Applying Fair Reward Divisions to Collaborative Work

by

Gregory Lawrence d'Eon

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2019

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes first-authored content from two conference submissions:

Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. Paying Crowd Workers for Collaborative Work. Submitted to the *2019 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '19)*.

Greg d'Eon, Kate Larson, and Edith Law. The Effects of Single-Player Coalitions on Reward Divisions in Cooperative Games. Submitted to the *2019 ACM Conference on Economics and Computation (EC '19)*.

Abstract

Collaborative crowdsourcing tasks allow workers to solve more difficult problems than they could alone, but motivating workers in these tasks is complex. In this thesis, we study how to use payments to motivate groups of crowd workers. We leverage concepts from equity theory and cooperative game theory to understand the connection between fair payments and motivation. Based on findings from a systematic literature review, we identify how the implications of equity theory relate to the Mechanical Turk ecosystem. Then, we use a realistic audio transcription task to evaluate how theoretically fair payments affect crowd workers. Further, we carry out two experiments to find how people's perceptions of fair rewards differ from cooperative game theory's fairness axioms. Our findings have important implications for designing collaborative work and directing future research on perceptions of fairness.

Acknowledgements

First, I'd like to thank my advisors, Edith Law and Kate Larson. When Kate says that she appreciates any interesting research, she really means it. Thank you for giving me the freedom to find my own direction while keeping me from going too far off the rails.

Second, Waterloo is full of brilliant faculty members, and I'm grateful that I had a chance to work so closely with them. Thanks to Dan Vogel, Ed Lank, and Keiko Katsuragawa, who have always been willing to give support and advice to anybody in the HCI group; to Robin Cohen for her detailed feedback as a reader on this thesis; and to Joslin Goh for helping me learn how statistics work in the real world.

Thank you to all of the students I've been lucky enough to work with at Waterloo. You've made such a huge impact on my life for the past two years. A special thanks to Edith's group: Alex, Mike, Sasha, Sangho, Jessy, Nalin, Louis, Colin, Will, and Graeme; to Kate's group: Vijay, Ben, Valerie, Adam, Josh, and Alan; and to Nikhita, Johann, Blaine, Lisa, and everyone else in the HCI lab for being such incredible friends.

Finally, thank you to my family, Normand, Janet, and Jason. Your unconditional support gives me the confidence to constantly keep pushing myself.

Dedication

To my family and my friends, for believing in me and telling me that I should too.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Teamwork is crucially important to high-quality work. In organizations, work is overwhelmingly structured around teams [48]. While coordinating groups of people can be costly, teams of people bring together diverse experience and complementary skills that no individual member could provide, allowing organizations to be more flexible and adaptable. A similar emphasis on group-based work is also present in science [96]: over time, research has shifted from an individual endeavour to a collective effort. Collaboration allows groups of people to make better decisions and solve more complex problems than they could alone.

One field that has grown to incorporate group work is microtask crowdsourcing. Crowdsourcing platforms like Amazon Mechanical Turk allow human workers to complete tasks that are difficult or impossible for computers. For example, a typical task might have workers transcribe text from an image of a shopping receipt. Classically, these tasks have been done by individual workers with no communication or cooperation, but collaborative tasks have been used to solve new problems by relying on contributions from multiple workers. These group-based tasks have had workers cooperate in a pipelined workflow [7, 47], deliberate about their answers [13, 83], and brainstorm as a team [60, 101]. These techniques allow crowdsourcing systems to solve more difficult problems by enabling interactions between teams of workers.

To ensure that crowd workers do good work, it is crucial to keep them motivated. For traditional, single-worker tasks, the question of worker motivation has been thoroughly studied in the crowdsourcing literature. Workers are primarily motivated by monetary rewards [41], and the impacts of different payment systems are well understood [67, 78, 30]. However, these findings may not transfer to collaborative tasks, where the problem of motivating workers through payments is more complicated. When working in a team,

workers can often see each others work, and this extra information may have a large impact on their motivation. For instance, the simplest payment strategy of paying all workers equally may not be suitable, as the most skilled workers could feel undervalued if they know others are earning the same wages.

In this thesis, we focus on the problem of motivating groups of workers for collaborative work. To understand this problem, we leverage concepts from both equity theory and cooperative game theory. First, the main idea of equity theory [1] is that people believe their outcomes, such as their wages or bonuses, should be proportional to their inputs, such as the amount of work they contributed; these judgements are made by comparing their inputs and outcomes to those of their teammates. These equity judgements are linked to motivation. If people feel that their skills, effort, or time are not being recognized by their rewards, they fix this equity balance by putting in less work. Based on this idea, we investigate the impacts of fair payment divisions in collaborative crowdsourcing tasks.

While equity theory prescribes one type of payment division, cooperative game theory [11] is another tool for understanding fair payments. In the context of transferable utility games, the Shapley value [85] is a method for dividing rewards between a group of cooperating agents. It satisfies four fairness axioms – symmetry, efficiency, null players, and additivity – and it is the only reward division that does so. Thus, in Chapter 4, we use both the proportional payments suggested by equity theory and the Shapley value as theoretically fair payments, and we evaluate how both of these payments are perceived by workers in a collaborative crowdsourcing task.

However, we do not take Shapley's axioms for granted. Empirical work on cooperative games has shown that people acting as impartial decision makers often violate the null player axiom [21], rewarding players for being present even if they provide no value to the group, resulting in rewards that are more "egalitarian" than the Shapley value. Further, this prior work only focused on a restrictive set of games. In Chapter 5, we investigate how people divide rewards in cooperative games, and we use our data to compare human reward divisions to the Shapley value and its axioms.

This work lies at the intersection of two fields: we study how collaborative crowdwork can be supported by understanding fair payment divisions, and we use this application domain to ground the payment divisions prescribed by cooperative game theory. Thus, this thesis makes three main contributions to the crowdsourcing and cooperative game theory literature:

- First, we present a comprehensive literature review of existing collaborative crowdsourcing tasks. We note similarities and differences between these tasks by identifying

the different types of information that workers have available to them. We use our categorization to find tasks that rely on collaborative crowd work and to show when equity theory's implications apply to these tasks.

- Second, we carry out two experiments using a crowdsourced audio transcription task to find how workers react to fair payments. In these experiments, we compare workers' perceptions of fairness when they are paid performance-based bonuses. Our results, which show that the theoretically fair proportional payments and Shapley values are recognized as being more fair than equal pay, inform the design of future collaborative crowdsourcing tasks.

- Third, we perform an empirical comparison between human reward divisions and the Shapley values in cooperative games. We use two experiments to study how people select rewards while acting as impartial decision makers. In contrast with previous work, we identify games where human rewards are unrelated to the Shapley values, and we show that humans violate both the null player and additivity axioms. These findings open a direction for future research on fairness axioms, helping artificial agents encode human standards of fairness.

The remainder of the thesis is organized as follows. Chapter 2 gives an overview of crowdsourcing and describes theoretical methods for paying groups of workers fairly. Chapter 3 reviews existing collaborative crowdsourcing tasks and their relationship to equity theoretic concerns. Chapter 4 describes a pair of experiments using a collaborative crowdsourcing task to evaluate the effects of fair payments on crowd work. Chapter 5 details a second set of experiments that investigate the differences between axiomatic definitions of fairness and human-selected rewards. Chapter 6 concludes the thesis by summarizing and proposing directions for future work.

Chapter 2

Background

In this chapter, we describe existing crowdsourcing platforms and the factors that affect worker motivation on these platforms. Then, we draw on concepts from social psychology to demonstrate the importance of fair payments in collaborative crowdsourcing tasks. Finally, we present a number of axiomatic definitions of fairness from cooperative game theory that can also be applied to this type of work. These ideas provide a theoretical foundation that we rely on in the remainder of this thesis.

2.1 Crowdsourcing

Crowdsourcing is the rapid mobilization of large numbers of people to accomplish global-scale tasks [46]. One of the most prevalent microtask crowdsourcing platforms is Amazon Mechanical Turk. On Mechanical Turk, requesters post work in the form of Human Intelligence Tasks (HITs). Workers accept these HITs, complete them, and submit their work to be reviewed by the requester. These HITs typically consist of short microtasks, which are tasks that take seconds or minutes to complete, and requesters usually pay a reward of a few cents for completing a task. Mechanical Turk acts as a marketplace, connecting workers and requesters.

Mechanical Turk gives requesters access to a large, global workforce. Difallah et al. [22] estimated the size of the worker population using capture-recapture techniques inspired by ecology. Their analysis of 85,000 survey responses over a 28 month period suggested that Mechanical Turk's worker population has 100,000 to 200,000 unique workers, with at least 2,000 workers active at any given point in time. They also found that, although the

majority of workers are from the United States (75%), there is a long tail of workers from other countries, including a significant number from India (16%). Finally, they showed that the worker population renews over time: while many workers leave the platform, tens of thousands of new workers arrive every year.

Using Mechanical Turk gives requesters the ability to use human computation for tasks that are difficult or impossible for computers to complete. Hara et al. [28] presented data collected through a Chrome extension from 2,676 workers completing 3.8 million HITs. They found that the majority of workers' tasks are "content creation", such as transcribing or tagging audio clips, images, or videos. Other common tasks include interpreting information (e.g., reviewing or rating images, articles or webpages) collecting data (finding company contact information), and completing academic surveys. The large scale of Mechanical Turk's workforce makes it possible to scale these tasks to an enormous extent. For one prolific example, the ImageNet dataset [80] consists of approximately 1.3 million images that were labeled by workers on Mechanical Turk. ImageNet had an enormous influence on research in machine learning and computer vision, leading to technical developments that would have been impossible without the huge scale of crowdsourced work.

We note that the term "crowdsourcing" has been used to describe many types of distributed work. In some situations, the crowd consists of a group of volunteers. Crowds of volunteers write articles on Wikipedia [45], answer programming questions on Stack Exchange and GitHub [92], contribute to citizen science efforts like Zooniverse [87], and spread news about crises on Twitter [89]. Crowdsourcing techniques have even been used to improve course materials for online classes based on students' input [94]. Other platforms, such as UpWork [15], use a workforce of expert freelance workers, such as graphic designers and programmers. Motivation is an important problem in all of these forms of crowdsourcing. However, in this thesis, we choose to focus on the problem of motivating workers through monetary payments, so we restrict our attention to paid, non-expert, microtask crowdsourcing platforms like Mechanical Turk.

2.2 Worker Motivation and Pay

The problem of picking payments for workers on Mechanical Turk has been studied extensively in the context of motivation. As Kaufmann et al. [41] found, monetary rewards are one of the most effective sources of motivation for crowd work. They surveyed 431 workers to find which factors of extrinsic and intrinsic motivation are most important to workers. Their results show that, while intrinsic factors such as enjoyment and skill variety influence the type of tasks that workers select, payment is the most important motivator.

Mason and Watts [67] described one of the earliest studies relating financial incentives and worker performance. In their experiments, they paid workers \$0.01, \$0.05, or \$0.10 for two different types of tasks. In the first task, workers arranged images from traffic cameras in chronological order; in the second, workers completed word search puzzles. They found that offering higher payments caused workers to complete more tasks, but had no reliable effects on workers' accuracy in either task.

Rogstadius et al. [78] extended this work by including an element of intrinsic motivation. In their experimental task, workers counted the number of infected blood cells in a series of images. They posted tasks with three levels of pay (\$0.00, \$0.03, or \$0.10) and with two different task backstories: one where the work was for a non-profit health organization, and the other for a private pharmaceutical company. They found that offering higher pay attracted more workers and caused workers to complete more tasks, but had no impact on their performance, confirming Mason and Watts' results. They also found that emphasizing the importance of the work by mentioning a non-profit organization led to higher accuracy.

Harris [29] performed a preliminary study of performance-based bonuses. They used two resume reviewing tasks, using ratings from an experienced HR hiring director as a gold standard. They tested several types of incentives: they either doubled workers' pay for matching the gold standard, halved the pay for significantly different answers, or both. Their data showed that workers spent more time on their tasks when they were offered performance-based payments. The results also suggested that the positive incentives – increased pay for correct answers – could lead to more accurate answers, but they did not find conclusive evidence of this finding.

Yin et al. [97] tested the effects of performance-based bonuses, where workers only received a bonus if their work met a pre-determined quality standard. They used two types of tasks. In their button-clicking task, workers received a bonus for alternately clicking between two buttons 400 times in 3 minutes. In their spotting-differences task, workers received a bonus for successfully identifying all 5 differences between two similar images. They found that workers put in more effort when their offered bonus was increased over time, due to an anchoring effect. However, the magnitude of the performance-based bonus alone did not affect work quality.

The most comprehensive work in this area is due to Ho et al [30], who explained the variance in these previous findings. Through a series of four experimental tasks, they found that performance-based payments improve work quality when three conditions hold. First, the performance-based bonuses must be relatively large compared to the task's base payment. Second, the bonus criteria must push workers to do more work without appearing to be unachievable. Third, the task must be effort-responsive: spending more time on the

task must lead to better work. This last condition is not true for some common tasks such as handwriting recognition, where workers can reach their best possible accuracy with little effort.

To our knowledge, no prior work has investigated the effects of performance-based payments for collaborative crowd work. In collaborative tasks, this problem becomes more complex: workers can often see each others' work, and this information may have a large impact on their motivation. In the next section, we discuss how this complexity can be tackled with the framework of equity theory.

2.3 Equity Theory

Worker motivation is more complex in collaborative tasks: the key difference is that it is easier for workers to compare themselves against each other. Thus, payments that are sensible for individual work might not translate easily to collaborative work. For instance, the most skilled workers in a team might feel undervalued if they are paid the same amount as their teammates. This idea is formalized by equity theory.

Equity theory [1] states that humans compare themselves to other people to decide whether they are being treated fairly. Humans believe that their outputs are equitable when

$$\frac{O_{self}}{I_{self}} = \frac{O_{other}}{I_{other}},$$

where I is one person's perceived input and O is their output. In other words, this relationship states that somebody that puts in twice as much work as their colleague should be rewarded twice as much. These outputs typically refer to some type of tangible reward, such as wages or bonuses. However, the inputs are not clearly defined. Depending on the situation, the inputs could be related to the amount of time spent working, the quantity of work done, or the quality of the work.

When workers do not believe that their outputs are equitable, they change their inputs to fix the discrepancy. In other words, overpaid workers will put in more effort, and underpaid workers will put in less effort. Workers might even quit their work in response to extremely unfair outcomes. In order to keep workers motivated in collaborative work, it is crucial to ensure that they do not feel underpaid compared with other members of the group.

Equity judgements are subjective, and it is important to recognize when workers' equity judgements are biased toward themselves. Ross and Sicoly [79] investigated these biases

through a series of five psychological experiments. They found that, in general, people remember more facts about their own work than their teammates' work, especially if they believe that their team was successful. Thompson and Loewenstein [90] presented similar findings from an experiment where participants role-played as negotiators bargaining over a strike. Their participants were able to recall more facts favouring themselves than their opponents. These egocentric biases could make it difficult to find payments that are considered to be equitable by all of the workers in a team.

In order to make equity judgements, workers also need enough information to make comparisons with their coworkers. While workers have knowledge of their own inputs and outputs, in some situations, they may not be aware of others'. In particular, complex collaborative systems sometimes hide information about each member's contributions to the team, and wages are not always transparent. We discuss when this information is available in the context of collaborative crowdsourcing tasks in Chapter 3.

Equity theory is also related to the concept of organizational justice [17]. Organizational justice decomposes the idea of fairness into four distinct components. Three of these components – procedural, interpersonal, and informational justice – refer to the procedures being applied, the level of respect that workers receive, and the amount of transparency that workers see when an organization is making decisions. However, the final component is distributive justice, which describes the fairness of workers' outcomes, and is directly related to equity theory. Studies with employees and students have also shown that distributive justice is related to workers' satisfaction about their outcomes [17]. This connection between organizational justice and equity theory provides additional measures that can be used to evaluate workers' perceptions of fairness.

2.4 Axiomatic Definitions of Fairness

Rather than describing fair payments through the psychological viewpoint of equity theory, an alternative approach is to encode fair outcomes through a number of mathematical axioms. This approach is best described in the language of cooperative game theory. In this section, we give an overview of cooperative games and fair values in these games; we defer precise, mathematical definitions of these values to Chapter 5.

Coalition	Reward
(nobody)	0
Alice	20
Bob	20
Charlie	0
Alice, Bob	60
Alice, Charlie	20
Bob, Charlie	20
Alice, Bob, Charlie	60

Table 2.1: An example of a transferable-utility game with 3 players named Alice, Bob, and Charlie. Alice and Bob are *symmetric*: they make the same marginal contribution to every coalition. Charlie is a *null player*: he contributes nothing to any of the coalitions.

