-game survey (i.e., which then would map player answers to a particular reward system).

For the updated *Café Flour Sack*, I opt the latter approach of asking players directly by adding a "favorite reward selection screen," which is accessible from the main start screen of the game. This screen gives players a short form description of each reward and then allows them to select a current favorite reward type. Figure 6.2 shows a screenshot of this interface, in which the customizable *avatar* is currently selected as the player's favorite reward type.

Players are required to select an initial favorite reward after being shown all four reward systems (as part of an upgraded tutorial) before proceeding through the rest of the game, thus giving them exposure to the reward systems before asking them to make a choice. Additionally, because a player's preference for certain game elements is not guaranteed to be the same throughout the entire duration of gameplay (i.e., players may no longer have the same favorite or wish to explore others), players are allowed to return to this screen at any point and reselect a different reward.

By tracking the player's current favorite reward, the updated *Café Flour Sack* now has the means of knowing what reward type to assign players for the *super task*. A personalized reward system can now use this information directly; a non-personalized version of the same reward system simply pick the reward type randomly (thus ignoring the player's specified preference). Figure 6.3 demonstrates the final breakdown for this experiment using the mechanics framework from Chapter 3.

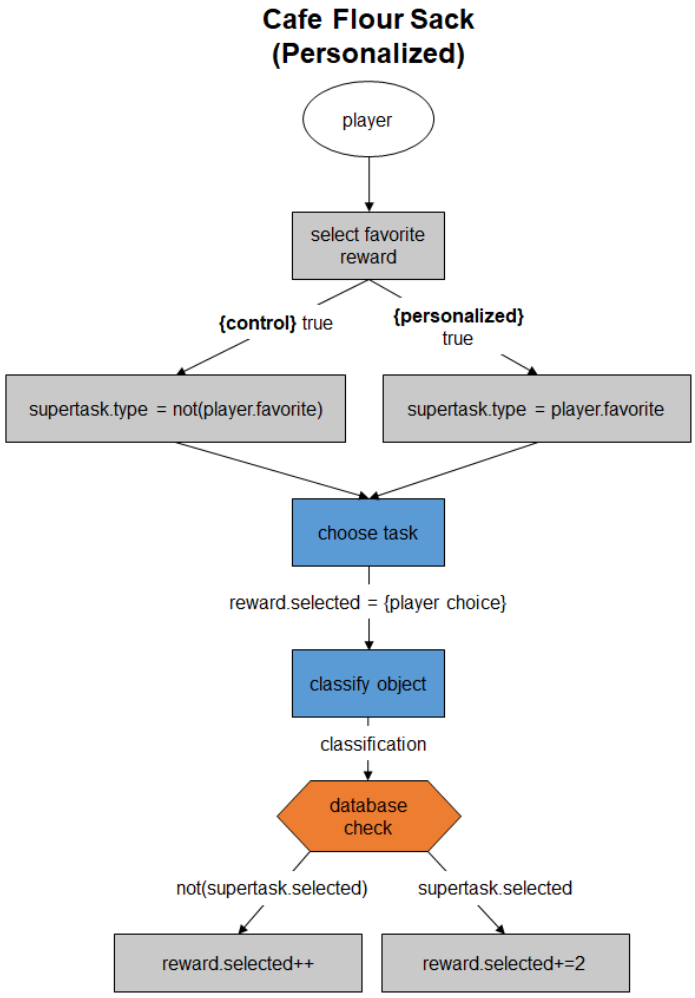

Figure 6.3: The breakdown of the personalized version of *Café Flour Sack's* mechanics and experimental conditions. Here, the selection of the reward type of the *super task* is the experimental condition (shown in boldfaced braces).

6.3 Methodology

The overall methodology for this experiment follows a similar flow to the previous studies conducted in Chapter 5. The experimental condition changed what the reward type of the *super task* would be. In the *control* condition, the *super task* would be randomly assigned to one of the two reward categories that were *not* the participant's currently selected favorite reward. In the *personalized* condition, the *super task* would always be assigned to be the participant's currently selected favorite reward. There were no visual or interactive

differences between the two versions (i.e., no additional screens, UI elements, or changes in interaction flow). The only difference between the versions would be which reward type the *super task* was assigned to. For example, Figure 6.1 shows a list of tasks where the *super task* is assigned to the avatar category. If the participant had selected the customizable avatar as their current favorite reward type (e.g., as shown in Figure 6.2), then this screen would reflect the *personalized* condition, wherein their favorite reward matches the category of the super task. In the corresponding *control* condition, the *super task* would instead be assigned randomly to either the leaderboard or narrative categories.

For this experiment, I recruited participants from a single population: crowdsourcing professionals (workers) from Amazon Mechanical Turk. This decision was made based on the results of the previous experiment in Chapter 5, which demonstrates the utility of using Amazon Mechanical Turk workers as well as potential interaction effects from using different player audiences. [4] *Café Flour Sack* was posted as a task (HIT) on the Mechanical Turk platform, where workers were shown a form with a set of instructions and a link to the external website hosting the game. [5] The HIT was distributed in batches of nine (i.e., nine workers participating on the same batch). Only one batch was available at any given time (to make the results easier to monitor and validate) and batches were distributed at various times of day to accommodate a global range of workers (i.e., to avoid the scenario wherein only workers awake in accommodating time zones could become part of the batch).

Upon accessing the game, workers (now referred to as participants) were shown consent information and asked to take a pre-game survey (identical to the one used in the followup study described in Section 5.5). After submitting the survey, participants were then randomly assigned to one of the two conditions—*control* or *personalized*—whose differences were previously described. As before, the game's backend servers generated the experimental condition randomly.

[4] Some global, logistical difficulties at the time of running the study also made using a student population prohibitively more difficult.

[5] All old versions of *Café Flour Sack* used in prior experiments were removed. This version is available as of the writing of this dissertation.

Gameplay began with a short tutorial round of five minigames (tasks), after which participants were given currency in all three eligible reward categories (minus the global tracker, which does not use currency). Participants were then directed to all four reward screens in order to view or spend those currencies before progressing further. Finally, participants were directed to the new screen that allowed them to specify their favorite of the rewards (minus the global tracker, which again has no currency). Participants were informed that they could change their favorite reward at any point by revisiting this new screen.

Participants were then asked to complete as many rounds as they desired for the remaining duration of the experiment. They were also informed that they could interact with as many or as few of the reward systems as desired. From the start of the tutorial round through the end of the experiment, participants were given a minimum required time of ten minutes (but were not explicitly informed how long this duration was; the consent form and task instructions stated that the entire study would take "around twenty minutes,").

For this experiment, participants were required to complete at least three rounds of tasks. This requirement ensured that participants would be exposed to a *super task* being made available at least once (on the third of those required rounds). Participants were not told of this requirement to ensure that they did not attempt to subvert any part of the experiment. Only after playing for the required duration and completing the required three tasks would the "Finish the Study" button appear. (Participants were told that if this button did not appear to try to complete more tasks.)

After clicking the "Finish the Study" button, participants were asked to fill out a post-game survey about the experience (and to provide demographic information if desired). Participants were then provided a code which could be used to provide proof of completion. Submitting this code would then compensate Mechanical Turk workers once the online submission was validated against logged telemetry data. [6] Workers were allowed to

[6] As previously, workers were paid $7 for completing the experiment; please note that this experiment was shorter in duration than the studies conducted in Chapter 5.

complete the study only once; duplicate attempts (e.g., through the use of multiple accounts per person) were removed when detected.

As before, I took additional steps to account for extrinsic motivations (i.e., monetary compensation), particularly since the compensation rate for participation was considerably higher than the average Mechanical Turk HIT. First, participants were required to play the game for at least a fixed amount of time *and* to complete at least a certain number of tasks. The exact time duration and number of required tasks were not directly disclosed to the participants. As noted above, this forced duration of play was again used to ensure that participants would not be incentivized to rush through the experiment as quickly as possible. Unlike the previous experiment, participants did need to complete a required number of tasks (to guarantee exposure to the *super task* mechanic).

For this experiment, I removed the "boredom button" from the game's menu. The previous studies used the button as a proxy for when players would have wished to quit the game due to boredom (i.e., a measure of retention), however no significant differences were detected between the experimental conditions and the participant audiences when measuring the time it took participants to press the button. While previous studies provided useful feedback above why players pressed the button (e.g., running out of reward content, lack of interest in the game, confusion, etc.), I chose to streamline the interface of the game and use the screen space to prioritize the button which would take participants to the favorite reward selection instead. Thus, in comparison to the previous studies, boredom was not a subjective metric measured in this experiment.

Finally, as before, real-time adjustments to both the leaderboards and the progress tracker using a simulation of fake players and results were used to normalize the experience for all players by providing the perception that other players were simultaneously playing the game, but without using previous participants' results. This would prevent any scenarios, such as later players being discouraged from social elements like the leaderboards, should an earlier player display extremely high results. Instead, all participants

would see similar sets of simulated social results.

6.4 Results

The study was conducted over the course of a week and a half, during which the game was made available online to workers on Amazon Mechanical Turk. I report on results from 74 participants who took part in the study. In actuality, I acquired data from 94 participants, 20 of which were determined to be duplicate users based on similarity of results. While the task on Amazon Mechanical Turk was set up so the workers could not repeat it (and were explicitly told not to), there is nothing stopping participants from using alternative user accounts to repeat the task. To detect duplicated users, I compared in-game user names (which participants enter on start so that in-game characters can address them by a provided name), in-game completion times, and survey results, and then excluded data for any users whose results were too similar to those previously provided. [7]

When divided by condition, 37 participants were assigned to the *control* condition and 37 participants were assigned to the *personalized* condition.

When considering self-reported population demographics, 22 participants self-reported as female and 52 self-reported as male. Figure 6.4 shows the breakdown of participant age. Most subjects fell into range of ages from 23–30. 61 participants reported that they played games regularly; 20 participants reported having played an HCG before.

The evaluation focuses on both subjective metrics related to the *player experience* and objective metrics related to the *task completion* between the *control* and the *personalized* (experimental) conditions. All data were treated as nonparametric in nature. Wilcoxon rank sum tests were utilized for continuous data analysis; Kruskal-Wallis tests were also utilized when the number of data points between the two conditions was tied. For categorical data, Pearson's Chi-Squared tests were utilized.

[7] The post-survey questions proved to be a very good litmus test for detecting duplicate results, as duplicate users would frequently repeat the same three favorite games—occasionally with identical spelling errors—when asked this question during the post-survey.

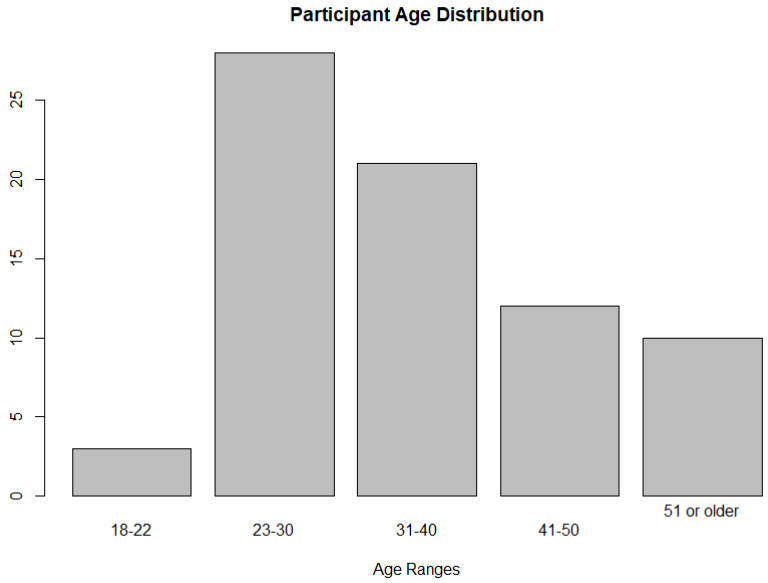

Figure 6.4: The age distribution for participants in the personalized *Café Flour Sack* study.

The subsequent sections report on these results; the discussion of their implications follows in Section 6.5.

6.4.1 Subjective Metrics—Player Experience

As in the previous studies with *Café Flour Sack*, the data contributing to the evaluation of the *player experience* consist of both player responses to questions on the post-game survey [8] and telemetry events from player interactions with the game.

First, I report on participants' survey results. These include responses regarding their favorite rewards (when ranked), as well the responses to post-survey questions regarding aspects of play such as engagement and perception of reward fairness.

Next, I report on participants interactions with different reward-related aspects of the game. I begin with reports on the duration of play within the reward system screens. Additionally, I report on participant interactions with choosing the option for each round, how

[8]The pre-game survey results were not used, although they did have the benefit of being an excellent means of helping to screen of the study results for participants who attempted to take the study twice, since repeat participants using different worker IDs would frequently fill out the study repeating the same value for all questions in the survey.

Table 6.1: Counts of participants' favorite rewards in the personalized *Café Flour Sack* study.

	Leaderboards	*Avatar*	*Narrative*	*Tracker*
Control	20	4	10	3
Personalized	17	7	7	6
Total	37	11	17	9

those choices aligned with participants' currently-selected favorite reward or the presence of a *super task*. Finally, I report on participant interactions with the favorite reward selection screen.

Favorite Rewards

As part of the post-game survey, participants were asked to rank the reward systems in order of preference. Table 6.1 shows the distribution of favorite (i.e., highest ranked reward) across the two conditions. In total, participants' favorite rewards were the leaderboards (37 participants), the unlockable narratives (17 participants), the customizable avatar (11 participants), and the global progress tracker (9 participants). This ordering held across both conditions (with the unlockable narratives being slightly more preferred in the *control* condition, but equivalently so with the customizable avatar in the *personalized* condition). No significant differences were detected between conditions.

Post-Survey Questions

The post-game survey asked participants various questions on their experience. Here, I report on five questions from the post-game survey in which players were asked to rate the statements on a 1–7 Likert scale, where "1" corresponded to "strongly disagree," 4 corresponded to "neither agree nor disagree," and 7 corresponded to "strongly agree."

The five statements were as follows:

1. "I found the reward systems **engaging**."

2. "I found the reward systems **frustrating**."

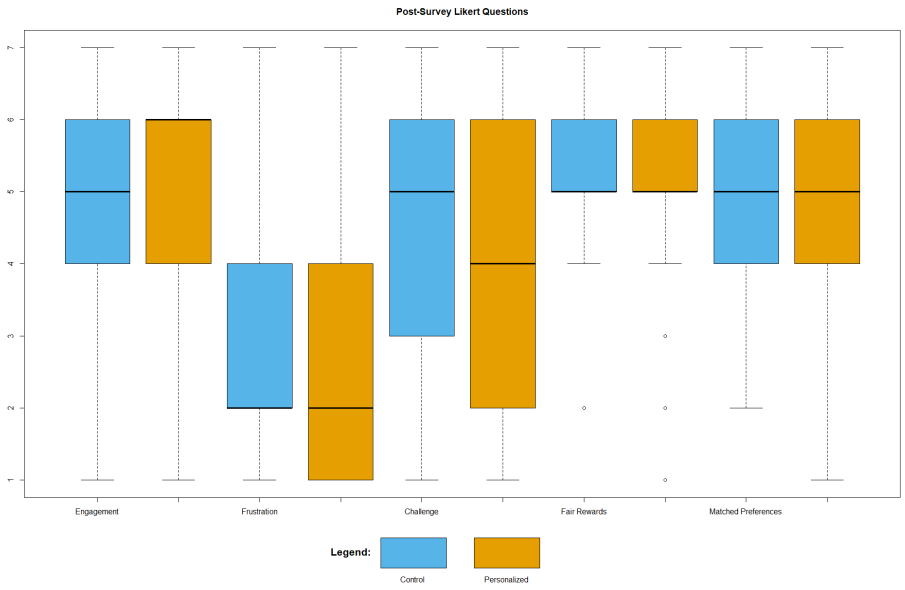

Figure 6.5: Score distributions for the five post-survey Likert questions in the personalized *Café Flour Sack* study. For each question, the data for the *control* condition are shown on the left and the data for the *personalized* condition are shown on the right (in blue and orange respectively).

Table 6.2: Mean scores for the five post-survey Likert questions in the personalized *Café Flour Sack* study.

	Engagement	*Frustration*	*Challenge*	*Fair Rewards*	*Matched Preferences*
Control	5.162	2.865	4.324	5.270	5.162
Personalized	5.00	2.595	3.973	5.216	4.865
Total	5.081	2.730	4.149	5.243	5.014

3. "I found the reward systems **challenging**."

4. "I felt like I was being **fairly rewarded** for the tasks I was completing."

5. "I felt like the rewards I was being given **matched my preferences** for the rewards in the game."

The mean scores for these questions, broken down by condition are broken down by condition in Table 6.2. The score distributions can be observed in Figure 6.5 with the *control* condition on the left and the *personalized* condition on the right for each question. For all five questions, no significant differences were detected between conditions.