2.4.1 Values for Cooperative Games

In cooperative game theory, a *transferable-utility game* consists of a fixed set of players and a *characteristic function*. The characteristic function describes the amount of reward that every subset of players (every *coalition*) could earn by working together. An example of a transferable-utility game is shown in Table 2.1. Then, one of the main questions of cooperative game theory is: if all of the players choose to work together (forming the *grand coalition*), how should they divide their team's reward among themselves? A *value* is a mapping from characteristic functions to tuples of rewards, assigning a reward to each of the players. For example, one of the simplest possible values is the equal division value, which divides the grand coaltion's reward evenly between all of the players.

One of the most well-known values is the *Shapley value* [85]. The Shapley value has a simple interpretation. A player's *marginal contribution* to a group is the amount of extra reward that the group earns by including the player; the Shapley value simply gives each player their average marginal contribution, with the average taken over all possible permutations of the players. This value is characterized by four fairness axioms:

- *Efficiency*: all of the reward earned by the grand coalition is divided across the players.
- *Symmetry*: if two players make the same marginal contributions to every coalition, then they receive the same reward.

- *Null players*: if a player makes no marginal contributions to any coalition, then they receive no reward.
- *Additivity*: when any two games are combined by summing the coalitions' values, the value of the combined game is the sum of the values for the individual games.

In fact, the Shapley value is the unique value that satisfies all four of these axioms. Thus, if these four axioms can be said to capture properties of fair reward divisions, then the Shapley value can be considered as an axiomatically fair value.

However, several authors have proposed alternative axioms that could be more representative of fair reward divisions. In particular, there are several modifications to the Shapley value that result from replacing the null player axiom while keeping the efficiency, symmetry, and additivity axioms. The first of these is the family of *egalitarian Shapley values* [36, 10], which are the convex combinations of the Shapley value and the equal division value. Egalitarian Shapley values satisfy a weaker version of the null player axiom: null players earn non-negative rewards. Another alternative is the *solidarity value* [74], which modifies the Shapley value by sharing players' marginal contributions with the other members of the group. The solidarity value gives players no reward if every coalition containing them earns no reward. Finally, these ideas are generalized by the family of *procedural values* [61, 77], which allow different levels of "egalitarianism" to be applied to different coalition sizes. In a sense, these alternative values can reward players for being part of a group, even if they make no tangible contribution to the group.

Transferable-utility games can also be thought of as bargaining processes between groups of rational players. This framing leads to the idea of *stable* reward divisions, where no players have any incentive to deviate from their bargaining outcome. Reward divisions motivated by stability include the *core* and its generalization to the *least-core*, the *nucleolus*, the *kernel*, and the *bargaining set* [11]. However, these reward divisions are much more complex than the axiomatic values described above. For instance, many games have no stable outcomes, so the core is empty in these games. In this thesis, we choose to leave these stability concerns and instead focus on axiomatic definitions of fairness.

2.4.2 Empirical Studies of Cooperative Games

A separate line of previous work has studied how humans act when participating in cooperative games. The earliest of these studies is due to Kalisch et al. [39], who described a number of experiments where participants bargained face-to-face in a series of 4- to 7-player games. They found that players were generally willing to split their rewards equally,

and players with the most power in a game rarely took full advantage of their position. However, their main focus was on the bargaining process, such as the speed of the negotiations, the effects of having participants sit around a table, and the resulting coalition structures.

In the following years, other experimental work had a similar focus to Kalisch et al's studies. Kahan and Rapoport [38] summarized many of these early results, and Maschler [66] published a comprehensive survey of these experimental papers. Most of this work is characterized by two features. First, it places an emphasis on the bargaining procedure: participants acted as players in the games, bargaining with others about forming coalitions and splitting rewards. With this emphasis, these experimental results are difficult to compare to the Shapley value axioms. Second, it focuses on *zero-normalized* games, where none of the 1-player coalitions can produce any value. These patterns have continued in more recent work in this area, which has focused on how the coalition structures are affected by limited communication [8] and innovative bargaining protocols [73], or how artificial agents can bargain effectively in these games by using supervised learning to predict whether people will accept an offer [102].

One experiment by Kahan and Rapoport [37] is particularly notable. Instead of using zero-normalized games, they controlled the rank-ordering of the single-person coalition values to either have the same order or the reverse order of the players' bargaining power. Their results show that the Shapley value is generally a good fit when all 3 players form a single coalition together. However, their analysis also includes situations where only two of the players formed a coalition, making it difficult to evaluate the accuracy of the Shapley values.

The experiment done by De Clippel and Rozen [21] is the most relevant to our work. In their experiment, a group of 3 "recipients" earned baskets of items by answering trivia questions. These items had no value alone (for example, a left shoe), but could be valuable when combined with another recipient's basket (for example, making a pair of shoes). Then, impartial "decision makers" saw how much value each group of recipients could earn and chose how to divide these rewards. They conclude that humans select convex combinations of the equal split and the Shapley value. To our knowledge, their work is the first where the participants dividing the rewards are impartial to the divisions. However, their games are zero-normalized, as the recipients could not earn any rewards alone, and it it is difficult to tell how their findings would generalize to games that are not zero-normalized.

We note some parallels between our research and other empirical work. In classical game theory, it is well known that humans deviate from concepts such as the Nash equilibrium; these deviations have been described and modelled in the field of behavioural game

theory [9, 95]. Similar findings also exist in the fair division literature [25]. However, these models are domain-specific and cannot be directly applied to cooperative games.

Chapter 3

Collaborative Crowdsourcing Tasks

In order to study how to pay groups of crowd workers, it is important to first understand the types of collaborative work that they do. In this chapter, we perform a literature review of existing collaborative crowdsourcing tasks. We characterize the different types of information that workers see during these tasks and identify work that is only possible with close collaboration. Then, we describe how this information allows workers to make equity judgements about their payments. Finally, we explore how these existing tasks pay groups of workers.

Before we begin, we first identify a precise definition of collaborative crowdsourcing. We follow Malone and Crowston [62], who define collaboration as "peers working together on an intellectual endeavor". Based on this, we take collaborative crowdsourcing to include *any crowdsourcing task where work from multiple workers is used to produce a single result.* Note that this is quite a broad definition: for example, it includes systems where answers from independent workers are aggregated without any interaction between the workers.

Note that collaborative tasks can also be competitive. To be precise, Malone and Crowston [62] state that cooperation indicates situations where actors share the same goals, while competition connotes one actor gaining from another's losses. Group work typically includes both of these elements: Davis [19] notes the extremes of pure cooperation or pure competition are rare. Most group-based crowdsourcing tasks also fall into quadrants 1 ("generate") and 2 ("choose") of McGrath's task circumplex [68]. While tasks in these quadrants are primarily cooperative, they also include elements of competition.

3.1 Literature Review

We first review the existing literature on collaborative crowdsourcing tasks. We use this literature review to identify the different types of information that are available to workers during collaborative tasks. These features allow us to characterize existing tasks into a number of distinct categories, each embodying a different level of interaction between the workers.

We performed our literature review using a snowball sampling process, which is a standard procedure for literature reviews [55]. Our search was seeded with Bernstein et al.'s Soylent [7]: as one of the first crowdsourced workflows, it represents one of the earliest and most recognized collaborative tasks. Then, we iteratively reviewed references in both directions by checking the reference lists and Google Scholar "cited by" lists. We kept all papers that described a collaborative crowdsourcing task. This process resulted in a total of 114 papers, the majority of which describe tasks for Mechanical Turk. We note that a small number of these tasks are intended for other platforms, such as professional crowdsourcing (e.g., Upwork) or citizen science (e.g., Zooniverse) platforms.

Through several rounds of informal iterative coding, we identified 4 factors that differentiate these collaborative tasks from each other. Each of these factors describes one type of information available to workers and the interactions that workers have during the task:

- *See others' work:* Do they see work completed by other workers on the *same* task, on an *other* task, or not at all?
- *Aware of others:* Do they know that other workers are involved in the task, or not?
- *Identify others' work:* Can they identify which other workers did each part of the work, or is the work anonymous?
- *Freely interact:* Can they have open, free-form conversations with other workers, or not?

Note that not every combination of these factors is possible. For instance, workers cannot *identify others' work* if they cannot *see others' work*. We used these four factors to classify each paper based on the interface elements and flow of information used in their tasks.

The distinct categories that we discovered are shown in Table 3.1. We identified four types of tasks that are relatively common, appearing in at least 10 publications. Characteristics and representative tasks for each of these categories are:

	See others' work	*Aware of others*	*Identify others' work*	*Freely interact*	# papers	Category
Common (10+ papers)	N	N	N	N	30	No information about others
	O	N	N	N	15	Workflows (no awareness)
	S	Y	N	N	18	Shared interfaces (anonymous)
	S	Y	Y	Y	17	Full collaboration
Uncommon (0-9 papers)	S	N	N	N	6	Iterative tasks
	N	Y	N	N	3	Aware of other workers
	O	Y	N	N	7	Workflows (with awareness)
	O	Y	Y	N	1	Subcontracting
	S	Y	Y	N	5	Structured deliberation; shared interfaces
	S	Y	N	Y	3	Anonymous chat
Not MTurk	N	Y	N	Y	1	Solo work with chat room
	O	Y	N	Y	3	Workflows with chat
	O	Y	Y	Y	4	Professional workflows

Table 3.1: The categories of collaborative crowdsourcing tasks that we found in our literature review. For the *See others' work* factor, workers can see others' work for the same task (S), another task (O), or not at all (N). For the other three factors, the collaboration is either present (Y) or not (N).

- *No information about others:* Tasks that require input from multiple workers, but do not have any form of interaction between the workers. This category does not involve collaboration, but it is included for completeness. It does include some real-time crowdsourcing tasks such as Adrenaline [6], where workers complete tasks simultaneously, but have no information about their coworkers.

- *Workflows with no awareness:* Each worker's job depends on data from previous workers, but the data's source is not mentioned. For example, in the final step of Soylent's find-fix-verify workflow [7], workers are asked to confirm writing quality without being told that the sentences were rewritten by other Turkers.

- *Anonymous shared interfaces:* Workers contribute to a common, shared interface, but cannot directly communicate or identify which workers performed each part of the work. This approach has been used to control arbitrary GUIs [53], plan complex itineraries [100], and write creative stories [43].

- *Full collaboration:* Workers closely interact as a team. Typically, this type of collaboration is achieved using a shared writing space, such as Google Documents or Etherpads, or using a chatroom such as a Slack workspace. This type of task is often associated with creative thinking [60], complex problem solving [101], or deliberation [83, 14].

We also identified six types of collaboration that are less common, but still present in previous work:

- *Iterative tasks:* A series of workers perform the same task, but are given previous results as a starting point or for inspiration. This approach works well for image segmentation [42, 40] and for some types of brainstorming [56, 86].

- *Aware of other workers:* The task interface mentions that other workers are completing the same task, but does not show their work. This technique is used to motivate workers in tasks that otherwise consist of individual work [31, 88].

- *Workflows with awareness of workers:* This category includes workflows where the presence of previous workers is explicitly mentioned [44, 27]. It also includes divide-and-conquer workflows [50, 49], where workers decide how complex tasks should be divided.

- *Subcontracting:* Morris et al. [71] proposed a workflow where workers choose to divide complex tasks through "subcontracting". They suggest that this system could include real-time chat to facilitate assistance between workers.

- *Structured deliberation and shared interfaces:* Some deliberation workflows only allow specific, structured communication between workers [13, 57]. Additionally, in some shared interfaces, it is possible for workers to see what each member of group is doing [52, 33].

- *Anonymous chat:* A small number of tasks involving chat interfaces show all messages coming from the anonymous "crowd" user [32].

Finally, we noted three other styles of collaboration that appear on other platforms, but have not appeared in microtask crowdsourcing. These three categories allow workers to communicate with each other, but vary the amount of cooperative work that they are involved in.

To emphasize the utility of collaboration, we point out four types of problems that are enabled using *structured deliberation*, *shared interfaces*, or *full collaboration* in this prior work. First, while individual workers are capable of some simple creative tasks, several creative writing tasks depend on workers having open discussions with each other [60, 82]. Second, workers are better at solving difficult cognitive tasks when they can communicate with each other to understand their team's strengths and weaknesses [16, 101]. Third, when tasks have subjective or unclear guidelines, deliberation can help workers converge on decisions [13, 83, 14]. Finally, collaborative environments help workers quickly divide tasks on the fly when it is difficult to automatically divide a job into microtasks [52, 63]. These tasks, which are only possible through close worker interaction, highlight the power of collaborative crowdsourcing.

3.2 Making Equity Judgements

Equity theory states that people compare themselves to their colleagues to decide whether they are being treated fairly. These comparisons are related to motivation: when people think they are undervalued, they restore the equity balance by putting in less work. However, making these comparisons requires knowledge of others' inputs and outputs. In collaborative crowdsourcing tasks, when can workers make these types of equity judgements?

First, workers must be able to see others' inputs. The availability of this information depends on the four factors that we identified about collaborative tasks. In tasks where workers have no knowledge of each other, they cannot compare inputs. However, when workers can *see others' work* and are *aware of others*, they can get a sense of the range of inputs that other workers are providing, giving them an approximate point of comparison. When workers can *identify others' work*, they can also make specific judgements about individual teammates. These extra pieces of information help workers to judge whether their payments are equitable in collaborative work.

Workers also need to have access to others' outputs (payments) to make equity comparisons. This information is much more readily accessible than others' inputs. Workers often discuss their pay public forums, such as Reddit's r/mturk or TurkerNation, or track

their wages on task reviewing websites, such as Turkopticon [35] or TurkerView. Many workers also rely on personal connections, and it is common for them to discuss wages [98]. These channels can give workers an idea of the payment range for a task.

Additionally, requesters can make this payment information transparent, allowing workers to see their teammates' exact rewards. Several authors have suggested that this added transparency would be beneficial, advocating for Amazon to make this information visible to workers. Martin et al. [64] concluded that additional market transparency would aid in minimizing Turkers' "work to make Turking work", helping them focus on their tasks. These impacts are magnified by the global nature of crowdwork [65]. Fieseler et al. [24] also advocated for increased transparency about workers' payments. They posited that this information would combat feelings that requesters are being deceptive about their workers' pay, making workers more loyal and improving trust and intrinsic motivation. Payment transparency could also help workers cope with unclear instructions by helping them recognize work that requesters marked as high- or low-quality. Overall, making payment information available would improve relations between workers and requesters, benefiting both parties.

3.3 Existing Payment Systems

Existing collaborative crowdsourcing tasks have used a variety of mechanisms to pay groups of workers, and these mechanisms are generally quite ad-hoc. Here, we discuss this wide range of payment methods.

The majority of collaborative tasks assign the same payment to every worker in a team. The most common method is to pay all workers a fixed, flat reward for completing a HIT, regardless of their performance. Some work has also given performance-based bonuses to teams, with each member receiving an equal bonus. Equal bonuses have been used to reward teams for converging to the correct answer [16], resolving disagreements [83], or performing above-average [82, 101]. These types of payment systems do not recognize differences between workers' contributions.

A number of tasks have paid workers based on their level of participation. These payments are used to incentivize workers to contribute to their group. In these tasks, the number of actions [54, 33], the amount of time spent in the interface [63], or the amount of chat interaction [59] have been used as participation metrics. It is difficult to ensure that these payments motivate workers to complete high-quality work. For instance, paying workers bonuses solely for suggesting chat messages [33] could cause workers to submit many low-effort suggestions.

Other tasks have paid different bonuses to team members based on the quality of their work in a variety of different ways. With ground truth answers available, some tasks have paid bonuses to individual workers for selecting the correct answer [23, 14]. Huang and Fu [31] found that paying bonuses to workers who outperformed their partner could also lead to increased effort. Without access to ground truth, these payments become more complex. In these cases, bonuses have been calculated based on workers' level of agreement with the crowd [53], the influence of their work on an algorithm's output [51, 40], or the subjective judgements of their teammates [81]. In general, these payments are ad-hoc: there is no existing systematic method of paying groups of workers based on the quality of their work.

Finally, some tasks have had workers take on distinct roles within their teams. Previous work has paid bonuses to workers for acting as a team lead [81] or a manager [101]. Reputation systems have also been used to give skilled workers access to higher-paying tasks [93]. Here, workers who take on more pivotal or influential roles are rewarded for their extra work.

Chapter 4

Fair Payments for Collaborative Crowd Work

In Chapter 3, we argued that equity considerations are important for motivating workers in collaborative crowdsourcing tasks. In this chapter, we aim to understand the practical importance of these equity theoretic issues in the Mechanical Turk ecosystem. We leverage ideas from equity theory and cooperative game theory to motivate two types of fair payments. Then, we use two experiments to evaluate how perceptive workers are to fair and unfair payments. We conclude the chapter by discussing the implications of our results for future collaborative crowdsourcing tasks.