Table 6.3: Mean duration (in seconds) of time spent in a single view for all four reward systems across both participant audience type and experimental condition in the *Café Flour Sack* study.

	Leaderboards	*Avatar*	*Narrative*	*Tracker*
Control	8.877	10.717	29.380	7.622
Personalized	9.611	13.379	33.000	11.786

Table 6.4: Choice counts showing whether or not participants selected the round option corresponding to their currently-selected favorite reward in the personalized *Café Flour Sack* study.

	Favorite Selected	*Favorite Not Selected*
Control	130	101
Personalized	125	115
Total	255	216

Duration of Play

Like the previous studies involving *Café Flour Sack*, this study had a fixed duration, so I again focus on where and how players spent their time during the required play time. As Amazon Mechanical Turk workers are incentivized to participate for financial reasons, observing play duration assumes that participants were working to complete tasks as quickly, but adequately as possible. So while the study duration was well within the time limit Amazon Mechanical Turk imposes for completing and submitting task results (before a worker's "lock" on that HIT expires and another worker is allowed to accept it), it was once again insufficient to look at the total duration of play as an indicator of engagement.

Once again, metrics of interest are related to rewards, and specifically how long players spent in each of the different reward systems menus. Table 6.3 shows the mean time spent per view of each reward menu, broken down by experimental condition. Despite the suggestions above that participants spent more time in reward menus in the *personalized* condition, these differences were not significant.

Table 6.5: Choice counts for rounds with a *super task* showing whether or not participants selected the option corresponding to their currently-selected favorite reward in the personalized *Café Flour Sack* study.

	Favorite Selected	*Favorite Not Selected*
Control	23	32
Personalized	29	29
Total	52	61

Choosing Favorite Rewards and Super Tasks

As favorite reward type plays a large role in the experimental conditions of this study, I sought to understand the impact, if any, a participant's current favorite reward type had on their choice of the three options available at the start of each round.

When observing whether or not participants picked round options that corresponded to their currently-selected favorite reward, participants did not always pick the option that corresponded to their favorite. In the *control* condition, participants selected the option with their favorite reward 56% of the time. In the *personalized* condition, participants selected the option with their favorite reward 52% of the time. Table 6.4 demonstrates the breakdown; no significant differences were detected between conditions.

Regarding rounds in which *super tasks* were available, participants in the *control* condition selected the *super task* option 42% of the time. In the *personalized* condition, participants selected the *super task* option 50% of the time. Table 6.5 shows this breakdown; again, no significant differences were detected between conditions.

Altogether, these results suggest that the experimental condition had no effect on whether or not participants picked round options corresponding to their currently-selected favorite reward type.

Changing Favorite Rewards

As participants had the ability to change their current favorite reward at any point during the experiment, one may raise questions such as how frequently did participants change

Table 6.6: Counts of initial choices for participants' favorite reward type when interacting with the favorite reward selection screen during the tutorial of the personalized *Café Flour Sack* study.

	Leaderboards	*Avatar*	*Narratives*
Control	20	13	4
Personalized	17	16	4
Total	37	29	8

their rewards and between which kinds of rewards did these changes occur. The game did not explicitly explain what the purpose of changing one's favorite reward type might have on the game (as this was the primary difference between the two conditions). Thus, participants were forced to infer its purpose, if they chose to interact with the favorite reward selection screen at all.

As part of the tutorial, participants were forced to interact with this screen to pick an initial, favorite reward type from among the three available rewards presented: the leaderboards, the customizable avatar, and the unlockable narratives. This choice occurred after participants were asked to visit all three reward menus to understand how the rewards worked, and thus could be expected to have a minimal understanding of each system. Table 6.6 shows the breakdown of these initial choices, with participants in both conditions overall preferring (in order) the leaderboards, the avatar, and the narratives. No significant differences were observed between the two experimental conditions.

Following this initial interaction, players were not required to change their favorite reward again, but could do so whenever they wished. Altogether, players in the *control* condition utilized this ability a total of 72 times, for a mean 1.946—nearly twice—times per participant. Players in the *personalized* condition, however, utilized this ability a total of 179 times, for a mean 4.838 times per participant, over double that of players in the previous condition. However, these differences in the number of times a favorite reward was changed were not significant.

6.4.2 Objective Metrics—Task Completion

To evaluate the *task completion*, I focus on the same three metrics observed in the previous studies using *Café Flour Sack*: the correctness of the task answers, the number of tasks completed, and the timing of task completion. These metrics reflect the typical considerations of task providers. For an actual human computation task, different metrics might be prioritized over others depending on the task requirements; here however, I present all metrics equally.

Correctness of Completed Tasks

As in previous studies, task correctness for a given task is the ratio of correctly-assigned ingredients to the total number of ingredients in the task, as determined by the gold-standard solution set. A task was considered correct if 75% (a corresponding ratio of 0.75) or more of its ingredients belonged the given recipe. Participants in the *control* condition demonstrated an average mean task correctness of 0.693. Participants in the *personalized* condition demonstrated an average mean task correctness of 0.697. The differences between these two conditions were not significant.

I conducted further analysis using two-way ANOVAs with aligned rank transforms [104], considering factors such as participants' favorite reward types, and whether or not task was part of a *super task* option. However, these factors did not have an effect on task correctness.

Number of Completed Tasks

Additionally, I examined the number of tasks completed per participant between the two experimental conditions. These results were further broken down into three categories: the *total* number of tasks completed, the number of *correct* tasks completed, and the number of *incorrect* tasks completed. When considering total tasks, participants in the *control* condition completed a mean 33.081 tasks, while participants in the *personalized* condition

completed a mean 34.784 tasks. When considering only correct tasks, participants in the *control* condition completed a mean 21.243 correct tasks, while participants in the *personalized* condition completed a mean 22.676 correct tasks. Finally, when considering only incorrect tasks, participants in the *control* condition completed a mean 11.838 incorrect tasks, while participants in the *personalized* condition completed a mean 12.108 incorrect tasks. Overall, participants in the *personalized* condition completed more tasks, but these differences are not significant.

Timing of Completed Tasks

Finally, I focused on the amount of time (in seconds) it took players to complete tasks. As with the observations into the number of tasks completed, these results are broken down into the completion times of *total* tasks, *correct* tasks, and *incorrect* tasks.

When considering total tasks, participants in the *control* condition had a mean completion time of 12.348 seconds, while participants in the *personalized* condition had a mean completion time of 11.789 seconds. When considering only correct tasks, participants in the *control* condition had a mean completion time of 11.440 seconds, while participants in the *personalized* condition had a mean completion time of 10.983 seconds. Lastly, when considering only incorrect tasks, participants in the *control* condition had a mean completion time of 13.977 seconds, while participants in the *personalized* condition had a mean completion time of 13.300 seconds. Overall, participants in the *personalized* condition appear to have slightly faster average times for task completion, but again, these differences are not significant.

6.5 Discussion

Overall, the results of the study show no noticeable differences between the two conditions: *control* and *personalized*. Some potential, but not significant, trends were observed. For subjective (*player experience*) metrics, participants rated aspects of the game such as en-

gagement, reward fairness, and preference matching slightly higher in the *control* condition than in the *personalized* condition. Additionally, *control* participants were more likely to select round options corresponding to their favorite reward. By comparison, participants in the *personalized* condition found the game less frustrating, less challenging, and were more likely to select super task options corresponding to their favorite reward. *Personalized* participants also interacted with the favorite reward selection screen nearly twice as much as *control* participants. For objective (*task completion*) metrics, participants in the *personalized* condition were more correct at solving tasks, completed more tasks, and completed tasks faster than their counterparts in the *control* condition. However, nearly all of these differences were small and these differences were not significant, which ultimately suggests there was no difference between the two conditions.

Taken literally, these results suggest that personalized reward systems are no better than a non-personalized system at improving either *task completion* or *player experience*. An uncharitable reading of the results might even prompt the question of then why even bother with personalized reward systems at all (given that they are typically more time-consuming and resource-intensive to implement, even for an implementation like that in *Café Flour Sack* which does not rely on a complex player modeling system to infer player preferences)?

At the very least, it is worth noting that the personalized reward systems in *Café Flour Sack* demonstrated no negative effects on *task completion*, which is often a concern for such systems. The results show that participants in the *personalized* condition did in fact use the favorite reward selection screen more frequently; in spite of this interaction arguably taking away time from solving tasks, the *task completion* results from *personalized* participants trended higher. Further investigation would be required to see if these results might in fact still hold across a larger or different audience of participants.

I instead propose that rather than dismissing personalized reward systems outright, that these results might have been a consequence of the variation between the two conditions being too subtle. One particularly telling, though not statistically-significant, observation

comes from the post-game survey. In the *control* condition, participants rated the statement "I felt like the rewards I was being given matched my preferences for the rewards in the game" higher than participants in the *personalized* version. This result, notably the fact that the *control* version of *Café Flour Sack* was not adapting to participants' favorite rewards (i.e., deliberately giving them the opposite) suggests that the experimental condition may have been too subtle for this particular study, especially when compared to the previous *Café Flour Sack* study in which participants in the *choice* condition did perceive they had more choice compared to participants in the *random condition*.

So why might this particular implementation of providing personalized rewards for more work-intensive tasks have been too subtle to demonstrate a noticeable effect?

For this personalized study, the only visual difference between the two versions of *Café Flour Sack* was which reward option the *super task* was assigned to. In-game, this change occurred once every three rounds when the *super task* was present and visually changed the set of round option buttons the participants saw (i.e., Figure 6.1). While I consider the single screen difference between the *random* and *choice* conditions from the previous *Café Flour Sack* study to be similarly subtle, this difference was exposed every round—thus three times more frequently than the experimental condition with the *super task*.

Another potential issue may have been the short study duration, which is both a consequence of its deployment to Amazon Mechanical Turk and feedback from the previous study, in which some participants expressed free-form feedback that the play duration was too long. Many HCGs solving classification tasks, like the one utilized in *Café Flour Sack*, are set up as puzzle games with short play durations, particularly when compared against optimization tasks (e.g., consider *Foldit* wherein players typically complete multiple tutorials before being given an actual protein folding problem, thereby giving them a longer period of time to become comfortable interacting with the game). While this makes *Café Flour Sack* similar to existing classification HCGs, it is possible that personalized reward systems (which benefit from extended play duration) may not be as impactful in games

where the player is not intended to repeat play sessions or interact for extended periods of time.

In combination, the limited exposure of the experimental condition with the change to the reward type of the *super task* in combination with short duration of study (again compared to the previous *Café Flour Sack* studies) could have resulted in the variation being too subtle to create any noticeable effect between the two versions. However, these results do not completely invalidate personalized reward systems as designed in *Café Flour Sack*. Potential follow-up explorations to this particular investigation could provide *super tasks* more frequently.[9] Alternatively, one might consider looking at HCGs with longer tasks, particularly (scientific) data optimization tasks (which often require more thorough training or tutorialization, thereby resulting in more interaction with game systems). The benefit of examining or using HCGs with more involved or time-consuming gameplay sessions would also permit the use of more complex player typologies or player modeling systems, which require extended periods or multiple sessions of player interaction with the game in order to properly understand player preferences.

Finally, I wish to address the role that participant audience may have played in the study results. As the previous study with *Café Flour Sack* demonstrated, audience does in fact have an impact on the results. For this personalized *Café Flour Sack* study, I chose to utilize workers from Amazon Mechanical Turk based on the results of the previous *Café Flour Sack* studies, in which Amazon Mechanical Turk participants demonstrated high *task completion* results in line with prior research such of that of Sabou *et al*. [103]. Such prior work does suggest the interchangeability of paid crowdsourcing platforms and HCGs, particularly when it comes to *task completion* metrics, although questions still remain around how motivation for monetary compensation may affect motivation for play and engagement.

One observation from this particular study was that there was a high percentage of participants who attempted to repeat the study using different Amazon Mechanical Turk user

[9]Such a change would necessitate a sufficiently larger amount of reward content than *Café Flour Sack* currently provides.

accounts. While I was able to detect these instances, 20 of the logged 94 participants were detected to be duplicates, suggesting that around 21% of the participants did attempt to re-complete the study despite the instructions explicitly asking otherwise.[10] By comparison, such an incident was only detected once during the original *Café Flour Sack* study. This would suggest that monetary compensation was an overwhelmingly strong motivator for participation (again in line with concerns that Sabou *et al.* highlighted in their work, but did not fully explore). [11] While it is unclear what the exact effect of overwhelming extrinsic motivation for monetary compensation may have had on the results, I hypothesize that it may have at the very least, distracted or taken away from extrinsic motivations for in-game rewards. Ultimately, I consider these issues around player audience a good cautionary takeaway for running HCG research on paid crowdsourcing platforms and future work is needed to truly understand just how extrinsic motivation for monetary compensation may interact (or worse, interfere) with extrinsic motivation for in-game rewards.

6.6 Conclusions

In this chapter, I describe a modified version of the *Café Flour Sack* human computation game and its use in a study of personalized reward systems. The study compared two versions of *Café Flour Sack*: a *control* version which did not personalize players' preferred rewards for more involved tasks *super tasks* and a *personalized* version which always assigned players' preferred rewards for more involved tasks. I finish with a discussion of the study results and what factors may have contributed to the lack of difference between the conditions, as well as some potential future directions for work.

This chapter began with the question of understanding how using personalization based

[10] Anecdotally, I noticed that this problem appeared to be more exacerbated during certain times of the day when releasing batches of HITs on the Amazon Mechanical Turk platform (as I made concerted efforts to distribute tasks at different times of the day to account for a more global audience of Amazon Mechanical Turk workers). I opted **not** to deliberately avoid these particular times of day to avoid biasing player audience selection, and instead focused on detecting and removing duplicates instead.

[11] I also acknowledge that this study was conducted during a global pandemic, during which job loss was global concern and may have inflated the motivation for monetary compensation.

on player preferences in reward systems would affect *task completion* and *player experience* metrics. The results of the study suggest personalization had no significant effects on either the *task completion* and *player experience*. While the results trended towards higher *task completion* metrics for the *personalized* condition with mixed (split) results for *player experience* metrics between, none of these results were statistically significant. These results suggest that preference-focused, personalized rewards systems may have no difference on *task completion* metrics, primarily in that they do not appear to adversely affect or distract from task solutions. However, it is not the case that preference-focused, personalized reward systems generate more positive *player experience* metrics.

While this might suggest that personalized reward systems may not be as effective as initially hypothesized or do not appear to be a useful avenue of exploration for HCGs, I emphasize that this is not the case. It is true that personalized reward systems involve more initial work to implement, but the study with this personalized version of *Café Flour Sack* is just the first exploration of personalization in the context of reward systems. Given the complex role that rewards have regarding player feedback, player motivation, and fair compensation for human computation work, it is possible that other aspects of reward systems—to name a few: reward function (i.e., for what in-game actions are rewards given), reward distribution (i.e., how in-game rewards are distributed), and other unexplored reward types (e.g., badges, in-game powerups, intrinsic rewards)—may have larger effects. It may also be the case that personalization and adaptive techniques may be more effective (with regards to both *task completion* and *player experience*) for different tasks (particularly HCGs with tasks necessitating longer or repeated play sessions which could take better advantage of more complex player modeling techniques) or different combinations of game elements.

Ultimately, very little is known about how personalization and adaptive techniques may play a role in HCGs. However, as player modeling and techniques for inferring player preferences become more sophisticated (in combination with growing support for logging and understanding player data), more and more mainstream, digital games are looking at

personalization and adaptive techniques as a way to keep players engaged with their experiences. Additionally, personalization and adaptive techniques enable systems where time-consuming, rote tasks such as tuning parameters or tweaking content can automatically assist with this process, which is of particular benefit to HCG developers who do not have the time or resources of industry-scale studios. In order for HCGs to remain competitive with the growing number of games that rely on personalization and adaptive techniques, HCGs must keep up and explore new ways to make their games more effective. This chapter is only the start of what may hopefully become future investigations into personalization and adaptive techniques, not just for reward systems, for HCG mechanics and game elements altogether.

CHAPTER 7

CONCLUSIONS

From classifying cat pictures to folding protein configurations to sketching silly doodles, human computation games have been used to provide playful solutions to complex problems that remain difficult and intractable for computers solve. Yet in spite of these successes, making games remains a complicated, time-consuming, and multidisciplinary process, one which is not always straightforward or possible for scientists, researchers, and task providers to undertake. Making an HCG comes with even more complications, the most noteworthy of which is that HCGs have two primary, often contradictory, goals. The first goal is to solve some underlying human computation *task*, which may sometimes be mundane, ambiguous, and in many cases, difficult to translate to engaging game mechanics. The second goal is to provide an engaging player experience, since without players, an HCG is incapable of generating sufficient solutions to the underlying task. At best, the game *mechanics* of an HCG work towards both of these goals without interference to create an engaging, effective experience. However, at worst, an HCG is either an unenaging, insipid experience optimized wholly around the task—one which players will not play—or an experience so engaging that players are too distracted by other elements—one which does not solve the task.