4.1 Motivating Groups with Fair Payments

Workers on Mechanical Turk are primarily motivated by monetary rewards [78], and the impacts of various payments are well understood for individual work [30]. However, the problem of motivating workers with payments is more complex when workers collaborate: in these situations, there is an additional issue of choosing payments that workers perceive as being equitable. For example, paying all workers in a team equally might make the best-performing workers feel undervalued for their work. In this section, we detail how workers can make equity judgements on Mechanical Turk, propose two theoretically fair payment methods, and describe how to measure workers' perceptions of fairness.

4.1.1 Fair Payments

In this work, we focus on a specific set of payment systems. We suppose that a requester posts a group-based task where a team of workers earns a collective payment together. This payment could be fixed, as in many existing tasks, or it could include a performance-based bonus for the team. Then, the challenge of this system is to divide the group's payments among the individual workers.

The most basic payment method is to simply pay all workers equally. This method is the default in micro-task crowdsourcing: usually, workers received a fixed, pre-determined payment for submitting a task. However, equal payments do not recognize varying levels of skill and effort between workers in the group. Thus, we use equal payments as our control, and we propose two alternative group payment methods based on concepts from the literature.

The first alternative is to pay workers according to equity theory. In order to ensure that each equity judgment is satisfied, the ratio of each worker's output to input must be equal. This constraint leads to *proportional* payments, where the pay for worker i is

$$O_i = c \cdot I_i,$$

where c is the amount of pay per unit of work. We note that there is still some subjectivity in the measurement of one unit of work, as the input I could depend on several different metrics, such as work quality or quantity, or time spent on the task. In our experiments, we choose to measure workers' inputs by the quantity of correctly finished work.

A second type of fair payment is the *Shapley value* described in Chapter 2.4.1. A transferable utility cooperative game consists of a set of players N and a characteristic function $v(C)$ which describes the amount of reward that every possible group of players could earn by working together in a coalition $C \subseteq N$. Then, the Shapley value divides the team's total reward $v(N)$ between the players by giving player i a reward of

$$\phi_i = \sum_{C \subseteq N \setminus \{i\}} \frac{|C|!(|N| - |C| - 1)!}{|N|!} \left(v(C \cup \{i\}) - v(C)\right).$$

Intuitively, this is the average amount of value a player contributes by joining any possible coalition. This reward division satisfies four fairness axioms – efficiency, symmetry, null players, and additivity – and it is the only reward division that does so. Note that the value of these coalitions can be purely hypothetical. In our tasks, workers collaborate as if they are a member of the grand coalition N, and they have no control over this team

structure: they cannot choose to form a smaller coalition by removing their teammates from the coalition.

It is important to note that these theoretical methods cannot be applied to all types of work: both of them require a clear definition of workers' inputs. In some crowdsourcing tasks, there are no straightforward ways to compare workers. One example is in deliberation tasks, where describing an individual worker's contributions would require a deep understanding of the deliberation process. In this type of work, an alternative method for payment division is to ask workers how valuable their teammates are. Algorithms for combining workers' subjective reports have been studied in the social choice literature [25]. However, these methods must recognize workers' conscious or unconscious biases toward themselves [79, 90] and stop workers from colluding with each other to increase their payments. We discuss how future work can investigate these worker-determined payments in Chapter 6.2.

4.1.2 Measuring Perceptions of Fairness

In order to evaluate these theoretically fair payments, we need a method for measuring workers' perceptions of fairness. One way to compare group payments is to explicitly ask workers whether their payments are fair. Organizational justice is a construct that measures employees' perceptions of fairness in a workplace. Colquitt [17] summarized this literature by describing four different components of justice and a set of questions designed to measure each of these components. One of these components is distributive justice, which specifically focuses on the fairness of workers' outcomes. Colquitt showed that distributive justice is correlated with satisfaction: workers tend to be most satisfied with their outcomes when they feel that the distribution is equitable.

However, humans are not perfect at recognizing fairness: in fact, they are often significantly biased toward themselves [69]. There are multiple reasons for this effect. One reason is that people believe that their work is more valuable because they remember more facts about their own work than their colleagues. Another reason is that people may react more strongly to being underpaid than to being overpaid. Recognizing these biases is central to understanding the whole picture of workers' fairness perceptions.

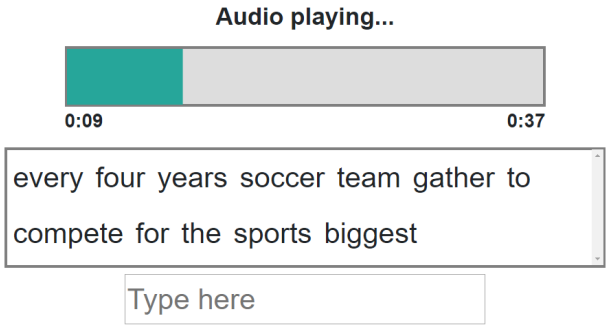

Figure 4.1: The audio transcription interface. Workers listened to short audio clips and typed the words they heard in real time. Each audio clip ended with 7 seconds of silence to allow workers to finish typing.

4.2 Study 1: Performance-Based Bonuses

In the previous section, we defined proportional payments and Shapley values, and we showed that these payments should be perceived as being more fair and should elicit more worker effort than equal payments. We performed a crowdsourced study to examine whether these effects can be observed in a real collaborative task. Specifically, this study attempts to answer three questions:

- **Question 1:** Do workers perceive proportional and Shapley value payments as being more fair than equal payments?
- **Question 2:** Are workers' fairness perceptions biased toward themselves?
- **Question 3:** Do workers put in more effort when they are paid fairly?

4.2.1 Method

To answer our three questions, we had workers complete a collaborative audio transcription task. We split performance-based bonuses between teams of workers using various bonus divisions, and we evaluated workers' fairness perceptions and performance levels based on these payment methods.

Participants

We hired participants from Mechanical Turk. We posted HITs with the title "Transcribe audio with a team of workers" and offered a base payment of $1.75. In the HIT instructions, we estimated that the HIT would take approximately 25 minutes, and we stated that workers would receive a performance-based bonus with a typical value of $1. We required workers to have at least 1000 approved HITs with a 95% or higher approval rate.

Teams

After workers accepted the HIT, we placed them into a 'virtual' team with two previous participants. We selected these teammates by drawing randomly from the pool of workers that had finished the experiment. To initialize this pool of workers for the first participants, we used data from workers that completed an earlier exploratory version of the experiment. We ensured that workers could only be selected as teammates twice. We also informed workers that their data may be re-used to serve as teammates to other workers in future batches of HITs. It was clear to the workers that their teammates had already completed the task, and were not currently using the interface.

Task

For our experimental task, we used a real-time audio transcription task based on Scribe [51]. Workers were not allowed to pause or replay the audio, as if the transcript was required in real time. We have several reasons for using this type of task. First, it is a difficult task, and workers need to focus to produce high-quality transcripts. Second, it is impossible for a single worker to produce a perfect transcript, motivating the need for multiple workers to complete the same task. Third, it is easy to learn, as many workers are familiar with regular audio transcription tasks. Finally, it is realistic: this interface could be used for a real-time captioning task. Our transcription interface is shown in Figure 4.1.

Procedure

In the experiment, workers first filled out a consent form and completed an interactive tutorial about the interface. Then, they performed 14 rounds of the task. In each round, they transcribed a short audio clip that we manually selected from podcast episodes.[1] We

[1] We used podcasts from http://freakonomics.com/.

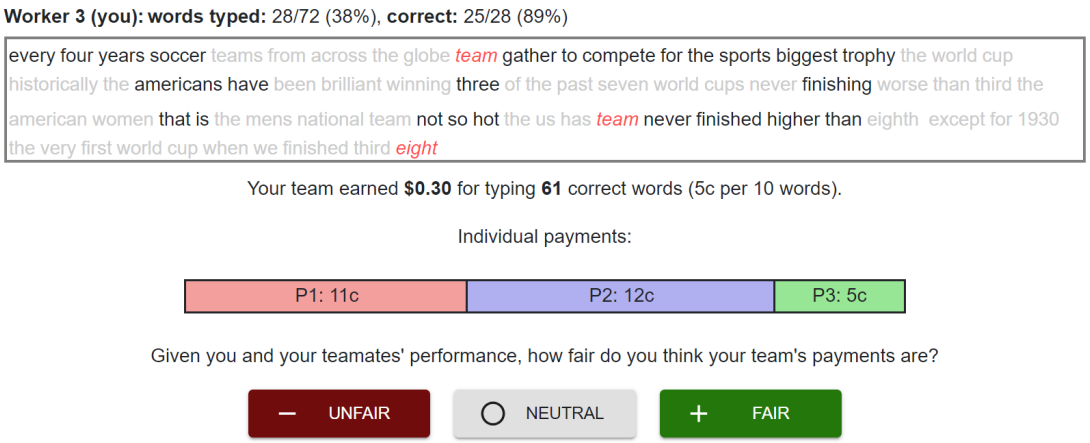

Figure 4.2: The bonus payment screen, showing an example of one worker's transcript. Workers saw the full text that their teammates typed, how these transcripts compare to the ground truth, and the exact bonus that each teammate received. (Workers could see all three team members' transcripts; to save space, we only show one here.)

used podcasts for our audio clips because there were high-quality transcripts available as a source of ground truth. The audio clips varied from 21 to 31 seconds with a median length of 28 seconds. We added an additional 7 seconds of silence to the end of each clip to allow workers to finish typing. We processed each word that workers typed by removing all punctuation and converting the text to lowercase. Then, at the end of each audio clip, we compared workers' transcripts to the ground truth with a word-level Myers diff [72], which allowed us to check whether workers typed each word correctly.

Bonuses

After each audio clip, we showed workers how well each member of their virtual team performed. We summarized each worker's performance by displaying both the number of words typed and the number of correct words. We also showed workers the full diff output, with correct words in black, incorrect words in red, and untyped words in gray, allowing them to interpret these results. Next, we counted the number of words in the ground truth transcript that were correctly typed by at least one worker. We calculated a total bonus payment of 5 cents for every 10 words that the team collectively typed correctly. We selected this bonus scale so a typical group would earn a bonus of 20 to 30 cents per round. The payment screen is shown in Figure 4.2.

After calculating the group's bonus, we divided it between the three workers. We placed teams into one of four experimental conditions:

- EQUAL: We gave each worker one third of the group's bonus. This method is the control, as it is similar to the default of paying a fixed HIT reward.
- PROPORTIONAL: We counted the number of words that each worker typed correctly. Then, we gave each worker a bonus proportional to the number of correct words that they typed. This method is fair according to equity theory.
- SHAPLEY: We computed the bonuses that each of the 8 possible subsets of the workers would have earned by collaborating on the task. Then, we paid workers with the Shapley values, using these bonuses as the characteristic function. This method is fair according to cooperative game theory.
- UNFAIR: We gave 50% of the bonus to the worker who typed the smallest number of words correctly, and we gave 25% of the bonus to the other two workers. We used this method as a manipulation check to understand how workers would react to payments that are clearly unfair.

In all four cases, we rounded bonuses down to the nearest cent. We displayed the transcripts and bonuses to workers in a payment screen at the end of each round, shown in Figure 4.2. Finally, we asked workers to rate the division of bonuses as 'Fair', 'Neutral', or 'Unfair' before proceeding to the next audio clip.

Post-Study

After transcribing all 14 audio clips, workers filled out a post-study survey. In the survey, we asked five 5-point Likert scale questions about the bonus payments. We adapted these questions from Colquitt's distributive justice and satisfaction measures [17]. Specifically, we asked workers whether their payments were appropriate, justified, acceptable, and satisfying, and whether the bonuses reflected the effort they put into the task. We also asked workers about their demographics, how they selected their fairness ratings, whether they enjoyed the task, and their feelings about working in a group with other workers. Lastly, after workers submitted the HIT, we granted bonuses to all three of the team members – both the participant and the two virtual teammates.

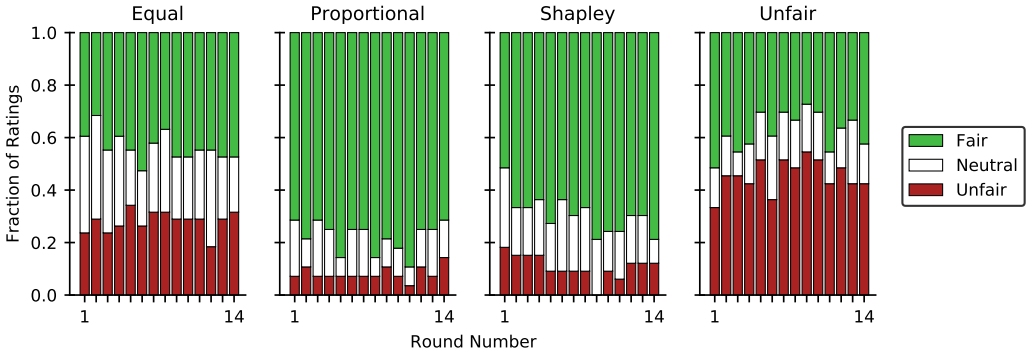

Figure 4.3: Workers' fairness ratings for each of the 14 rounds of the experiment. Workers in the PROPORTIONAL and SHAPLEY conditions rated their payments as being more fair than workers in the EQUAL or UNFAIR conditions.

Measures

Overall, we controlled two independent variables. The payment condition (EQUAL, PROPORTIONAL, SHAPLEY, or UNFAIR) was a between-subjects variable, and the round number was a within-subjects variable, with each worker completing all 14 rounds of the task. For dependent variables in each round, we measured the number of words that each worker typed, the number of these words that were correct, and their fairness rating (fair, neutral, or unfair). We also recorded workers' responses to the post-study questions, including their justice ratings and their responses to the open-ended survey questions.

4.2.2 Results

A total of 132 workers completed the HIT. We removed 2 workers who typed 0 words in the first round of the task. The number of workers in each condition varied from 28 to 38 workers; we confirmed that these conditions were not significantly unbalanced with a chi-squared test ($p = 0.65$). Workers typed an average of 29.23 words per round ($\sigma = 10.55$), with 24.48 of these words being marked as correct ($\sigma = 9.87$). In total, they earned an average bonus of 99.41 cents ($\sigma = 34.82$).

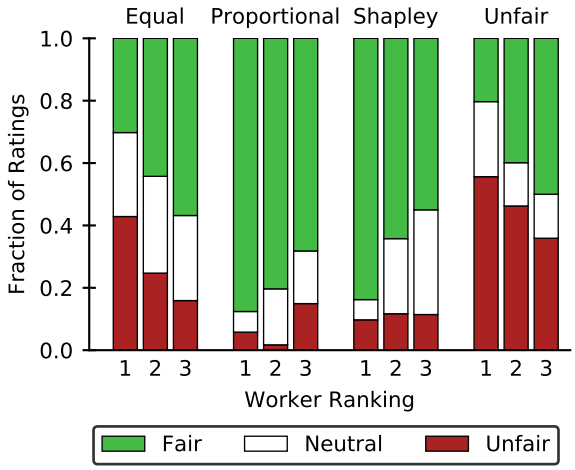

Figure 4.4: Fairness ratings for each condition, split by workers' ranking in the team. The best worker in each round has a ranking of 1, and the worst has a ranking of 3.

Fairness Ratings

Each worker submitted one fairness rating for each of the 14 rounds in the main experiment. These ratings are plotted in Figure 4.3. This plot shows that workers are most likely to rate their payments as fair in the PROPORTIONAL and SHAPLEY conditions. To confirm these differences, we fit a proportional odds model to these ratings using the workers' conditions as a factor. This model showed that ratings in the EQUAL condition were significantly more negative than the PROPORTIONAL ($p < 0.001$) and SHAPLEY conditions ($p = 0.002$), but not significantly different from the UNFAIR condition. Thus, the answer to our first research question is yes: workers do recognize theoretically fair payments as being more fair than equal payments.

Worker Bias

We also investigated the amount of bias in workers' fairness ratings. To do this, we split workers' ratings across all rounds into three groups: whether they were the best, the middle, or the worst worker in their team for each round. The distribution of ratings for each condition and team position is shown in Figure 4.4. This plot suggests that workers' perceptions of fairness change based on their abilities, relative to their teammates.

We confirmed these biases by adding a measure of the workers' relative skill levels to our proportional odds model. For each round, we calculated the skill difference between

the participant and their two teammates as

$$\begin{aligned}\textsc{Skill Difference} = 2 \cdot \textsc{Words Correct}_{worker} \\ - \textsc{Words Correct}_{teammate1} \\ - \textsc{Words Correct}_{teammate2}.\end{aligned}$$

This quantity is positive when the participant types more correct words and negative when they type fewer correct words than their teammates. After adding this factor to the model, the results showed that Skill Difference had a negative effect in the Equal ($p = 0.006$) and Unfair ($p < 0.001$) conditions: workers with more skill than their teammates thought that these payments were less fair. On the other hand, it had a positive effect in the Shapley condition ($p < 0.001$), where workers felt their pay was more fair when they had more skill than their teammates. Finally, Skill Difference had no significant effect in the Proportional condition.

Justice Ratings

Workers' answers to the five post-survey Likert scale questions had a high level of internal reliability (Cronbach's $\alpha = 0.92$). We aggregated these answers into a single justice score for each participant by taking the average of the five answers. The resulting justice scores are shown in Figure 4.5. This boxplot shows that the score distributions are not the same: workers in the Proportional and Shapley conditions never give very low scores. However, the median scores in the Equal, Proportional, and Shapley conditions are quite similar.

We used non-parametric statistics to analyze these ratings.[2] A Kruskal-Wallis test revealed that the condition had a significant effect on the justice scores: $H(3) = 18.42$, $p < 0.001$. We performed post-hoc Mann-Whitney tests with a Holm-Bonferroni correction and found significant differences between the Proportional and Unfair conditions ($p < 0.001$) and between the Shapley and Unfair conditions ($p = 0.01$). All other comparisons were not significant. This analysis shows that workers responded more favourably to the theoretically fair payments than to the unfair payments.

The differences between workers' justice scores in each condition were quite small. This effect contrasts with the fairness rating analysis, where the differences between the four conditions were more clear. This effect may be caused by the timing of these questions.

[2] We first fit a one-way ANOVA model to the ratings, but a Shapiro-Wilk test showed that the residuals were not normally distributed ($p < 0.05$).

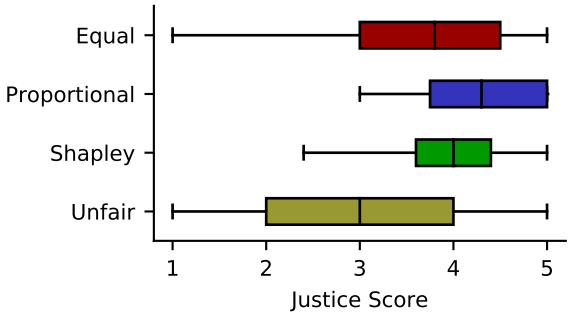

Figure 4.5: A boxplot of workers' justice scores in each of the conditions. Workers had higher justice scores in the PROPORTIONAL and SHAPLEY conditions than in the UNFAIR condition; no other comparisons were significant.

Table 4.1: Workers' average change in performance between the first and the last round. Workers in the PROPORTIONAL and SHAPLEY conditions improved more than in the EQUAL condition, but no comparisons were significant.

Condition	WORDS TYPED	WORDS CORRECT
EQUAL	M=4.03, $\sigma = 5.10$	M=4.97, $\sigma = 5.03$
PROPORTIONAL	M=5.43, $\sigma = 6.33$	M=5.93, $\sigma = 5.14$
SHAPLEY	M=6.82, $\sigma = 6.54$	M=7.30, $\sigma = 6.75$
UNFAIR	M=3.87, $\sigma = 6.16$	M=4.68, $\sigma = 6.95$

In the post-survey, workers may have considered their bonus payment for the entire experiment and answered whether it is fair, compared to typical Mechanical Turk wages. As our task paid more than the median wage on Mechanical Turk—approximately $2 per hour [28]— workers may have tended to answer more positively than expected. Alternatively, workers may have been hesitant to select the "extreme" answer of 1 for the justice questions.

Effort

We recorded two performance metrics in each round: the number of words each worker typed and the number counted as correct. These metrics are affected by many factors, including the length and difficulty of the audio clips, as well as the workers' skill and effort

levels. We chose to consider each worker's change in performance between the first and last rounds. Comparing these changes between conditions allows us to control for workers' skill levels, isolating their effort and learning rates as they become accustomed to the interface and style of the audio clips. The average changes are shown in Table 4.1. These values suggest that there may be a small difference in performance improvements between the conditions, with workers improving by 1 to 3 more words in the PROPORTIONAL and SHAPLEY conditions. With workers averaging approximately 24 words per audio clip, an improvement of this size might be practically relevant for large-scale tasks with hundreds or thousands of HITs.

To analyze these differences, we fit two binomial regression models: one to WORDS TYPED and another to WORDS CORRECT. In both of the models, we fit the workers' final round performance, using their condition and first round performance as factors. For both models, we found a main effect of first round performance ($p < 0.001$), but no main effects of condition or interaction effects. In other words, we could not detect any significant differences in performance changes between the conditions. To validate this result, we compare our results to previous work on bonus payments for crowdsourcing tasks. Ho et al. [30] found that workers corrected 1 additional error out of 12 when they were paid with appropriate bonuses. This improvement – an increase of less than 10% – was only detected with large samples of up to 1000 workers due to large variances in workers' skill levels. We suggest that studying workers' effort requires more accurate measurements of their baseline skill and tasks with less variation in their individual performance.

Survey Responses

Workers had a variety of explanations for their fairness ratings. Many workers mentioned making direct comparisons between the number of words or accuracy of their teammates. Others explicitly referred to the effort that they put into the task. Another common theme was the difficulty of the task: several workers were surprised that real-time audio transcription was so difficult. In particular, workers who thought they performed poorly often said that they were happy to get any bonus at all. We note that these feelings might affect workers' opinions about their payments: if they believe that they did poorly in the task, then they might be less critical of their bonuses.

Workers had diverse opinions about how enjoyable the task was. Negative comments tended to mention how frustrating, difficult, tedious, or weird the task was. Positive comments described the task as fun, challenging, or different from usual HITs. Workers were also split about the competitive aspect of the task: some workers enjoyed the competition, while others thought it was stressful to compare themselves against their team.

Many workers were positive about the idea of working in a team. They described it as being motivating and fun, while helping them to earn larger bonuses. They also mentioned that having multiple workers do the same task can make for useful feedback, allowing them to learn from each other. The negative comments argued that teamwork was more stressful, and some workers disliked the idea of relying on others. In this vein, some workers said that they would enjoy teamwork as long as their teammates were better than them.

4.3 Study 2: External Ratings

In our first study, we examined how workers respond to different payment methods for collaborative work. Now, in our second study, we used an independent group of workers to review the bonus payments from the first study. We used this second set of opinions to look for additional biases in the original workers' fairness ratings.

4.3.1 Method

Participants:

We hired participants from Mechanical Turk by posting HITs with the title "Review work done by other workers". We offered a HIT payment of $1.50 with no bonus. The HIT instructions gave a time estimate of 12 minutes. We required workers to have at least 1000 approved HITs with a 95% or higher approval rate. We also ensured that workers who completed the first experiment could not participate.

Task:

In the second study, workers did not complete any audio transcriptions. Instead, we showed them transcripts from previous teams of workers and asked them to rate how fair the bonus payments were. We used the same bonus payment screen except for minor modifications to the text (e.g., we changed "you and your teammates" to "the workers").

Procedure:

After workers accepted the HIT, they accepted a consent form and completed a tutorial. In the tutorial, we explained the real-time audio transcription task so that workers understood

the difficulty of the work. We also showed workers the bonus payment screen and asked comprehension questions about the transcript displays and bonus divisions. Then, workers were shown a total of 16 rounds from the audio transcription tasks. For each worker, we picked 3 random rounds from each of the 4 payment divisions. We also selected 1 fixed round for each payment division to show to every worker. These 16 rounds were randomly ordered. For each round, they clicked on one of three buttons, labelled "Fair", "Neutral", and "Unfair". As an attention check, we randomized the positions of the "Fair" and "Unfair" buttons in every round.

At the end of the study, workers filled out a post-study survey. We asked about their demographics, their reasoning for their fairness ratings, and whether they would like to rate or be rated by other workers in crowdsourcing tasks. Finally, workers submitted the HIT.

4.3.2 Results

A total of 79 workers completed the HIT. We removed 16 workers that averaged less than 5 seconds per round, leaving 63 workers. After this filtering step, we did not find any workers that clearly ignored the task instructions. For clarity, in this section we refer to the new participants as the external raters, and we refer to the participants from Study 1 as the original workers.

Fairness Ratings:

Workers submitted a total of 1008 ratings: 756 on the randomly selected rounds and 252 on the fixed rounds. We found that the original workers' ratings on the 4 fixed rounds that we selected were not representative of typical ratings in each condition, so we chose to focus only on the randomly selected rounds. The aggregates of these ratings are shown in Figure 4.6. This plot suggests that raters were generally more critical than the original workers, rating "Unfair" more often. This effect is strongest for the Equal and Unfair payments.

We first analyzed the external raters' ratings alone for each condition. To do this, we fit a proportional odds model to the ratings using only the payment method as a factor. This model shows significant differences between the Equal payments and each of the other three payment methods (all $p < 0.001$). Post-hoc tests with a Holm-Bonferroni correction showed significant differences between each of the conditions ($p = 0.002$ for Proportional – Shapley; all other comparisons $p < 0.001$). The directions of these post-hoc tests show

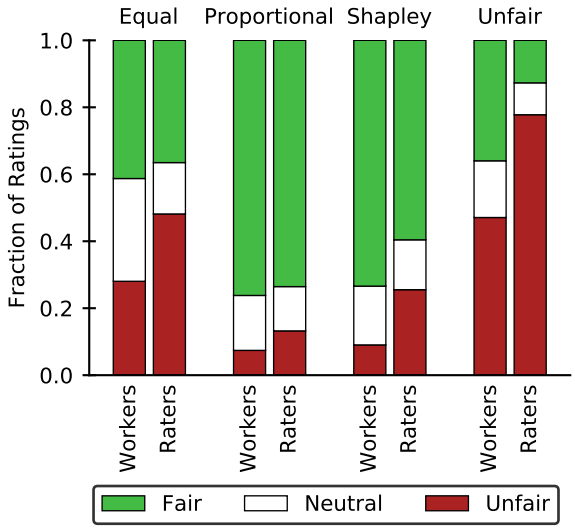

Figure 4.6: Fairness ratings from the original workers in study 1 (left), compared with the external raters in study 2 (right). Raters were more critical of the EQUAL, SHAPLEY, and UNFAIR payments.

that the PROPORTIONAL payments were rated as the most fair, followed by SHAPLEY payments, then EQUAL payments, with UNFAIR payments being rated as the least fair.

We also compared the original workers' fairness ratings with the external raters' to check for differences between these two groups of workers. For each condition, we performed a paired Wilcoxon signed-rank test between the two sets of ratings. These tests showed that the external raters found the payments less fair than the workers for the EQUAL ($p = 0.004$), SHAPLEY ($p < 0.001$), and UNFAIR ($p < 0.001$) conditions. We found no significant difference in the PROPORTIONAL condition ($p = 0.22$), suggesting that the external raters and the workers shared similar opinions about these payments.

We suggest several possible reasons for the differences in ratings between the two groups of workers. First, the original workers only saw one type of payment, while the external raters saw all four types. Workers may be more critical of EQUAL pay if they are aware of the other, theoretically fair payments. Second, external raters are not biased in the same ways that the original workers are. It is easier for raters to honestly judge whether a payment is fair because they do not benefit from payments that reward any of the team members.

Survey Responses:

The external raters judged fairness using similar criteria to workers in the first study. Most of the responses mentioned comparing the team members' numbers of words typed, correct words, or accuracy. A few raters were more interested in effort, and looked more carefully at the words that the team members typed in order to gauge how hard they were working. Some workers explicitly referred to their overall wages on Mechanical Turk, with one worker citing "hourly wage... how much I work to eat."

The majority of the workers were positive about the idea of rating each others' work, as long as they were paid to do it. Most workers were also happy to have their work judged by others. One worker pointed out that this is already close to their job: requesters can judge every HIT that they submit. However, several participants disagreed, saying that this felt invasive and that it would be hard to trust the raters. Finally, one response said that it would be stressful having to worry about performance ratings on top of already low pay.

4.4 Discussion

In this chapter, we studied how crowd workers are motivated by different payment divisions for group-based work. We identified two theoretically fair payments, motivated by equity theory and cooperative game theory, and evaluated how these payments can affect worker motivation in the crowdsourcing ecosystem. In our first study, we found that workers who were paid theoretically fair bonuses—that is, proportional to quality of their work or calculated with the Shapley values—reported their payments as being more fair than equal bonuses. Furthermore, our second study showed that this effect is even stronger for external raters that were not involved in the tasks. We also discovered that workers were positive about tasks that involve working with or evaluating other workers. Finally, our performance metrics suggest that workers might exert slightly more effort when they are paid with these fair bonus divisions, but we do not have conclusive evidence of this effect. In this section, we discuss the implications of our findings for future collaborative crowd work.

4.4.1 The Impacts of Payments and Transparency

In Chapter 3, we showed that collaborative tasks with close interactions between workers can be used to solve complex problems. By allowing these close interactions, teams of

workers can combine their skills in creative writing or cognitive tasks, work around subjective instructions, or divide work on the fly. However, in these tasks, it is essential to ensure that workers feel that they are paid equitably. Our experimental results showed that workers are receptive to fair and unfair payments, with the most skilled workers in the teams being the most sensitive to unfair payments. In order to keep workers motivated and satisfied with their rewards, it is crucial to pay workers relative to their contributions to the team.

Ethically, paying workers fairly for their work is the right thing to do. However, Fieseler et al. [24] posited that treating workers fairly and transparently is not simply a question of ethics. They proposed several features that crowdsourcing platforms – such as Mechanical Turk – could implement for the benefit of both workers and requesters. Allowing communication between workers would decrease feelings of isolation, help workers set time commitments and effort levels, and clarify task descriptions. Additionally, payment transparency would help to mitigate feelings that requesters are lying or deceiving workers. Together, these effects would lead to more committed workers with increased trust, satisfaction, and intrinsic motivation. We argue that collaborative tasks are an opportunity for requesters to reap these benefits now. Rather than relying on the platform to take action, requesters can implement tasks with explicit collaboration and public payment information. As long as requesters are conscious of paying fairly, these tasks are an excellent opportunity to build trust and reputation with workers, and ultimately to produce better results. However, help from the platform is still necessary to support workers by ensuring that requesters reveal accurate and truthful information about work and pay.

Workers need knowledge of their teammates' work and wages to make equity judgements, so requesters might think that they can sidestep the issue of fair payments by withholding this information. We reiterate that it is impossible to keep payments secret. Crowd workers have a basic social need for communication [26]; when they are not provided with communication channels, they seek to reproduce these channels in both public forums and private relationships. These external communication lines give workers a way to exchange payment information, and these discussions can often be more speculative than truthful. On top of requesters' moral duty to treat workers fairly, we also believe it is in requesters' best interests to communicate with workers on public platforms like Turkopticon and TurkerView, or even publicize payment information themselves.

There is also an opportunity here for crowdsourcing platforms to make an impact. Mechanical Turk hides most information about its workers, and requesters – particularly inexperienced ones – may interpret this anonymity as a signal that workers do not communicate with each other. This lack of information can encourage opportunistic or exploitative behaviour from requesters [24]; in fact, even well-intentioned requesters cannot

correct their actions unless they know that their workers are unhappy. Platforms can combat this behaviour by improving transparency on their marketplace: for instance, by displaying requesters' historical wages on the workers' interface. Although many workers already rely on external tools that provide this information, building these features into the platform would send a clear signal to requesters that they should be conscious about treating workers equitably.

A substantial number of workers from our studies were intrinsically motivated by working with others, describing the teamwork as enjoyable, motivating, and fun. For these workers, it would be useful to provide a consistent source of collaborative work. Gray et al. [26] proposed splitting crowd work into two separate streams, with one stream permitting collaboration in tasks that do not require independent responses. We suggest that this idea can be taken another step further. Rather than simply allowing workers to communicate, this stream of work can be designed to leverage the benefits that workers and requesters receive from transparent teamwork.

4.4.2 Fair Payment and Effort

In our main experiment, we did not find conclusive evidence that workers exert more effort when they are paid using theoretically fair methods. It is possible that there truly is no effect: Ho et al. [30] suggest that performance-based payments may not affect worker effort if the bonuses are too small, relative to the task's overall pay, or if the task is not effort-responsive. We used a relatively small bonus compared to our base payment so that even the lower-performing workers could earn close to minimum wage in our experiment. However, there are several other possible explanations for our results.

First, real-time audio transcription tasks are not perfectly suited for measuring a worker's skill and effort. Our metrics, which are related to typing speed and accuracy, have a large amount of variance between audio clips. Future studies on this topic should consider tasks where the quality of workers' output is more consistent, and should more carefully measure workers' initial skill levels – for example, using a longer qualification task.

The other reasons are factors that could affect workers' motivation and actions. We paid workers close to minimum wage, which is substantially higher than a typical task on Mechanical Turk [28]. We also told workers that they would be paid bonuses. Knowledge of a bonus might reduce workers' fear of having their work rejected, as bonuses on Mechanical Turk can only be paid after approving a HIT. Without this knowledge, workers might have worked harder to ensure their work is accepted, even if their bonuses are not motivating.