Given that most HCG developers are experts at the task they wish to solve rather than professional game developers, many HCGs are task-focused experiences that lack the engagement necessary to sustain these games beyond the initial acquisition of results. This creates an ominous prospect in a games industry that where every year, more and more games become available for potential players to interact, where HCGs must compete for the time and attention of players who may choose to avoid HCGs entirely given their lackluster reputation for engagement. While some HCGs [20, 4, 25] have managed to elevate

themselves into mature platforms and others have managed to foster successful collaborations with large-scale, big-budget games [24, 30], not all potential HCGs may be so fortunate. Design knowledge specific to HCGs remains limited—at best anecdotal—with few generalizable guiding principles, thus resulting in games that mimic old patterns and risk failing to engage newer player audiences. Realistically, the average task provider may look at this current state of HCG design and ask why they should risk a lengthy, complicated development process to end up with a game that may not even be effective or engaging. As a result, HCGs continue to remain a relatively unexplored area of design study, even when compared with other serious game domains like education and training.

In this thesis, I address a specific subset of human computation game mechanics: *reward* mechanics. Reward mechanics are the game mechanics responsible for providing players feedback about their effort in solving tasks, typically in the form of in-game digital rewards or catering to intrinsic motivation for participating in the crowdsourcing process. Given that HCGs typically do not compensate their players monetarily compared with paid, online crowdsourcing platforms (e.g., Amazon Mechanical Turk, Crowdflower), *reward* mechanics are very important to HCGs for their role in compensation and providing players an engaging experience in gratitude for their participation in the human computation process.

In Chapter 3, I propose a framework for understanding and visualizing the *mechanics* of human computation games. This framework breaks down HCGs into three types of mechanics: *action*, *verification*, and *feedback* (rewards), which align the core gameplay loop with the steps of the human computation process. I propose this framework alongside a methodology for evaluating variations in HCG mechanics that considers the impact of variations on both the *player experience* and the *task completion*—the dual design goals which HCGs must optimize for. The descriptions of this framework and methodology are illustrated using examples of existing, successful HCGs; throughout subsequent chapters, I use both framework and methodology to contextualize and design controlled experiments.

In Chapter 4, I begin investigations into *reward* mechanics using a between-subjects study comparing collaborative and competitive reward functions. This study simultaneously examined singleplayer and multiplayer game mechanics, and was accomplished using an HCG developed specifically for this purpose: *Gwario*, a *Super Mario Bros.*-inspired platformer that uses a commonsense knowledge task to test design hypotheses. Additionally, I report on the results of a survey sent to HCG experts (i.e., developers) to understand expert opinions on the variations on the game mechanics explored in the study. The study and interviews show, among many results, that *collaborative* reward mechanics result in higher *task completion* metrics, while *competitive* reward mechanics yield higher *player experience* metrics and that *collusion* between players enabled higher task accuracy.

In Chapter 5, I continue investigations into *reward* mechanics, this time looking at multiple reward types and the impact of offering players a *choice* of reward using a between-subjects study. The study also examined different HCG audiences by comparing the results between a population of crowdsourcing experts and a population of student players. Additionally, I report on the results of a series of semi-structured interviews conducted with a student player population to better understand multiple reward systems. The study shows, among other results, that *choice* of reward demonstrated benefits to both *player experience* and *task completion* metrics.

In Chapter 6, I end with investigations into personalized reward systems, specifically what impact personalizing the type of reward a player receives might have on more effort-intensive tasks. The study results using a personalized reward system are inconclusive; personalized reward systems did not show any significant negative effects on *task completion* compared with non-personalized reward systems, but showed no benefits to *player experience* either.

Taken altogether, these results show that variations in HCG reward systems do have an impact on *player experience* and *task completion*. Specifically, I show that various properties of reward systems can in fact, improve the *player experience* without compromising

task completion metrics such as solution quality, such as *offering players choice of reward type*—and may occasionally improve *task completion* metrics, too. In other cases, there are nuances, such as in the case of *collaborative* and *competitive* reward systems. Prior work [70, 73] has shown the inclusion of *competitive* reward systems results in no difference in *task completion* metrics, but the results of Chapter 4 suggest that *competitive* reward mechanics resulted in lower task accuracy than the *collaborative* version of those reward mechanics.

In the process of these investigations into reward systems, I also explored a number of other variations not simply limited to *feedback* mechanics. Chapter 3 provides an example of how to adapt an HCG to a platformer game (i.e., exploring platformer *action* mechanics), a rarely-explored avenue of HCG design (although one of increasing interest given how many recent HCGs [24, 30, 31] have been integrated into large-scale, entertainment-oriented games not originally intended for human computation). That chapter also demonstrates results (from both the study and HCG expert interviews) that contradict contradict existing HCG design wisdom [50, 90], such as how *allowing players to collude or communicate* proved to be a predictor of task accuracy. Such results suggest that there is in fact a need to continue updating HCG design knowledge and not to blindly rely on existing (and often dated) design tenets. Additionally, Chapter 5 looks into the unexplored area of player audience, showing that different audience do respond differently to variations in reward mechanics. For example, by *offering players choice of reward type*, student players performed closer to expert crowdsourcers at metrics such as task correctness and rate of task completion, compared with *offering random reward types* where the differences between the two audiences were far greater.

Ultimately, all of these results—both rewards mechanics and other game aspects—reiterate that human computation game design knowledge is not and should not be a static handful of old design patterns and generalizations based on anecdotes. HCG design knowledge can and should be based on empirically testing and evaluating variations of game

elements when possible while considering the effects on both *player experience* and *task completion* metrics. By understanding how the inclusion of certain game elements in various HCG contexts work, HCG developers will have better information about how to build these games in a way that gives them more confidence that their games will be both engaging and effective.

Finally, this thesis only addresses one aspect of HCG mechanics in the context of only several kinds of human computation tasks. In the context of reward mechanics specifically, there is plenty of room in the future to explore other aspects of reward mechanics such as different reward types not explored in this thesis (e.g., badges, other forms of feedback for intrinsic motivation), across a wider variety of tasks. Specifically, this dissertation focused on classification and commonsense knowledge tasks; how these results might generalize to data optimization and scientific discovery tasks is an open question. Furthermore, some investigations explored in this thesis, such as personalized reward systems, remain inconclusive.

The future remains kind to the need for human computation games. Artificial intelligence and machine learning become more ubiquitous every year, and will continue to require data sets and information where humans are currently the optimal solution solvers or creators. Scientific optimization and citizen science efforts will continue to encounter new scenarios—diseases, novel protein configurations, and untackled DNA sequences, to name a few—that humans remain the experts at navigating. Games become more and more accessible across existing and novel platforms, and the tools to build these experiences will be more readily available for novices and non-experts to use. With all of these aspects in their favor, HCGs do not deserve to fade into obscurity when there remain (and will likely always be) computationally-intractable problems and the digital tools to build the experiences to tackle them. But only by giving HCG developers and task providers the best, most rigorous design knowledge about how to build these games to address both the needs of tasks and players can these games reach their true potential.

Appendices

APPENDIX A
FULL SCRIPT FOR THE INTERVIEW TO THE *CAFÉ FLOUR SACK* FOLLOWUP STUDY

The following text was printed on paper and kept on hand during the interview with participants. Handheld recording began with the interviewer greeting and ended with the closing; participants' questions following the closing were not recorded or considered part of the interview.

Interviewer Greeting: Thank you for agreeing to conduct this interview with me. I'd like to ask you a couple of questions about the experience you had. If you have any feedback for us, there will be time at the end of the interview to ask.

A.1 General Questions about Rewards in Games

1. What kinds of rewards to you enjoy in games?
    (a) *(If rewards given)* What motivates you to interact with these rewards? (Alternative phrasing: Why do you like these rewards?)
2. Have you ever felt like certain rewards made you less motivated to play a game?

A.2 Human Computation Games

3. Do you play human computation games? Have you ever tried one?
    (a) *(If positive response given)* Can you describe what motivates you to play these games?

(b) *(If neutral or negative response given)* Can you explain or describe why not? What might motivate you to play one of these games?

A.3 Rewards in the Context of the Study's HCG (*Café Flour Sack*)

4. How do you think the rewards in the HCG you just played (*Café Flour Sack*) compared to rewards in other games?

5. What was your favorite reward system in the HCG you just played (*Café Flour Sack*) and why?

6. What was your least favorite reward system in the HCG you just played (*Café Flour Sack*) and why?

7. *(Optional, as this may be answered by prior questions)* Can you give me a preference ranking of the reward systems in *Café Flour Sack*? *(For each, ask why if possible.)*

A.4 Future Rewards

8. What kinds of rewards would you like to have seen in a human computation game (including this one in the study)?

   (a) *(If positive response given)* Are these the kinds of rewards that you regularly interact with in games that you frequently play or enjoy?