Finally, many workers mentioned that they found the task fun, interesting, and different. Workers that are intrinsically motivated might work hard regardless of their teammates' bonuses. Tedious, uninteresting tasks such as Yin et al.'s button-clicking task [97] would help to isolate the effects of bonuses on workers' effort. Longer tasks, taking an hour or more, would also help to capture these effects.

4.4.3 Limitations

We evaluated our payment divisions on an audio transcription task, where it is easy to view and compare results from several workers. These comparisons would be possible in basic content creation tasks, which are one of the most common types of HITs on Mechanical Turk [28]. However, there may not be a simple way to represent the amount of work that each worker has contributed to the team in more complex tasks. For instance, in a creative writing task, it is difficult to determine how valuable each worker was to the team's thought process. We cannot generalize our results to tasks without clear performance metrics.

In our main experiment, we simulated a team environment by comparing workers' transcripts against previous participants. This style of task is similar to existing crowdsourcing workflows, but it is quite different from tasks with real-time team interactions. Working with a team in real time may be more motivating, but it could make workers more frustrated or anxious as they are forced to work at the team's pace. More work is required to understand the impacts of real-time interaction.

Finally, we did not control for the location of the workers in our experiment. The majority of workers on Mechanical Turk are located in the United States, but an appreciable number live in other countries, with the largest group being from India [26]. It is possible that there are significant cultural differences between these worker populations that we have not studied here.

Chapter 5

Human Perceptions of Fairness

In Chapter 4, we used the Shapley value as a method for computing fair payment divisions between workers. The Shapley value is an attractive payment method for several reasons. It does not depend on subjective judgements of workers' inputs, and it is theoretically fair – at least, as long as its four axioms are believed to be fair. However, our experimental results showed that external, unbiased raters considered proportional payments to be more fair than the Shapley value. This finding hints at a fundamental difference between Shapley's axioms and human understandings of fairness.

In this chapter, we take a deeper look at human perceptions of fairness in the context of cooperative games. We use two controlled experiments to see how people divide rewards in fictional cooperative games, where they are impartial to the outcome. We compare our results to De Clippel et al. [21], who suggested that human reward divisions only violate the null player axiom when considering zero-normalized games. In contrast, our experiments show that people also violate the additivity axiom, and their reward divisions are often unrelated to the Shapley value.

5.1 Values for Cooperative Games

We begin this chapter by formally defining the cooperative game theory concepts that we described in Chapter 2.4.

5.1.1 Cooperative Games

A *transferable utility game* $G = (N, f)$ consists of a set of players $N = \{1, 2, \ldots, n\}$ and a characteristic function $f : 2^N \to \mathbb{R}$. This characteristic function assigns a reward $f(C)$ to each coalition $C \subseteq N$. We typically require $f(\emptyset) = 0$ – a coalition with no players earns no reward. In this paper, we restrict our attention to transferable utility games with $n = 3$ players, so we often refer to the characteristic function f as a "game". Also, we often write the set $\{i\}$ as i and the set $\{i, j\}$ as ij – for example, $C \cup i$ means $C \cup \{i\}$.

A player i's *marginal contribution* to a coalition $C \subseteq N \setminus i$ is the amount of reward that the player brings by joining the coalition

$$mc(i, f, C) = f(C \cup i) - f(C).$$

Two players i and j are *symmetric* if they contribute the same amount to all coalitions:

$$mc(i, f, C) = mc(j, f, C) \quad \forall C \subseteq N \setminus ij.$$

A *null player* is one who contributes nothing to any coalition:

$$mc(i, f, C) = 0 \quad \forall C \subseteq N \setminus i.$$

A game is *monotonic* if all possible marginal contributions are non-negative: for all $C \subset N$ and all $i \notin C$, $mc(i, f, C) \geq 0$. A game is *zero-normalized* if no player can earn a non-zero reward by working alone: for all $i \in N$, $f(i) = 0$.

For the sake of a running example, Table 5.1 shows a tabular representation of a 3-player game. In this example, Alice can earn a small amount of reward of 15 units by working alone, while Bob cannot. If Bob works together with Alice, they can earn a much larger reward of 60 units together. Charlie is a null player: adding him to any of the coalitions does not increase its value.

5.1.2 Values

A *value* is a function $v : \mathbb{R}^{2^N} \to \mathbb{R}^N$ that assigns a reward $v_i(f)$ to each of the players i in the game f. We will focus on *efficient* values, where $\sum_i v_i(f) = f(N)$ – all of the group's reward is allocated to the players. Perhaps the simplest value is the *equal division value* $ED(f)$, where each player receives an equal fraction of the total:

$$ED_i(f) = \frac{f(N)}{n}$$

Coalition C	**Reward** $f(C)$
(nobody)	0
Alice	30
Bob	0
Charlie	0
Alice, Bob	60
Alice, Charlie	30
Bob, Charlie	0
Alice, Bob, Charlie	60

Table 5.1: A 3-player transferable utility game.

The game in Table 5.1 has an equal division value of $[20, 20, 20]$.

The most celebrated value is the Shapley value [85], which is the unique value $Sh(f)$ that satisfies four axioms:

- *Symmetry*: if players i and j are symmetric in f, then $Sh_i(f) = Sh_j(f)$.
- *Efficiency*: the players' rewards sum to $f(N)$: $\sum_i Sh_i(f) = f(N)$.
- *Null players*: if player i is a null player in f, then $Sh_i(f) = 0$.
- *Additivity*: if f and g are two games, define a new game $(f + g)(C) = f(C) + g(C)$ for all coalitions C. Then, $Sh_i(f + g) = Sh_i(f) + Sh_i(g)$.

This value can be computed by rewarding each player the amount of value they bring to a coalition, averaged over all possible orders of building the coalitions:

$$Sh_i(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(n - |C| - 1)!}{n!} mc(i, f, C)$$

For example, for the game in Table 5.1, consider the order [Charlie, Alice, Bob]. Charlie, working alone, earns no reward. When Alice joins him, she adds 30 units of reward to the group's value. When Bob joins the pair, he brings the group up to a total reward of 60 units, adding another 30 units. Repeating these calculations for all 6 possible permutations of the players and averaging each player's contributions gives a Shapley value of $[45, 15, 0]$. Note that Charlie, as a null player, earns no reward.

More recently, alternative values have been proposed. One is the family of *egalitarian Shapley values* [36, 10], which is the set of convex combinations of the equal division and Shapley values:

$$Sh^{\alpha}(f) = \alpha Sh(f) + (1-\alpha)ED(f)$$

Here, the parameter α describes a social norm of equality: $\alpha = 0$ gives the equal division, while $\alpha = 1$ recovers the Shapley value. For example, with $\alpha = 0.5$, the egalitarian Shapley value for the game in Table 5.1 is $[32.5, 17.5, 10]$. Another is the *solidarity value* [74], which is

$$Sol_i(f) = \sum_{C \ni i} \frac{(n-|C|)!(|C|-1)!}{n!} A^f(C)$$

where $A^f(C) = \frac{1}{|C|}\sum_{i \in C} mc(i, f, C)$ is the average marginal contribution of any player to C. For the game in Table 5.1, the solidarity value is $[30, 17.5, 12.5]$. Nowak suggests that the solidarity value is more human, capturing some subjective psychological aspects of the game, while the Shapley value is the "pure economic" solution.

Note that each of these values describes a single reward vector for every game. While there are other methods for describing reward vectors, such as the core, the kernel, and the bargaining set, these are often more focused on the *stability* of the proposed reward vectors [12], and they can be multi-valued or empty for some games. We choose to leave these and instead focus on single-valued solution concepts.

5.1.3 Procedural Values

The key difference between these values is that they vary the amount of reward that each player receives for their marginal contributions to each coalition. This idea is generalized by the family of *procedural values*.

A procedural value $P^s(f)$ is described by a tuple of $n-1$ parameters $s = (s_1, s_2, \ldots, s_{n-1})$. Each term in this tuple is a measure of equality: when a player joins a coalition of size k, they keep a fraction s_k of their marginal contribution, and the remaining fraction $(1-s_k)$ is split equally among the coalition's other players. To simplify calculations, we denote $s_0 = s_n = 1$. These procedural values can be computed as

$$P^s(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(n-|C|-1)!}{n!} \left[s_{|C|+1} f(C \cup i) - s_{|C|} f(C) \right].$$

To help understand the effects of varying the s parameters, we describe a method for decomposing a value into several components. First, we define the games f^k for $1 \leq k \leq n$

as

$$f^k(C) = \begin{cases} f(C), & |C| = k \\ 0, & |C| \neq k \end{cases}$$

Then, we define $d^k(f) = Sh(f^k)$. Each of these d^k vectors represents the differences between the players' marginal contributions, only considering coalitions of size k. Note that $d^n(f) = ED(f)$, and for $k < n$, $\sum_i d_i^k = 0$ – adding d^k to a value preserves efficiency. For example, take the game in Table 5.1. Considering the 1-player coalitions, the first vector is $d^1(f) = [10, -5, -5]$: Alice can earn some reward alone, but Bob and Charlie cannot. For the 2-player coalitions, the vector is $d^2(f) = [15, 0, -15]$: Alice is the most productive in a pair, while Charlie is not useful to any of the pairs, and Bob is in between these two. This decomposition allows the procedural values for any game to be written as a vector sum; for a 3-player game,

$$P^s(f) = ED(f) + s_1 d^1(f) + s_2 d^2(f).$$

We also overload notation and write $d^{Sh}(f) = Sh(f) - ED(f)$ so the egalitarian Shapley values are

$$Sh^\alpha(f) = ED(f) + \alpha\ d^{Sh}(f)$$

We use procedural values to design our games and interpret our results in this paper for two reasons. First, the family of procedural values includes all of the values described previously: $ED(f)$ has $s_k = 0$, $Sh(f)$ has $s_k = 1$, $Sh^\alpha(f)$ has $s_k = \alpha$, and $Sol(f)$ has $s_k = \frac{1}{k+1}$. Second, De Clippel et al. [21] found the egalitarian Shapley values to be a good model for their results; procedural values are a natural way to generalize this idea to non-zero-normalized games.

5.2 Experiment 1

In our first experiment, we studied how people's reward divisions are affected when the players' marginal contributions stem from the 1-player or the 2-player coalitions.

5.2.1 Games

For our first experiment, we aimed to extend the results from De Clippel et al [21]. One natural way to do this is to construct a variety of games with identical Shapley values, but divide the players' contributions between the d^1 and d^2 vectors in different ways. Varying

the games in this way allows us to understand how much weight people put on the different coalition sizes. We expected people to pay more attention to the values of the individual players acting alone than to the values of the 2-player coalitions.

We used 11 games in our first experiment. To construct these games, we chose three Shapley values that represent different rank-orderings of the players. First, in 1-WORSE games ($Sh = [25, 25, 10]$), player 3 is less valuable than the other two. Second, in 1-BETTER games ($Sh = [30, 15, 15]$), player 1 is more valuable than the others. Third, in DISTINCT games ($Sh = [30, 20, 10]$), all three players have different values.

For each of these Shapley values, we created three games. In the SOLO games, all of the 2-player coalitions were all worth 60 units of reward, and the differences between the players were caused by their individual values. In the PAIR games, the single-player coalitions were all worth 0 reward, and the pairs had different values. Note that the PAIR games are zero-normalized. In the BOTH games, the players' marginal contributions were split between both coalition sizes. These games have

$$\begin{aligned} \text{SOLO:} \quad & d^1(f) = d^{Sh}(f); \quad d^2(f) = 0 \\ \text{BOTH:} \quad & d^1(f) = d^2(f) = \frac{d^{Sh}(f)}{2} \\ \text{PAIR:} \quad & d^1(f) = 0; \quad d^2(f) = d^{Sh}(f) \end{aligned}$$

We arbitrarily set the precise values of the coalitions to make the games monotonic. We also added two additional games: one purely additive game with a Shapley value of $[10, 20, 30]$, and one symmetric game where $Sh(f) = ED(f) = [20, 20, 20]$. These 11 games are listed in Table 5.2.

5.2.2 Method

Participants: We hired participants from Mechanical Turk. We posted human intelligence tasks (HITs) with the title "Divide rewards in fictional scenarios (10 mins)" and offered a payment of $1.25 USD. We required workers to have at least 1000 approved HITs with a 95% or higher approval rate. We restricted the HIT to workers located in the United States, and we used Mechanical Turk's qualification system to ensure that workers could not accept the HIT multiple times.

Task: During the experiment, participants were presented with a series of scenarios about three fictional characters – Alice, Bob, and Charlie – playing a video game online. Each of these scenarios was associated with a cooperative game, which describes how many

	Characteristic function								Shapley value		
Condition	∅	1	2	3	12	13	23	123	1	2	3
1-Worse-Solo	0	40	40	10	60	60	60	60	25	25	10
1-Worse-Both	0	15	15	0	45	30	30	60			
1-Worse-Pair	0	0	0	0	45	15	15	60			
1-Better-Solo	0	40	10	10	60	60	60	60	30	15	15
1-Better-Both	0	15	0	0	45	45	30	60			
1-Better-Pair	0	0	0	0	45	45	15	60			
Distinct-Solo	0	40	20	0	60	60	60	60	30	20	10
Distinct-Both	0	20	10	0	60	50	40	60			
Distinct-Pair	0	0	0	0	60	40	20	60			
Symmetric	0	20	20	20	40	40	40	60	20	20	20
Additive	0	10	20	30	30	40	50	60	10	20	30

Table 5.2: The 11 games used in experiment 1. All games have $f(\emptyset) = 0$ and $f(123) = 60$.

gold pieces every coalition could earn by working together. We displayed this information in a colour-coded table, which listed every combination of players and the amount of gold that the group could earn. Then, we told workers that the three characters all chose to work together, and we asked how the gold should be divided. Workers entered their responses by adjusting three sliders and clicking the submit button. The interface disabled the submit button as long as there was a surplus, only allowing efficient responses to be submitted. The experiment interface is shown in Figure 5.1.

Procedure: After workers accepted the HIT, they filled out a consent form and completed a brief tutorial. In this tutorial, we described the interface and asked comprehension questions about the reward displays. Then, workers completed 11 rounds of the task, with each round corresponding to one of the 11 games. We presented the 11 games to participants in a random order. We also randomly labelled players 1, 2, and 3 as Alice, Bob, and Charlie in each game.

At the end of the experiment, workers filled out a post-study questionnaire. Here, we asked participants about their age and gender. Then, we asked three open-ended questions: what factors they considered while dividing the rewards, whether they thought the solo or pair values were more important, and whether they thought other participants would have split the rewards differently. We defer our discussion of the post-study questionnaire to the

Figure 5.1: The task interface. Participants were presented with a tabular representation of the game and asked to divide the total reward between the three players. The "submit" button was only enabled when the entire reward was allocated.

final section of this paper. Finally, after completing this questionnaire, workers received a confirmation code and submitted the HIT.

5.2.3 Results

A total of 100 workers completed the HIT. We noticed a number of workers who submitted low-quality answers (for example, submitting [30, 30, 0] in the SYMMETRIC condition). To remove these low quality answers, we filtered out 21 workers who spent less than 5 seconds on at least one scenario. We also removed 4 additional workers who repeatedly submitted nonsensical answers, such as $[1, 1, 58]$ in DISTINCT-BOTH. The remaining 75 workers spent a median of 16.1 seconds per game. We confirmed that this filter criteria was appropriate by checking the rewards in the SYMMETRIC game. After filtering, the most extreme reward in this game was [20, 22, 18], differing from an equal division by 2 gold pieces.

For each game, each participant submitted one reward division. These rewards are plotted in Figure 5.2. Each of these plots shows the distribution of selected rewards, along with the equal division (red circle) and the Shapley value (blue circle) for every game.

There are several key features to note about these plots. First, most rewards in all games are close to affine combinations of the equal division and the Shapley value. On this line, there are a few key points where most rewards land. The most common is the equal division. The exact frequency of the equal division varies between games, but it is always at least 25 of the 75 participants. Then, the Shapley value also appears frequently in many games. Other common points include rewards half or double the distance from the equal division to the Shapley value.

Next, we focus on the differences between the 9 main games in this experiment. To more easily visualize these differences, for each reward, we calculated the the sum of the absolute differences – in other words, the L^1 norm – from the equal division. The distributions of these distances are shown in Figure 5.3. These distributions show substantial differences between the Solo, Both, and Pair games for each Shapley value. In all three cases, the rewards that are furthest from the equal division appear most often in the Solo games. Then, in the Both and Pair games, many of these people move toward a more equal division. For instance, in 1-Better-Solo, 14 participants submitted rewards that were approximately 40 gold pieces away from an equal division; in 1-Better-Both, only 3 such rewards remained.

We confirmed these trends using non-parametric within-subjects statistical tests. First, we performed Friedman tests to test whether the distances between the rewards and the equal division are different in the Solo, Both, and Pair games. We found significant differences between these games for all three Shapley values (all $p < 0.001$). Then, we performed pairwise Wilcoxon signed-rank tests with a Holm-Bonferroni correction for multiple comparisons. For all three Shapley values, we found a significant difference between the Solo and Pair conditions (all $p < 0.001$) and between the the Both and Pair conditions ($p < 0.01$). We also found a significant difference between the Solo and Both conditions for the 1-Better Shapley value ($p < 0.001$). These results confirm that people give more equal reward divisions in the Pair games and more unequal rewards in the Solo games.

5.3 Experiment 2

Our data from the first experiment appears to be consistent with De Clippel et al. [21]. However, the set of games in this first experiment was still quite restricted. In each of the

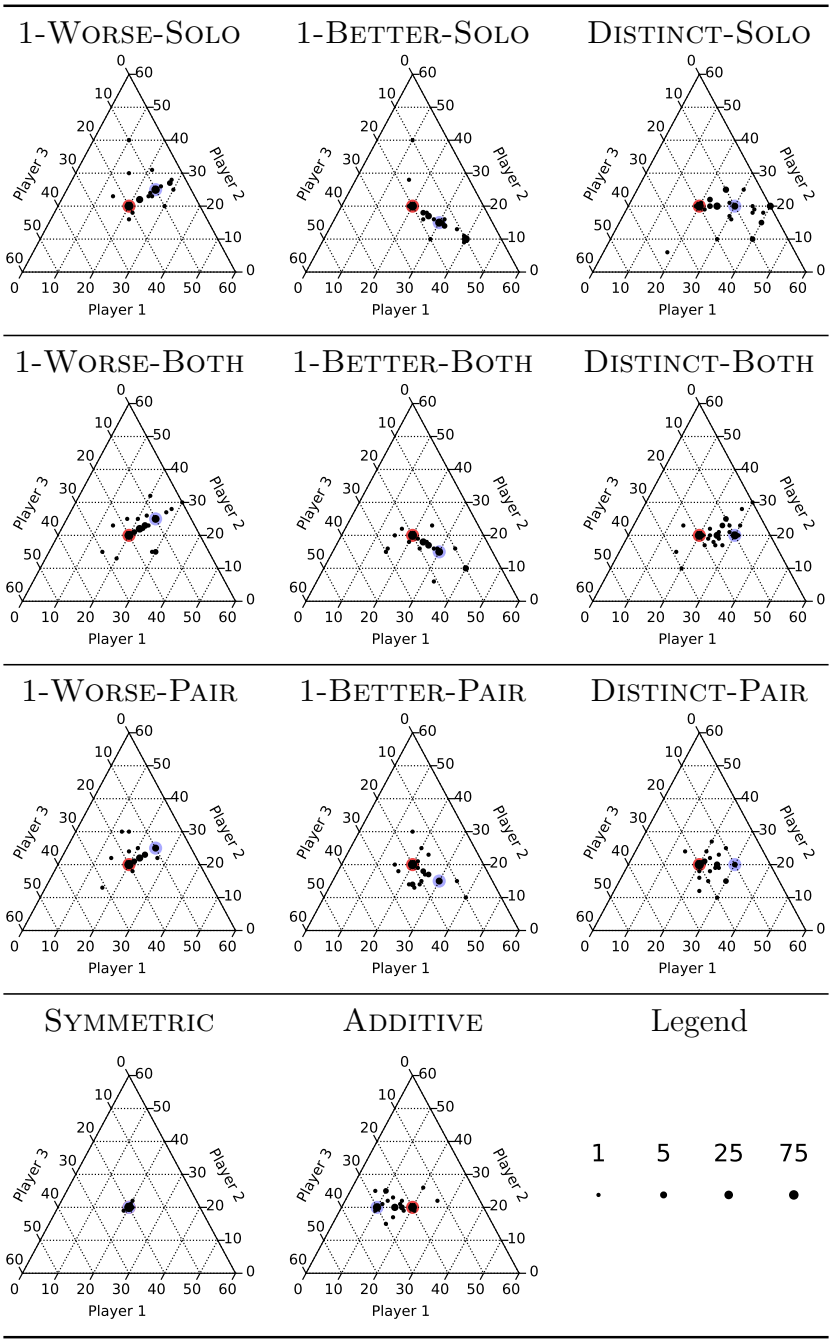

Figure 5.2: The rewards that participants submitted for each game in Experiment 1. On each plot, $ED(f)$ is circled in dark red, and $Sh(f)$ is circled in light blue.

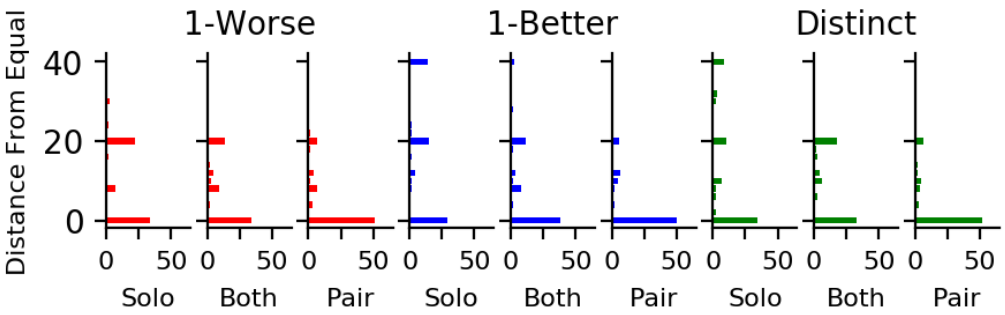

Figure 5.3: Histograms showing the distances between the submitted rewards and the equal division. Rewards far from an equal split are most common in the SOLO games, and equal divisions are most common in the PAIR games.

games, either d^1 or d^2 was 0, or d^1 and d^2 were colinear. We now turn our focus to games where these conditions do not necessarily hold. Specifically, we attempt to find games where human-chosen rewards differ from the egalitarian Shapley values, and we investigate how different properties of the 1-player coalition values affect these differences.

5.3.1 Games

We used 17 games in our second experiment. We picked two of the Shapley values, 1-WORSE and 1-BETTER, from the first experiment. Then, we selected d^1 vectors that do not point towards these Shapley values. For 1-WORSE, we used vectors of the form

$$d^1_{\text{1-Worse}} = [2x, -x, -x]$$

and for 1-BETTER, we used vectors of the form

$$d^1_{\text{1-Better}} = [x, x, -2x]$$

In other words, the 1-player coalition values in these games are "misleading". In these 1-WORSE games, the solo coalitions give the appearance that player 1 is more valuable than player 2; the reason why the Shapley value assigns the same reward to them is because player 2 is more productive when working in a pair with player 3. Similarly, in the 1-BETTER games, players 1 and 2 are both valuable while working alone, but player 1 is better at collaborating with player 3.

	Characteristic function								Shapley value		
Condition	∅	1	2	3	12	13	23	123	1	2	3
1-Worse-Zeros2	0	2	0	0	40	10	12	60	25	25	10
1-Worse-Zeros5	0	5	0	0	40	10	15	60			
1-Worse-Zeros10	0	10	0	0	40	10	20	60			
1-Worse-Sum30	0	20	5	5	60	30	45	60			
1-Worse-Sum45	0	25	10	10	60	30	45	60			
1-Worse-Sum60	0	30	15	15	60	30	45	60			
1-Better-Zeros2	0	2	2	0	38	40	10	60	30	15	15
1-Better-Zeros5	0	5	5	0	35	40	10	60			
1-Better-Zeros10	0	10	10	0	30	40	10	60			
1-Better-Sum30	0	15	15	0	45	60	30	60			
1-Better-Sum45	0	20	20	5	45	60	30	60			
1-Better-Sum60	0	25	25	10	45	60	30	60			
1-Null-Zeros	0	20	0	0	60	20	0	60	40	20	0
1-Null-Sum40	0	30	10	0	60	30	10	60			
1-Null-Sum50	0	35	15	0	60	35	15	60			
1-Null-Sum60	0	40	20	0	60	40	20	60			
Symmetric	0	20	20	20	40	40	40	60	20	20	20

Table 5.3: The 17 games used in experiment 2.

For each Shapley value, we selected 6 games using these vectors. In 3 of these games (Zeros2, Zeros5, and Zeros10), we gave values of 0 to some of the players and values of 2, 5, or 10 to the others. We thought that participants would give disproportionately low credit to players who could not earn any reward alone. In the other 3 games (Sum30, Sum45, and Sum60), we chose values for the players that summed to 30, 45, or 60. For example, in the 1-Worse-Sum30 game, the individual players' values are 20, 5, and 5, respectively. We were curious if people would place more weight on these values as they became larger.

We also included 4 games with a null player, where $Sh = [40, 20, 0]$. We refer to this value as 1-Null. In these four games, we gave player 2 a reward of 0 (1-Null-Zeros) or we had the individual rewards sum to 40, 50, or 60 (1-Null-Sum40, 1-Null-Sum50, and 1-Null-Sum60), respectively). We also included the Symmetric game from experiment 1. All 17 games are listed in Table 5.3.

5.3.2 Method

We used the same method as Experiment 1 with some minor adjustments. Due to the increased number of games, we posted HITs with the title "Divide rewards in fictional scenarios (15 mins)" and offered a payment of $1.75 USD. Otherwise, the rest of the procedure was unchanged. We continued to use the same qualification system, ensuring that no participants from the first experiment could access the HIT.

5.3.3 Results

A total of 100 workers completed the experiment. We used the same filtering criteria as in the first experiment, removing all workers that completed any round in 5 seconds or less. We also manually removed 4 workers with low-quality submissions. After filtering, we were left with 74 workers. These remaining workers spent a median of 15.1 seconds per game. We checked the filter's quality with the SYMMETRIC game; all 74 workers submitted a reward of $[20, 20, 20]$ for this game.

Each participant submitted a reward division for all 17 games. We split this data across two figures. Figure 5.4 shows the rewards for the 1-WORSE and 1-BETTER games; Figure 5.5 has the 1-NULL games.

Compared with the data from our first experiment, these rewards show some striking differences. In almost all of the games, the majority of the rewards that participants selected are not affine combinations of equal divisions and the Shapley values. Further, in the 1-WORSE and 1-BETTER games, the Shapley values are quite uncommon. In fact, in four of the 1-BETTER games, no participants chose the Shapley values. However, there is still a clear linear pattern to the rewards in most games. In the 1-WORSE games, most of the rewards lie between the equal division and the value $[60, 0, 0]$; in the 1-BETTER games, they lie between the equal division and $[30, 30, 0]$. These two trends are the directions of the d^1 vectors for both sets of games. The 1-NULL-ZEROS game appears to be similar to the 1-WORSE games, with many of the responses giving a disproportionately high amount of reward to player 1. Lastly, the other 1-NULL games have more rewards close to the Shapley values.

In order to describe these trends, we used principal component analysis (PCA). For each of the games, we computed the main principal component of the game's data. With our data, this principal component can serve as a d vector – its elements are guaranteed to sum to zero. We note that PCA is sensitive to outliers, but we elect to use it for this analysis because most of the outliers have already been removed from the data. The computed

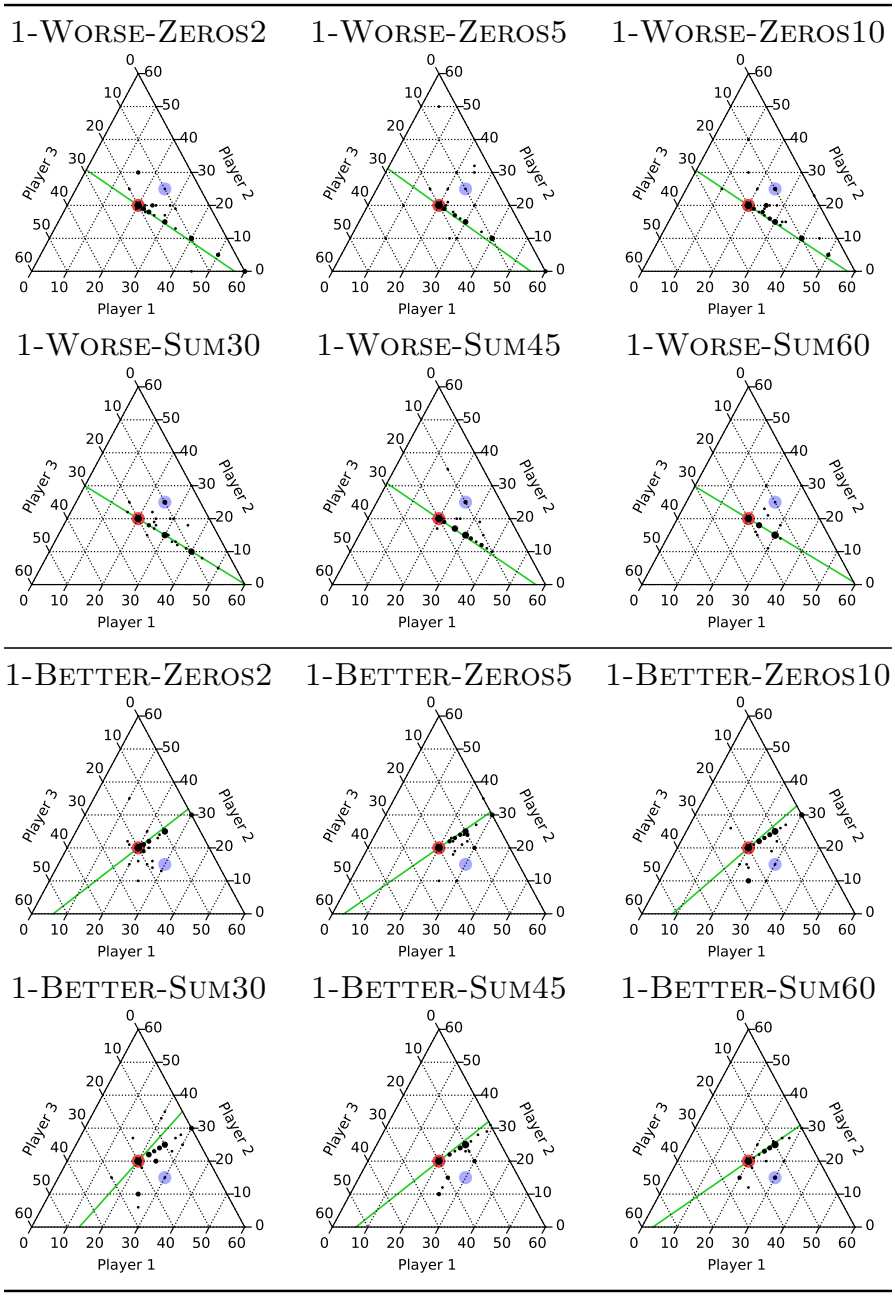

Figure 5.4: The rewards that participants submitted for each of the 1-Worse and 1-Better games in Experiment 2. In each plot, $ED(f)$ is circled in dark red, and $Sh(f)$ is circled in light blue. Green lines indicate the direction of the main PCA component. In all 12 games, the PCA component is close to $d^1(f)$, but far from $d^{Sh}(f)$.

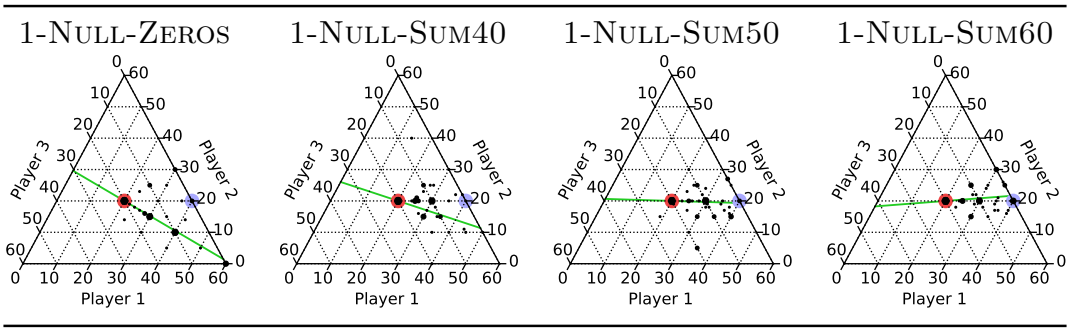

Figure 5.5: The rewards that participants submitted for the 1-Null games in experiment 2. In each plot, $ED(f)$ is circled in dark red, and $Sh(f)$ is circled in light blue. Green lines indicate the direction of the main PCA component. The rewards approach the egalitarian Shapley values as the 1-player coalition values become closer to 60.

principal components are plotted as green lines in Figures 5.4 and 5.5. Due to the high number of participants selecting equal splits in all games, we plotted these components as passing through the equal division. These components are highly consistent, with nearly identical directions in each 1-Worse game and in each 1-Better game. They also show the differences between the 1-Null games, where the components steadily shift from the extreme value in 1-Null-Zeros towards the set of egalitarian Shapley values in 1-Null-Sum50 and 1-Null-Sum60.