   (b) *(If positive or neutral response given)* Are there any other rewards that you think could be implemented in a human computation game that weren't available here?

Interviewer Closing: Finally, do you have any other feedback for me? This is also an opportunity to ask me any additional questions you have about the study.

REFERENCES

[1] E. Law and L. von Ahn, "Human computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121, 2011.

[2] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04, Vienna, Austria: ACM, 2004, pp. 319–326, ISBN: 1-58113-702-8.

[3] E. Law and L. von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09, Boston, MA, USA: ACM, 2009, pp. 1197–1206, ISBN: 978-1-60558-246-7.

[4] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, *et al.*, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, no. 7307, pp. 756–760, 2010.

[5] J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Yoon, A. Treuille, R. Das, and E. Participants, "RNA design rules from a massive open laboratory," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2122–2127, 2014.

[6] K. Tuite, N. Snavely, D.-y. Hsiao, N. Tabing, and Z. Popović, "Photocity: Training experts at large-scale image acquisition through a competitive game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, Vancouver, BC, Canada: ACM, 2011, pp. 1383–1392, ISBN: 978-1-4503-0228-9.

[7] M. Bell, S. Reeves, B. Brown, S. Sherwood, D. MacMillan, J. Ferguson, and M. Chalmers, "Eyespy: Supporting navigation through play," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09, Boston, MA, USA: ACM, 2009, pp. 123–132, ISBN: 978-1-60558-246-7.

[8] A. Anthropy and N. Clark, *A Game Design Vocabulary*. Addison-Wesley, 2014.

[9] S. Cooper, A. Treuille, J. Barbero, A. Leaver-Fay, K. Tuite, F. Khatib, A. C. Snyder, M. Beenen, D. Salesin, D. Baker, and Z. Popović, "The challenge of designing scientific discovery games," in *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, ser. FDG '10, Monterey, California: ACM, 2010, pp. 40–47, ISBN: 978-1-60558-937-4.

[10] P. Lessel, M. Altmeyer, L. V. Schmeer, and A. Krüger, ""enable or disable gamification?": Analyzing the impact of choice in a gamified image tagging task," in

Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12, ISBN: 9781450359702.

[11] W. Mason and D. J. Watts, "Financial incentives and the "performance of crowds"," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, ser. HCOMP '09, Paris, France: ACM, 2009, pp. 77–85, ISBN: 978-1-60558-672-4.

[12] K. Schrier, "What's in a name?: Naming games that solve real-world problems," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, ser. FDG '17, Hyannis, Massachusetts: ACM, 2017, 50:1–50:4, ISBN: 978-1-4503-5319-9.

[13] M. Krause and J. Smeddinck, "Human computation games: A survey," in *2011 19th European Signal Processing Conference*, IEEE, Aug. 2011, pp. 754–758.

[14] S. Thaler, K. Siorpaes, E. Simperl, and C. Hofer, "A survey on games for knowledge acquisition," *Rapport technique, STI*, p. 26, 2011.

[15] E. P. P. Pe-Than, D. H.-L. Goh, and C. S. Lee, "A typology of human computation games: An analysis and a review of current games," *Behaviour & Information Technology*, 2013.

[16] L. von Ahn, R. Liu, and M. Blum, "Peekaboom: A game for locating objects in images," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '06, Montréal, Québec, Canada: ACM, 2006, pp. 55–64, ISBN: 1-59593-372-7.

[17] S. Hacker and L. von Ahn, "Matchin: Eliciting user preferences with an online game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09, Boston, MA, USA: ACM, 2009, pp. 1207–1216, ISBN: 978-1-60558-246-7.

[18] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, 2008.

[19] L. Barrington, D. Turnbull, and G. Lanckriet, "Game-powered machine learning," *Proceedings of the National Academy of Sciences*, vol. 109, no. 17, pp. 6411–6416, 2012. eprint: http://www.pnas.org/content/109/17/6411.full.pdf.

[20] R. Simpson, K. R. Page, and D. De Roure, "Zooniverse: Observing the world's largest citizen science platform," in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 1049–1054.

[21] K. Siorpaes and M. Hepp, "Games with a purpose for the semantic web," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 50–60, 2008.

[22] M. Krause, A. Takhtamysheva, M. Wittstock, and R. Malaka, "Frontiers of a paradigm: Exploring human computation with digital games," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, ser. HCOMP '10, Washington DC: ACM, 2010, pp. 22–25, ISBN: 978-1-4503-0222-7.

[23] R. Hodhod, M. Huet, and M. Riedl, "Toward generating 3D games with the help of commonsense knowledge and the crowd," in *Experimental AI in Games Workshop*, 2014.

[24] D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, and E. Lundberg, "Deep learning is combined with massive-scale citizen science to improve large-scale image classification," *Nature Biotechnology*, vol. 36, no. 9, pp. 820–828, 2018.

[25] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, J. Waldispühl, and P. Players, "Phylo: A citizen science approach for improving multiple sequence alignment," *PLoS One*, vol. 7, no. 3, e31362, 2012.

[26] G. Mehta, C. Crawford, X. Luo, N. Parde, K. Patel, B. Rodgers, A. K. Sistla, A. Yadav, and M. Reisner, "Untangled: A game environment for discovery of creative mapping strategies," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 6, no. 3, 13:1–13:26, Oct. 2013.

[27] O. Tremblay-Savard, A. Butyaev, and J. Waldispühl, "Collaborative solving in a human computing game using a market, skills and challenges," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '16, Austin, Texas, USA: Association for Computing Machinery, 2016, pp. 130–141, ISBN: 9781450344562.

[28] J. Roskams and Z. Popović, "Power to the people: Addressing big data challenges in neuroscience by creating a new cadre of citizen neuroscientists," *Neuron*, vol. 92, pp. 658–664, 2016.

[29] A. Singh, F. Ahsan, M. Blanchette, and J. Waldispühl, "Lessons from an online massive genomics computer game," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 5, 2017.

[30] *Play borderlands science today!* https://borderlands.com/en-US/news/2020-04-07-borderlands-science/, 2020.

[31] C. Clark, I. Greenberg, and M. Ouellette, "A model for integrating human computing into commercial video games," in *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, 2018.

[32] K. Tuite, R. Banerjee, N. Snavely, J. Popović, and Z. Popović, "Pointcraft: Harnessing players' fps skills to interactively trace point clouds in 3d," in *10th International Conference on the Foundations of Digital Games*, 2015.

[33] Y.-l. Kuo, J.-C. Lee, K.-y. Chiang, R. Wang, E. Shen, C.-w. Chan, and J. Y.-j. Hsu, "Community-based game design: Experiments on social games for commonsense data collection," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, ser. HCOMP '09, Paris, France: ACM, 2009, pp. 15–22, ISBN: 978-1-60558-672-4.

[34] *Quick, draw!* https://quickdraw.withgoogle.com/, Accessed: 2020-06-06.

[35] D. Ha and D. Eck, "A neural representation of sketch drawings," *CoRR*, vol. abs/1704.03477, 2017. arXiv: 1704.03477.

[36] C.-H. Yu, Z. Qin, F. J. Martin-Martinez, and M. J. Buehler, "A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using artificial intelligence," *ACS nano*, vol. 13, no. 7, pp. 7471–7482, 2019.

[37] K. Siu, A. Zook, and M. O. Riedl, "A framework for exploring and evaluating mechanics in human computation games," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, ser. FDG '17, Hyannis, Massachusetts: ACM, 2017, 38:1–38:4, ISBN: 978-1-4503-5319-9.

[38] ——, "A framework for exploring and evaluating mechanics in human computation games," *CoRR*, vol. abs/1706.03311, 2017. arXiv: 1706.03311.

[39] K. Salen and E. Zimmerman, *Rules of Play: Game Design Fundamentals*. Cambridge Mass.: MIT Press, 2003.

[40] T. Fullerton, C. Swain, and S. Hoffman, *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*. Morgan Kaufmann, 2008.

[41] J. Schell, *The Art of Game Design: A Book of Lenses*. Elsevier/Morgan Kaufmann, 2008.

[42] S. Björk and J. Holopainen, *Patterns in Game Design (Game Development Series)*. Charles River Media, 2004.

[43] N. Falstein. "The 400 project." (2006).