To make a formal comparison between the data and the Shapley values, we used bootstrapping to find confidence intervals for the angles of the PCA components. Specifically, to compute one bootstrapped estimate of the angle for a game, we sampled 74 points with replacement from the dataset of 74 rewards, and we recalculated the PCA component on our resampled data. We repeated this process 10000 times for each game to get a distribution of the PCA angles, and we took the middle 99% of these angles as the confidence interval. These intervals are shown in Figure 5.6. Only three of these confidence intervals contain the $d^{Sh}(f)$ vector: 1-Null-Sum40, 1-Null-Sum50, and 1-Null-Sum60. However, almost all of them contain the $d^{1}(f)$ vector; the only exception is 1-Better-Sum30, where it is 0.7° outside of the interval. The consistency of these directions strongly suggests that the egalitarian Shapley values are not a good model for human-selected rewards in these games.

Many of the confidence intervals in Figure 5.6 are quite tight, suggesting that most participants have similar reactions to the games. Some examples of these tight intervals are in the 1-Worse-Sum30 and 1-Better-Sum60 games, where most of the data falls neatly on the PCA line. However, there are several games with wider intervals, indicating

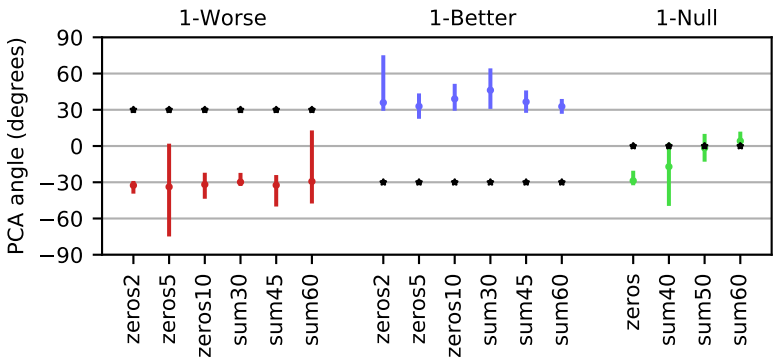

Figure 5.6: Bootstrapped 99% confidence intervals for the angles of each of the PCA components. For each game, the angle of the vector $d^{Sh}(f)$ is indicated with a black point. All but one (1-Better-Sum30) of the confidence intervals contain $d^{1}(f)$; only the rightmost three intervals contain $d^{Sh}(f)$.

that participants submitted a wider variety of rewards. Two of these games are particularly interesting.

First, in 1-Worse-Sum60, the confidence interval is wide because of a lack of extreme rewards: no participant gave more than 35 gold to player 1. This data contrasts with 1-Worse-Sum30, where it was much more common to give rewards close to $[40, 10, 10]$. This difference suggests that participants use the players' individual values as a baseline for their rewards; if these values are large, there is little room to vary the leftover reward between the players. Second, 1-Better-Sum30 has a large confidence interval due to a small number of rewards close to $[30, 0, 30]$. While these rewards might be dismissed as outliers, there is a possible explanation: in this game, $f(13) = 60$, so players 1 and 3 can obtain the entire reward without the help of player 2. It is plausible that some participants put a larger amount of weight on coalitions that can obtain the full team's value.

5.4 Discussion

The results from both of our experiments suggest that the single-player coalitions in cooperative games play an important role in people's decisions on how to divide the rewards. In the rest of this section, we provide additional insights and analyses as to how participants made reward division decisions. First, we analyse the post-study questionnaire data so as to better understand the stated rationales of participants decisions. We then compare

our experimental data in order to see how participant decisions aligned with the axioms used to define the Shapley value and with the rewards predicted by procedural values. We conclude with a discussion of the validity of our results and directions for future work.

5.4.1 Post-Study Questionnaires

The responses to the post-questionnaires in both experiments were similar, so we discuss both together in this section.

We asked participants what factors they considered when splitting the rewards and whether the values of the solo or the pair coalitions were more important. Approximately 40% of the responses explicitly mentioned basing their rewards on the individual players. Many of these responses explained that these individual values represented the skill or effort levels of the players. A few participants also mentioned that the solo coalitions are easier to understand, and that it is harder to know which players are the biggest contributors to the larger coalitions. Then, roughly 20% of the participants said that they split the rewards equally. Generally, they justified this choice by saying that the players chose to complete the quest as a group, so the amounts that they could have made alone are irrelevant – the only fair way to divide the rewards was an even split. The remaining responses gave a variety of answers with vague mentions of "fairness" or "equity".

We also asked participants if they thought others would select different rewards; approximately 65% said yes. Many participants said that the problem of allocating rewards has no objective answer, and they expected that others might simply consider different components of the scenarios. A number of participants who submitted equal rewards correctly suggested that others might focus on the solo coalition values, and vice versa. One response mentioned that people's selected rewards might differ because they have different calculation abilities. The remaining responses either suggest that others would split the rewards in a similar to manner to themselves or gave a non-committal answer.

5.4.2 Shapley Value Axioms

The Shapley values are characterized by four fairness axioms. In both of our experiments, we only allowed participants to submit efficient rewards, but we made no restrictions related to the other three. Did participants obey these axioms?

Symmetry: Six of the games in Experiment 1 have two symmetric players. In the three 1-Worse games, players 1 and 2 are symmetric; in the 1-Better games, players

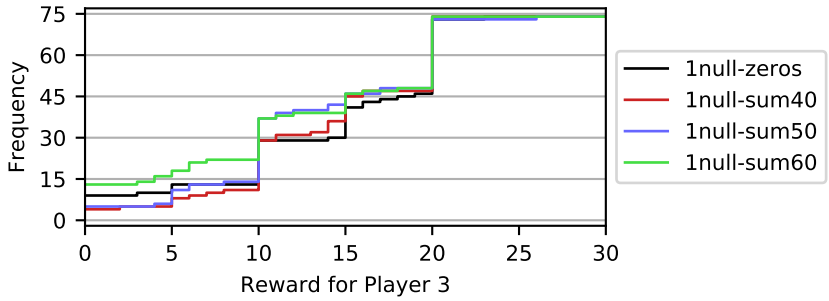

Figure 5.7: Cumulative distributions of the rewards that were assigned to null players in Experiment 2. The height at each reward r indicates the number of participants who gave no more than r gold pieces to the null player.

2 and 3 are symmetric. To check whether the rewards obeyed symmetry, we used paired Wilcoxon signed-rank tests to test whether the two symmetric players received different rewards. We found no significant differences between these rewards in any of these six games (all $p > 0.1$). Thus, we cannot reject the hypothesis that participants obey the symmetry axiom.

Null Players: In all four of the 1-Null games in Experiment 2, player 3 is a null player. It is clear from Figure 5.5 that a majority of players give a positive reward to player 3, breaking the null player axiom. To help quantify this behaviour, Figure 5.7 shows the cumulative distribution of the rewards that people assigned to player 3 in each of these games. This plot shows that, even in the best case (1-Null-Sum60), only 14 of the 74 participants satisfied the null player axiom; in the other games, this proportion is even smaller. Instead, many of the participants tended to assign small, round rewards to the null player, with rewards of 5, 10, and 15 being most common. (Note that the large jump at 20 is mainly caused by the large number of equal divisions.) While participants tend to recognize that null players contribute little to the group, they rarely go so far as to assign no reward to these null players.

Additivity: Several of the games in Experiment 2 are closely related. For instance, between the 1-Worse-Sum30 and 1-Worse-Sum45 games, the only difference is that all of the solo coalitions' values have been increased by 5. This change is equivalent to adding the game

$$f(C) = \begin{cases} 5, & |C| = 1 \\ 0, & |C| \neq 1 \end{cases}$$

It is difficult to argue that any value other than $[0, 0, 0]$ is reasonable for f: $f(N) = 0$, and all three players are symmetric. Then, to satisfy additivity, participants must select the same rewards for all three of the 1-Worse-SumX games and for all three of the 1-Better-SumX games.

We used 6 within-subjects Friedman tests to check if participants violated additivity. For instance, one of these tests checked whether participants assigned the same rewards to player 1 in the 1-Worse-Sum30, 1-Worse-Sum45, and 1-Worse-Sum60 games. 5 other similar tests compared the rewards for player 2 or 3 and for both the 1-Worse-SumX and 1-Better-SumX games. In the 1-Worse games, we found that the rewards allocated to players 1 and 3 varied significantly between the games (both $p < 0.01$). Our results also approached significance in the 1-Better games, where the rewards assigned to player 1 ($p = 0.08$) and player 3 ($p = 0.07$) were inconsistent. These results imply that people may not obey additivity in all cases – simply adding a constant to each of the solo coalitions has a significant impact on their reward divisions that cannot be explained by additive models.

5.4.3 Models for Human Rewards

Motivated by the concept of procedural values, we found that the values of the solo coalitions have a larger impact on human-selected rewards than the pair coalitions' values. Now, we ask: can predictions from procedural values be used to accurately describe our participants' rewards?

To answer this question, we tried finding procedural values that accurately described each participant's rewards. However, it would be too strict to require an exact match between the procedural value predictions and the actual rewards. To allow for some error, we searched for parameters (s_1, s_2) such that the predicted values differed from each participants' rewards by no more than a threshold t in any player's reward. We note one caveat: this goodness of fit measurement has a slight bias, as it is easier to satisfy this condition for rewards that are closer to an equal division. For every participant, we tested all combinations of parameters $s_1 \in [-1, 2]$ and $s_2 \in [-2, 2]$ in steps of 0.01; we report our results using thresholds of $t = 2$ and $t = 5$.

We had little success fitting these procedural values to individual participants. In experiment 1, we found 25 participants that submitted an equal division for every game, so they could be described by $s = (0, 0)$. However, few of the other participants could be described by any procedural value. With a threshold of $t = 2$, we only found a good fit for one participant with $s = (0.3, 0.2)$; increasing this threshold to $t = 5$, 19 participants

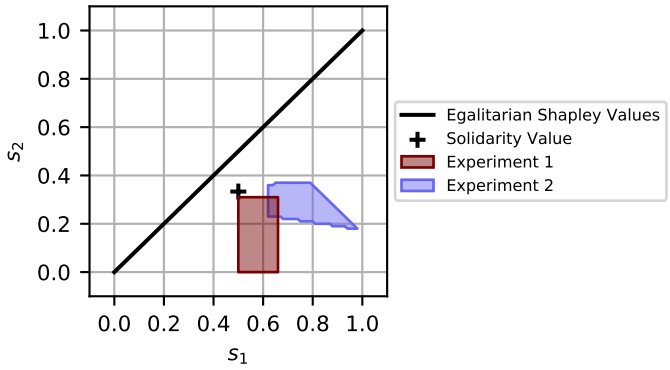

Figure 5.8: Regions of (s_1, s_2) that describe the population averages for each experiment. Experiment 1 averages were fit using a threshold of $t = 2$; Experiment 2 averages used $t = 4$. The regions overlap, but neither one contains the egalitarian Shapley values or the solidarity value.

had a suitable set of parameters. The remaining 31 participants could not be described by any procedural values. We found similar results in experiment 2. Here, 19 participants submitted equal divisions; otherwise, procedural values only described 2 and 8 participants at $t = 2$ and $t = 5$ respectively. We conclude that procedural values are generally not suitable for describing individual people's reward divisions.

We repeated this process using the population averages for each game instead of individual participants' rewards. Surprisingly, we found that the averages are described fairly well by procedural values. In experiment 1, we found a set of good s-values at $t = 2$; in experiment 2, we also found good fits at $t = 4$. The regions of these suitable s parameters are plotted in Figure 5.8. We note that these regions overlap around the point $(0.65, 0.25)$, but neither region contains the egalitarian Shapley values or solidarity value.

There are several possible reasons why few participants can be described accurately by procedural values. One potential reason for this is that people may be performing a completely different type of calculation. As one participant mentioned explicitly in their questionnaire, the values of the single-player coalitions might serve as an initial source of information. Then, depending on how informative these values are, some people might feel that they understand the relationships between the players' power in the game without considering the two-player coalitions. There are many possible features that could cause people to quickly make a judgement about the game. Two examples are games where the single-player values sum to $f(N)$, or where one of the single-player coalitions has

$f(i) = f(N)$. In this model, instead of a straightforward sum across all of the coalitions, rewards might be calculated using a complex algorithm that relies on a number of simpler heuristics – similar to Selten's equal division payoff bounds for bargaining [84].

Another possible explanation is that there could be a non-linear utility function involved in the computation. If each person has a utility function $u(x)$, then they could be computing rewards for the game f_u defined by

$$f_u(C) = u(f(C))$$

for all coalitions C. Then, they could be computing procedural values on f_U instead of f. For example, a proper choice of utility function might help to explain the violations of additivity that we found in our data. However, there is a large space of possible models for these utility functions, so we choose to leave this problem for future work.

5.4.4 Limitations and Validity

This work gave new insights about the heuristics that people use to divide rewards in cooperative games. However, we could only study a limited number of games. What other factors might influence these reward divisions?

First, we chose a limited number of Shapley values and designed games around these values. We believe that other values could still expose new metrics that people apply to these games. For instance, in cases where the Shapley values are extremely far from an equal division, it would be interesting to see whether people are more or less willing to select these extreme values. Also, we intentionally chose games with round numbers: we used a total of 60 gold pieces so that division by 3 was possible, and all of our Shapley values were based on multiples of 5. Testing games with non-round numbers could cause effects related to the prominence of these values [3]: for instance, a coalition value of 10 might appear disproportionately larger than a value of 9.

Second, our results might depend on the framing of the game. In our experiments, we gave a story of three people playing a video game online. One participant explicitly mentioned this story in the surveys, stating that it is most common for parties to split their loot evenly in these types of games, regardless of the members' contributions. It would be interesting to study how this behaviour changes for different scenarios. One way to do this is to replace the video game setting with a merger negotiation between a number of companies. Another way is to have participants divide losses or costs instead of rewards. Reframing the games in this way might induce more calculated, rational behaviour.

Third, we presented our games in a tabular format. This form is one of the easiest to explain to participants, but it makes some types of calculations difficult. For example, to check whether one of the players is a null player, one needs to manually compare several coalitions' values. This type of information might be more easily gained by representing the games in different ways, displaying MC-net-like rules ([34]) or players' skills for a coalitional skill game [4]. A more succinct representation is also necessary in order to extend this research to games with more than 3 players.

Finally, the population participating in the experiment might be a significant factor. We used workers from Mechanical Turk, who are generally focused on completing their tasks as quickly as possible. Furthermore, beyond our experiments' tutorials, it is difficult for us to measure how much comprehension workers had, and we cannot detect workers' effort aside from our simple filtering criteria. Despite these potential issues, we still see significant value in our results. First, rejected work can have severe consequences, so many workers are remarkably careful and honest. Second, regardless of these issues, our data shows clear trends indicating the consistency of these workers in almost all of the games. Thus, while running these experiments through crowdsourcing might explain the high rate of equal divisions in many games, we believe that our data successfully captures human heuristics for these games.

Chapter 6

Conclusion

Paid crowd work is no longer only a solo endeavour. Crowdsourcing systems are growing to tackle larger, more difficult problems that can only be solved by allowing workers to collaborate, and understanding how to support groups of workers in these tasks is crucial to the future of crowdsourcing. In this thesis, we contributed to this effort by focusing on the problem of paying crowd workers for collaborative work, drawing on concepts from equity theory and cooperative game theory. Fair pay is not simply requesters' moral responsibility: theoretical literature suggests that paying groups of workers fairly is positively linked to trust, satisfaction, and motivation.

In Chapter 3, we carried out a literature review of existing collaborative crowdsourcing tasks. We identified four distinct types of information that workers often have available to them during this type of work. We also found that some problems, such as creative writing, difficult cognitive tasks, and tasks with subjective guidelines, can only be solved with explicit collaboration, where workers can readily make equity judgements about their wages. However, these existing tasks lack a systematic payment structure, instead relying on ad-hoc payment methods.

In Chapter 4, we selected two fair payment division methods from equity theory and cooperative game theory. Then, we used two experiments to test the effects of these payment divisions in a team-based audio transcription task. We found that workers are biased in their fairness judgements, but are perceptive of fair and unfair payments. Our data also suggests that fair payments could lead to small increases in worker effort that would be significant in certain types of tasks. Based on our findings, we argue that requesters and platforms can improve worker motivation, trust, and satisfaction by paying groups of workers fairly and transparently.

In Chapter 5, we took a closer look at the differences between human reward divisions and the axiomatically fair Shapley value. To do this, we used two experiments to find how people divide rewards in cooperative games while acting as impartial decision makers. Our results showed striking systematic trends in people's reward divisions that were often unrelated to the Shapley value. Further, while prior work showed that human reward divisions violate the null player axiom, we showed that they also break the additivity axiom. Although we could not find a model that fully captures people's decisions, our results highlight a fundamental issue with the Shapley value's axioms and open up a direction for future research on these axioms. Understanding the breakdowns in these fairness axioms will be important for building artificial agents that properly reason about human perceptions of fairness.