[44] *Gamasutra*, https://www.gamasutra.com/, Accessed: 2018-01-01.

[45] *Game developers conference vault*, https://www.gdcvault.com/, Accessed: 2018-01-01.

[46] M. Seif El-Nasr, B. Aghabeigi, D. Milam, M. Erfani, B. Lameman, H. Maygoli, and S. Mah, "Understanding and evaluating cooperative games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10, Atlanta, Georgia, USA: ACM, 2010, pp. 253–262, ISBN: 978-1-60558-929-9.

[47] K. Hullett, "The science of level design: Design patterns and analysis of player behavior in first-person shooter levels," Ph.D. dissertation, University of California, Santa Cruz, 2012.

[48] D. R. Michael and S. L. Chen, *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade, 2005, ISBN: 1592006221.

[49] U. Ritterfeld, M. Cody, and P. Vorderer, *Serious Games: Mechanisms and Effects*. Routledge, 2009.

[50] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[51] J. Carranza and M. Krause, "Evaluation of game designs for human computation," in *Proceedings of the 2012 AAAI Workshop on Human Computation in Digital Games and Artificial Intelligence for Serious Games*, 2012.

[52] R. Hunicke, M. LeBlanc, and R. Zubek, "Mda: A formal approach to game design and game research," in *Proceedings of the AAAI Workshop on Challenges in Game AI*, vol. 4, 2004, p. 1.

[53] J. A. Miller, U. Narayan, M. Hantsbarger, S. Cooper, and M. S. El-Nasr, "Expertise and engagement: Re-designing citizen science games with players' minds in mind," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, 2019, pp. 1–11.

[54] P. Jamieson, L. Grace, and J. Hall, "Research directions for pushing harnessing human computation to mainstream video games," in *Meaningful Play*, 2012.

[55] K. Tuite, "GWAPs: Games with a problem," in *9th International Conference on the Foundations of Digital Games*, 2014.

[56] E. Cambria, T. V. Nguyen, B. Cheng, K. Kwok, and J. Sepulveda, "GECKA3D: A 3d game engine for commonsense knowledge acquisition," *CoRR*, vol. abs/1602.01178, 2016. arXiv: 1602.01178.

[57] *Mmos*, http://mmos.ch/, Accessed: 2020-12-30.

[58] K. Isbister and N. Schaffer, *Game Usability: Advancing the Player Experience*. Morgan Kaufmann, 2008.

[59] M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds., *Game Analytics*. Springer London, 2013.

[60] G. Wallner and S. Kriglstein, "Visualization-based analysis of gameplay data – a review of literature," *Entertainment Computing*, vol. 4, no. 3, pp. 143–155, 2013.

[61] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger, "Optimizing challenge in an educational game using large-scale design experiments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13, Paris, France: ACM, 2013, pp. 89–98, ISBN: 978-1-4503-1899-0.

[62] L. A. Whitlock, A. C. McLaughlin, W. Leidheiser, M. Gandy, and J. C. Allaire, "Know before you go: Feelings of flow for older players depends on game and player characteristics," in *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, ser. CHI PLAY '14, Toronto, Ontario, Canada: ACM, 2014, pp. 277–286, ISBN: 978-1-4503-3014-5.

[63] M. Lankes, T. Mirlacher, S. Wagner, and W. Hochleitner, "Whom are you looking for?: The effects of different player representation relations on the presence in gaze-based games," in *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, ser. CHI PLAY '14, Toronto, Ontario, Canada: ACM, 2014, pp. 171–179, ISBN: 978-1-4503-3014-5.

[64] M. W. McEwan, A. L. Blackler, D. M. Johnson, and P. A. Wyeth, "Natural mapping and intuitive interaction in videogames," in *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, ser. CHI PLAY '14, Toronto, Ontario, Canada: ACM, 2014, pp. 191–200, ISBN: 978-1-4503-3014-5.

[65] E. Andersen, Y.-E. Liu, R. Snider, R. Szeto, S. Cooper, and Z. Popović, "On the harmfulness of secondary game objectives," in *Proceedings of the 6th International Conference on Foundations of Digital Games*, ser. FDG '11, Bordeaux, France: ACM, 2011, pp. 30–37, ISBN: 978-1-4503-0804-5.

[66] E. Andersen, E. O'Rourke, Y.-E. Liu, R. Snider, J. Lowdermilk, D. Truong, S. Cooper, and Z. Popović, "The impact of tutorials on games of varying complexity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, Austin, Texas, USA: ACM, 2012, pp. 59–68, ISBN: 978-1-4503-1015-4.

[67] Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popović, "Trading off scientific knowledge and user learning with multi-armed bandits," in *Educational Data Mining*, 2014.

[68] A. Shannon, A. Boyce, C. Gadwal, and T. Barnes, "Effective practices in game tutorial systems," in *8th International Conference on the Foundations of Digital Games*, 2013.

[69] C.-J. Ho, T.-H. Chang, J.-C. Lee, J. Y.-j. Hsu, and K.-T. Chen, "Kisskissban: A competitive human computation game for image annotation," 1, vol. 12, New York, NY, USA: ACM, Nov. 2010, pp. 21–24.

[70] D. H.-L. Goh, R. P. Ang, C. S. Lee, and A. Y. K. Chua, "Fight or unite: Investigating game genres for image tagging," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 7, pp. 1311–1324, Jul. 2011.

[71] M. Sicart, "Defining game mechanics," *The International Journal of Computer Game Research*, vol. 8, no. 2, Dec. 2008.

[72] E. Adams and J. Dormans, *Games Mechanics Advanced Game Design*. New Riders Games, 2012.

[73] K. Siu, A. Zook, and M. O. Riedl, "Collaboration versus competition: Design and evaluation of mechanics for games with a purpose," in *9th International Conference on the Foundations of Digital Games*, 2014.

[74] K. Siu, M. Guzdial, and M. O. Riedl, "Evaluating singleplayer and multiplayer in human computation games," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, ser. FDG '17, Hyannis, Massachusetts: ACM, 2017, 34:1–34:10, ISBN: 978-1-4503-5319-9.

[75] K. Murayama, M. Matsumoto, K. Izuma, and K. Matsumoto, "Neural basis of the undermining effect of monetary reward on intrinsic motivation," *Proceedings of the National Academy of Sciences*, vol. 107, no. 49, pp. 20 911–20 916, 2010.

[76] E. Law, M. Yin, J. Goh, K. Chen, M. A. Terry, and K. Z. Gajos, "Curiosity killed the cat, but makes crowdwork better," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16, Santa Clara, California, USA: ACM, 2016, pp. 4098–4110, ISBN: 978-1-4503-3362-7.

[77] D. H.-L. Goh, E. P. P. Pe-Than, and C. S. Lee, "An investigation of reward systems in human computation games," in *Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II*, M. Kurosu, Ed. Springer International Publishing, 2015, pp. 596–607, ISBN: 978-3-319-20916-6.

[78] K. Siu and M. O. Riedl, "Reward systems in human computation games," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '16, Austin, Texas, USA: ACM, 2016, pp. 266–275, ISBN: 978-1-4503-4456-2.

[79] J. Gaston and S. Cooper, "To three or not to three: Improving human computation game onboarding with a three-star system," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17, Denver, Colorado, USA: ACM, 2017, pp. 5034–5039, ISBN: 978-1-4503-4655-9.

[80] W. Peng and G. Hsieh, "The influence of competition, cooperation, and player relationship in a motor performance centered computer game," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2100–2106, 2012.

[81] J. L. Plass, P. A. O'keefe, B. D. Homer, J. Case, E. O. Hayward, M. Stein, and K. Perlin, "The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation.," *Journal of Educational Psychology*, vol. 105, no. 4, p. 1050, 2013.

[82] K. Emmerich and M. Masuch, "Helping friends or fighting foes: The influence of collaboration and competition on player experience.," in *Proceedings of the Eighth International Conference on the Foundations of Digital Games*, ACM, 2013, pp. 150–157.

[83] R. R. Wehbe and L. E. Nacke, "Towards understanding the importance of co-located gameplay," in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '15, London, United Kingdom: ACM, 2015, pp. 733–738, ISBN: 978-1-4503-3466-2.

[84] R. L. Mandryk and K. M. Inkpen, "Physiological indicators for the evaluation of co-located collaborative play," in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '04, Chicago, Illinois, USA: ACM, 2004, pp. 102–111, ISBN: 1-58113-810-5.

[85] F. Ke and B. L. Grabowski, "Gameplaying for maths learning: Cooperative or not?" *British Journal of Educational Technology*, vol. 38, pp. 249–259, 2007.

[86] K. Siu, M. Guzdial, and M. O. Riedl, "Evaluating singleplayer and multiplayer in human computation games," *CoRR*, vol. abs/1703.00818, 2017. arXiv: 1703.00818.

[87] M. Peplow, "Citizen science lures gamers into sweden's human protein atlas," *Nature Biotechnology*, vol. 34, no. 5, pp. 452–452, 2016.

[88] J. Togelius, S. Karakovskiy, and N. Shaker, *2012 mario ai championship*, http://www.marioai.org/, 2012.

[89] M. Guzdial and M. Riedl, "Conceptually blended levels in a unity engine," in *Playable Experiences at Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2016.

[90] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?: Screening mechanical turk workers," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10, Atlanta, Georgia, USA: ACM, 2010, pp. 2399–2402, ISBN: 978-1-60558-929-9.

[91] N. Salleh, E. Mendes, and J. Grundy, "Empirical studies of pair programming for cs/se teaching in higher education: A systematic literature review," *IEEE Transactions on Software Engineering*, vol. 37, no. 4, pp. 509–525, 2011.

[92] N. Yee, "Motivations for play in online games," *CyberPsychology & behavior*, vol. 9, no. 6, pp. 772–775, 2006.

[93] R. Bartle, "Hearts, clubs, diamonds, spades: Players who suit muds," *Journal of MUD research*, vol. 1, no. 1, p. 19, 1996.

[94] R. D. Laws, *Robin's laws of good game mastering*. Steve Jackson Games, 2002.

[95] D. Choi and J. Kim, "Why people continue to play online games: In search of critical design factors to increase customer loyalty to online contents," *CyberPsychology & behavior*, vol. 7, no. 1, pp. 11–24, 2004.

[96] L. E. Nacke, C. Bateman, and R. L. Mandryk, "Brainhex: A neurobiological gamer typology survey," *Entertainment Computing*, vol. 5, no. 1, pp. 55–62, 2014.

[97] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke, "The gamification user types hexad scale," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '16, Austin, Texas, USA: ACM, 2016, pp. 229–243, ISBN: 978-1-4503-4456-2.

[98] A. K. Przybylski, C. S. Rigby, and R. M. Ryan, "A motivational model of video game engagement.," *Review of general psychology*, vol. 14, no. 2, p. 154, 2010.

[99] G. Richter, D. R. Raban, and S. Rafaeli, "Gamification in education and business," in T. Reiners and C. L. Wood, Eds. Springer International Publishing, 2015,

ch. Studying Gamification: The Effect of Rewards and Incentives on Motivation, pp. 21–46.

[100] C. S. Lee, D. H.-L. Goh, A. Y. Chua, and R. P. Ang, "Indagator: Investigating perceived gratifications of an application that blends mobile content sharing with gameplay," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 6, pp. 1244–1257, 2010.

[101] B. Magerko, C. Heeter, and B. Medler, "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*, p. 255, 2010.

[102] S. Thaler, E. Simperl, and S. Wolger, "An experiment in comparing human-computation techniques," *IEEE Internet Computing*, vol. 16, no. 5, pp. 52–58, Sep. 2012.

[103] M. Sabou, K. Bontcheva, A. Scharl, and M. Föls, "Games with a purpose or mechanised labour?: A comparative study," in *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, ser. i-Know '13, Graz, Austria: ACM, 2013, 19:1–19:8, ISBN: 978-1-4503-2300-0.

[104] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, Vancouver, BC, Canada: ACM, 2011, pp. 143–146, ISBN: 978-1-4503-0228-9.

[105] R. Hunicke and V. Chapman, "AI for dynamic difficulty adjustment in games," in *AAAI Workshop on Challenges in Game Artificial Intelligence*, 2004.

[106] A. Zook and M. O. Riedl, "Temporal game challenge tailoring," *IEEE Transactions on Computational Intelligence and Artificial Intelligence in Games*, 2014.

[107] N. Shaker, G. Yannakakis, and J. Togelius, "Towards automatic personalized content generation for platform games," in *Proceedings of the Sixth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, ser. AIIDE'10, Stanford, California, USA: AAAI Press, 2010, pp. 63–68.

[108] H. Yu and T. Trawick, "Personalized procedural content generation to minimize frustration and boredom based on ranking algorithm," in *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, ser. AIIDE'11, Stanford, California, USA: AAAI Press, 2011, pp. 208–213.

[109] B. Harrison and D. L. Roberts, "Analytics-driven dynamic game adaption for player retention in Scrabble," in *IEEE Conference on Computational Intelligence in Games*, IEEE, 2013.

[110] M. O. Riedl, A. Stern, D. Dini, and J. Alderman, "Dynamic experience management in virtual worlds for entertainment, education, and training," *International Transactions on Systems Science and Applicaitons, Special Issue on Agent Based Systems for Human Learning*, vol. 4, no. 2, pp. 23–42, 2008.

[111] H. Yu and M. O. Riedl, "Personalized interactive narratives via sequential recommendation of plot points," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 2, pp. 174–187, 2013.

[112] M. Booth, "The ai systems of left 4 dead. keynote," in *Fifth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE'09). Stanford, CA*, 2009.

[113] C. Hoge, "Gdc vault — helping players hate (or love) their nemesis," Game Developers Conference, 2018.

[114] E. Hastings, R. K. Guha, and K. Stanley, "Automatic content generation in the galactic arms race video game," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, pp. 245–263, 2009.

[115] A. Zook, E. Fruchter, and M. O. Riedl, "Automatic playtesting for game parameter tuning via active learning," in *9th International Conference on the Foundations of Digital Games*, 2014.

[116] A. Liapis, H. P. Martınez, J. Togelius, and G. N. Yannakakis, "Adaptive game level creation through rank-based interactive evolution," in *IEEE Conference on Computational Intelligence in Games*, vol. 4, Springer, 2013, pp. 71–78.

[117] M. Treanor, B. Schweizer, I. Bogost, and M. Mateas, "The micro-rhetorics of Game-O-Matic," in *7th International Conference on the Foundations of Digital Games*, 2012.

[118] M. Cook and S. Colton, "A Rogue Dream: Automatically generating meaningful content for games," in *Experimental AI in Games Workshop*, 2014.

[119] M. J. Nelson and M. Mateas, "An interactive game-design assistant," in *13th International Conference on Intelligent User Interfaces*, 2008.

[120] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player modeling," 2013.

[121] A. M. Smith, C. Lewis, K. Hullet, G. Smith, and A. Sullivan, "An inclusive view of player modeling," in *Proceedings of the 6th International Conference on Foundations of Digital Games*, 2011, pp. 301–303.

[122] D. Hooshyar, M. Yousefi, and H. Lim, "A systematic review of data-driven approaches in player modeling of educational games," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1997–2017, 2019.

[123] E. Loria and A. Marconi, "Reading between the lines–towards an algorithm exploiting in-game behaviors to learn preferences in gameful systems," in *International Conference on the Foundations of Digital Games*, 2020, pp. 1–12.

[124] R. Orji, R. L. Mandryk, and J. Vassileva, "Improving the efficacy of games for change using personalization models," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 24, no. 5, pp. 1–22, 2017.

[125] G. F. Tondello, R. Orji, and L. E. Nacke, "Recommender systems for personalized gamification," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 425–430.

[126] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, "Intelligent tutoring goes to school in the big city," 1997.

[127] Y.-E. Liu, C. Ballweber, E. O'rourke, E. Butler, P. Thummaphan, and Z. Popović, "Large-scale educational campaigns," *ACM Trans. Comput.-Hum. Interact.*, vol. 22, no. 2, 8:1–8:24, Mar. 2015.

[128] M. Rogers, W. Yao, A. Luxton-Reilly, J. Leinonen, D. Lottridge, and P. Denny, "Exploring personalization of gamification in an introductory programming course," in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, New York, NY, USA: Association for Computing Machinery, 2021, pp. 1121–1127, ISBN: 9781450380621.

VITA

Kristin Siu was born in 1988 and grew up near San Francisco, California. She received a B.S. in computer science from Carnegie Mellon University's School of Computer Science in 2010 and a Ph.D. in computer science from the Georgia Institute of Technology's School of Interactive Computing in 2021.

Currently, Kristin lives in Seattle, Washington, where she is a Senior Software Engineer at Microsoft working on the educational edition of *Minecraft*.

Outside of research and work, Kristin is an independent game developer; her most recently published side project was the time-looping Shakespearean game *Elsinore*. Kristin is an avid programmer, tea drinker, and lover of small cute animals, most specifically hamsters.