6.1 Broader Impacts

It is important to recognize the ethical issues that crowdsourcing researchers face. Crowdsourcing platforms incentivize low pay, with workers on Mechanical Turk making a median wage under US$2 per hour [28]. Finding new, difficult problems that workers can solve together could have the unfortunate consequence of attracting more low-paying requesters to the system. However, we believe that well-designed tasks with explicit teamwork are beneficial to the workers. Having workers cooperate can give them more information about their work, making it easier for them to avoid returning HITs or being rejected for misunderstanding a task – two of the biggest impacts on their hourly wage [28].

Emphasizing the role of teamwork in their tasks could also help workers become a stronger community. Turkers already rely on forums, discussion boards, social networks, and chat while working [26]. These communication channels help Turkers recognize good and bad tasks or requesters. We hope that emphasizing the role of teamwork in their tasks can help them become a stronger community with a more powerful voice to change the status quo of micro-task crowdsourcing.

Additionally, our proposed payments may appear to be in conflict with minimum wage standards. When one worker does not produce any useful input, both the proportional payments and Shapley values give no payment. This point could be an issue: sometimes, workers cannot complete their work due to factors out of their control, such as broken user interfaces or unclear instructions. However, both payments can adapt around this issue. For proportional payments, each worker's input can combine the amount of work they did with the amount of time they spent on the work, ensuring a minimum wage. Relaxing the Shapley value's null player axiom can also remove the requirement that null players

receive no reward – for instance, resulting in an egalitarian Shapley value. In both cases, it is possible to ensure a minimum wage for the workers as long as the group is given enough reward to pay its workers this minimum wage.

6.2 Future Work

The research presented in this thesis opens up several directions for future work; we discuss three of them here.

The first avenue for future work is to find a more precise description of human perceptions of fair payment divisions. When making equity judgements, equity theory does not specify exactly how each person's input should be quantified. These inputs could combine many components, such as skill, effort, or time, in countless ways. How much value do people put on each of these components? It seems likely that an answer to this question would be closely related to workers' egocentric biases; for instance, in our audio transcription task, fast typists might place more value on complete transcripts, while slow typists might appreciate effort or feel that they deserve their pay just for "showing up". While our experimental data could only give us a broad understanding of trends in workers' fairness perceptions, future studies could ask participants for more in-depth rationales to help identify how they make these equity judgements.

There is also space for future research on human fairness perceptions in terms of cooperative game theory. We found that people violate additivity when selecting reward divisions, but additivity is central to the uniqueness of the Shapley value. Can the Shapley value be modified to capture human fairness standards more faithfully? There are a myriad of possible modifications. One idea is to weaken the additivity axiom: a weak version of marginality [99] could be suitable, but other invariants could be uncovered by further analyzing our dataset. Another is to transform the game, applying a non-linear utility function or equalizing the players' contributions, before computing the Shapley value or a procedural value. People could also be modelled as performing a bounded number of calculations: perhaps their reward divisions can be described with a restricted form of stability or as the outcome of imperfect play in a bargaining game [76]. More work is required to test these conjectures.

The second direction for future research is to consider tasks where it is difficult to quantitatively measure the quality of each teammate's work. When the ground truth is not known, the quality of each worker's contributions could be estimated with crowd agreement scores [53] or peer prediction algorithms [70]. However, in other tasks such as

collaborative design work, there is no correct answer: workers may only have subjective opinions about how valuable their teammates are. This problem has received some theoretical attention as the "divide the dollar" game [20]. Practical systems have also used algorithms inspired by PageRank to divide credit between large teams of authors [91]. In collaborative crowdsourcing tasks, it would be challenging to guarantee that these mechanisms are robust against collusion – and further, to convince workers that these mechanisms are trustworthy.

The final direction for future work is to study tasks that involve collaboration between human workers and AI agents. Two examples of this type of task are Evorus [33], where chatbots suggest messages alongside human workers, and DreamTeam [101], where teams of workers are managed by Slack bots. In these tasks, workers could potentially feel that these AI agents are taking their work, lowering their pay. It would be interesting to study whether people's biases change in these scenarios – perhaps they make more unforgiving equity judgements when working alongside an AI. As human computation systems continue to incorporate more computational agents, it will be increasingly important to understand how these new types of tasks impact worker motivation.

References

[1] John Stacy Adams. Inequity In Social Exchange. In *Advances in Experimental Social Psychology*, volume 2, pages 267–299. 1965.

[2] Victor H. Aguiar, Roland Pongou, and Jean-Baptiste Tondji. A non-parametric approach to testing the axioms of the Shapley value with limited data. *Games and Economic Behavior*, 111:41–63, sep 2018.

[3] Wulf Albers and Glsela Albers. On the prominence structure of the decimal system. In *Advances in Psychology*, volume 16, pages 271–287. 1983.

[4] Yoram Bachrach and Jeffrey Rosenschein. Coalitional skill games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1023–1030, 2008.

[5] Sylvain Béal, Eric Rémila, and Philippe Solal. A decomposition of the space of TU-games using addition and transfer invariance. *Discrete Applied Mathematics*, 184:1–13, mar 2015.

[6] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, page 33, New York, New York, USA, 2011. ACM Press.

[7] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*, page 313, New York, New York, USA, 2010. ACM Press.

[8] Gary E. Bolton, Kalyan Chatterjee, and Kathleen L. McGinn. How communication links influence coalition bargaining: A laboratory investigation. *Management Science*, 49(5):583–598, may 2003.

[9] Colin Camerer. *Behavioural game theory: Experiments in strategic interaction*. Princeton Unversity Press, 2003.

[10] André Casajus and Frank Huettner. Null players, solidarity, and the egalitarian Shapley values. *Journal of Mathematical Economics*, 49(1):58–61, jan 2013.

[11] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational Aspects of Cooperative Game Theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, oct 2011.

[12] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011.

[13] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 2334–2346, New York, New York, USA, 2017. ACM Press.

[14] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 ACM annual conference on Human Factors in Computing Systems - CHI '19*, 2019.

[15] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2246–2266, Sep. 2016.

[16] D. Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A. Hearst. Structuring Interactions for Large-Scale Synchronous Peer Learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 1139–1152, New York, New York, USA, 2015. ACM Press.

[17] Jason A. Colquitt. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, 86(3):386–400, 2001.

[18] Peng Dai, Christopher H. Lin, Mausam, and Daniel S. Weld. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence*, 202:52–85, 2013.

[19] J H Davis, P R Laughlin, and S S Komorita. The social psychology of small groups: Cooperative and mixed-motive interaction. *Annual Review of Psychology*, 27(1):501–541, 1976.

[20] Geoffroy de Clippel, Herve Moulin, and Nicolaus Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139(1):176 – 191, 2008.

[21] Geoffroy De Clippel and Kareen Rozen. Fairness through the lens of cooperative game theory: An experimental approach. 2013.

[22] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 135–143, New York, NY, USA, 2018. ACM.

[23] Ryan Drapeau, Lydia B Chilton, and Daniel S Weld. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*, pages 32–41, 2016.

[24] Christian Fieseler, Eliane Bucher, and Christian Pieter Hoffmann. Unfairness by Design? The Perceived Fairness of Digital Labor on Crowdworking Platforms. *Journal of Business Ethics*, jun 2017.

[25] Ya'akov (Kobi) Gal, Moshe Mash, Ariel D. Procaccia, and Yair Zick. Which is the fairest (rent division) of them all? In *Proceedings of the 2016 ACM Conference on Economics and Computation - EC '16*, pages 67–84, New York, New York, USA, 2016. ACM Press.

[26] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. The Crowd is a Collaborative Network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 134–147, New York, New York, USA, 2016. ACM Press.

[27] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 2258–2270, New York, New York, USA, 2016. ACM Press.

[28] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, New York, New York, USA, 2018. ACM Press.

[29] Christopher Harris. Youre hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.

[30] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 419–429, New York, New York, USA, 2015. ACM Press.

[31] Shih-Wen Huang and Wai-Tat Fu. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 621, New York, New York, USA, 2013. ACM Press.

[32] Ting-Hao Kenneth Huang, Joseph Chee Chang, Saiganesh Swaminathan, and Jeffrey P. Bigham. Evorus: A Crowd-powered Conversational Assistant That Automates Itself Over Time. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17*, pages 155–157, New York, New York, USA, 2017. ACM Press.

[33] Ting-Hao (Kenneth) Huang, Walter S. Lasecki, Amos Azaria, and Jeffrey P. Bigham. "Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, pages 79–88, 2016.

[34] Samuel Ieong and Yoav Shoham. Marginal contribution nets: A compact representation scheme for coalitional games. In *Proceedings of the 6th ACM Conference on Electronic Commerce*, pages 193–202. ACM, 2005.

[35] Lilly C. Irani and M. Six Silberman. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 611, New York, New York, USA, 2013. ACM Press.

[36] Reinoud Joosten. *Dynamics, equilibria, and values*. PhD thesis, Maastricht University, 1996.

[37] James P Kahan and Amnon Rapoport. When you don't need to join: The effects of guaranteed payoffs on bargaining in three-person cooperative games. *Theory and Decision*, 8(2):97–127, 1977.

[38] James P. Kahan and Amnon Rapoport. *Theories of coalition formation*. Psychology Press, 1984.

[39] Gerhard K. Kalisch, John W. Milnor, John F. Nash, and Evan D. Nering. Some experimental n-person games. In Robert M. Thrall, Clyde H. Coombs, and Robert L. Davis, editors, *Decision Processes*, pages 301–327. 1954.

[40] Alexandre Kaspar, Genevieve Patterson, Changil Kim, Yagiz Aksoy, Wojciech Matusik, and Mohamed Elgharib. Crowd-Guided Ensembles: How Can We Choreograph Crowd Workers for Video Segmentation? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–12, New York, New York, USA, 2018. ACM Press.

[41] Nicolas Kaufmann, Thimo Schulze, and Daniel Viet. More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, New York, New York, USA, 2011. ACM Press.

[42] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P. Bigham, Jaime Teevan, Ece Kamar, and Walter S. Lasecki. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, pages 89–97, 2017.

[43] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 233–245, New York, New York, USA, 2017. ACM Press.

[44] Joy O Kim and Andres Monroy-Hernandez. Storia: Summarizing Social Media Content based on Narrative Theory using Crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 1016–1025, New York, New York, USA, 2016. ACM Press.

[45] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 37–46, New York, NY, USA, 2008. ACM.

[46] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM.

[47] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. CrowdForge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, page 43, New York, New York, USA, 2011. ACM Press.

[48] Steve WJ Kozlowski and Bradford S Bell. *Work groups and teams in organizations*, pages 333–375. Wiley Online Library, 2003.

[49] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, page 1003, New York, New York, USA, 2012. ACM Press.

[50] Anand Pramod Kulkarni, Matthew Can, and Björn Hartmann. Turkomatic: Automatic, Recursive Task and Workflow Design for Mechanical Turk. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2011.

[51] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*, page 23, New York, New York, USA, 2012. ACM Press.

[52] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. Apparition: Crowdsourced User Interfaces that Come to Life as You Sketch Them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1925–1934, New York, New York, USA, 2015. ACM Press.

[53] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, page 23, New York, New York, USA, 2011. ACM Press.

[54] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology - UIST '13*, pages 151–162, New York, New York, USA, 2013. ACM Press.

[55] Jesse D. Lecy and Kate E. Beatty. Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis. *SSRN Electronic Journal*, 2012.

[56] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, page 68, New York, New York, USA, 2010. ACM Press.

[57] Weichen Liu, Sijia Xiao, Jacob T. Browne, Ming Yang, and Steven P. Dow. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *ACM Transactions on Social Computing*, 1(1):1–26, jan 2018.

[58] Alan Lundgard, Yiwei Yang, Maya L Foster, and Walter S Lasecki. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–7, New York, New York, USA, 2018. ACM Press.

[59] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 260–273, New York, NY, USA, 2016. ACM.

[60] Ioanna Lykourentzou, Robert E Kraut, and Steven P Dow. Team Dating Leads to Better Online Ad Hoc Collaborations. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 2330–2343, New York, New York, USA, 2017. ACM Press.

[61] Marcin Malawski. Procedural values for cooperative games. *International Journal of Game Theory*, 42(1):305–324, feb 2013.

[62] Thomas W. Malone, Thomas W. Malone, and Kevin Crowston. The interdisciplinary study of coordination. *ACM Comput. Surv.*, 26(1):87–119, March 1994.

[63] Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J. Watts. An Experimental Study of Team Size and Performance on a Complex Task. *PLOS ONE*, 11(4):e0153048, apr 2016.

[64] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 224–235, New York, New York, USA, 2014. ACM Press.

[65] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. Turking in a global labour market. *Computer Supported Cooperative Work (CSCW)*, 25(1):39–77, Feb 2016.

[66] Michael Maschler. The bargaining set, kernel, and nucleolus. In Robert J. Aumann and Sergiu Hart, editors, *Handbook of game theory with economic applications*, volume 1, pages 591–667. Elsevier, 1992.

[67] Winter Mason and Duncan J Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '09*, page 77, New York, New York, USA, 2009. ACM Press.

[68] J.E. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[69] David M. Messick and Keith P. Sentis. Fairness and preference. *Journal of Experimental Social Psychology*, 15(4):418–434, 1979.

[70] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, 51(9):1359–1373, 2005.

[71] Meredith Ringel Morris, Jeffrey P. Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. Subcontracting Microwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 1867–1876, New York, New York, USA, 2017. ACM Press.

[72] Eugene W. Myers. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266, 1986.

[73] John F Nash, Rosemarie Nagel, Axel Ockenfels, and Reinhard Selten. The agencies method for coalition formation in experimental games. *Proceedings of the National Academy of Sciences*, 109(50):20358–20363, 2012.

[74] Andrzej S. Nowak and Tadeusz Radzik. A solidarity value for n-person transferable utility games. *International Journal of Game Theory*, 23(1):43–48, mar 1994.

[75] Andrzej S Nowak and Tadeusz Radzik. On convex combinations of two values. *Applicationes Mathematicae*, 24(1):47–56, 1996.

[76] David Prez-Castrillo and David Wettstein. Bidding for the surplus: A non-cooperative approach to the shapley value. *Journal of Economic Theory*, 100(2):274 – 294, 2001.

[77] Tadeusz Radzik and Theo Driessen. On a family of values for TU-games generalizing the Shapley value. *Mathematical Social Sciences*, 65(2):105–111, mar 2013.

[78] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *ICWSM International Conference on Web and Social Media*, pages 321–328, 2011.

[79] Michael Ross and Fiore Sicoly. Egocentric biases in availability and attribution, 1979.

[80] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[81] Niloufar Salehi and Michael S. Bernstein. Hive: Collective design through network rotation. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):151:1–151:26, November 2018.

[82] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 1700–1713, New York, New York, USA, 2017. ACM Press.

[83] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, nov 2018.

[84] Reinhard Selten. Equity and coalition bargaining in experimental three-person games. In Alvin E Roth, editor, *Laboratory Experimentation in Economics: Six Points of View*, pages 42–98. Cambridge University Press, 1987.

[85] Lloyd S. Shapley. A value for n-person games. In H. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*, pages 307–317. Princeton Unversity Press, Princeton, 2nd edition, 1953.

[86] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 937–945, New York, New York, USA, 2015. ACM Press.

[87] Robert Simpson, Kevin R. Page, and David De Roure. Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 1049–1054, New York, NY, USA, 2014. ACM.

[88] Maximilian Speicher and Michael Nebeling. GestureWiz: A Human-Powered Gesture Design Environment for User Interface Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–11, New York, New York, USA, 2018. ACM Press.

[89] Kate Starbird and Leysia Palen. "voluntweeters": Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1071–1080, New York, NY, USA, 2011. ACM.

[90] Leigh Thompson and George Loewenstein. Egocentric perceptions of fairness and interpersonal conflict. *Organizational Behavior & Human Decision Processes*, 51:176–197, 1992.

[91] Rajan Vaish, Snehalkumar (Neil) S. Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, James Davis, and Michael S. Bernstein. Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pages 829–843, New York, NY, USA, 2017. ACM.

[92] B. Vasilescu, V. Filkov, and A. Serebrenik. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *2013 International Conference on Social Computing*, pages 188–195, Sep. 2013.

[93] Mark E Whiting, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A Mullings, Yoni Dayan, Kristy Milland, Dilrukshi Gamage, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, Michael S Bernstein, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, and Angela Richmond-Fuller. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 1902–1913, New York, New York, USA, 2017. ACM Press.

[94] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, pages 379–388, New York, NY, USA, 2016. ACM.

[95] James R Wright and Kevin Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 901–907, 2010.

[96] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[97] Ming Yin, Yiling Chen, and Yu-An Sun. The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[98] Ming Yin, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. The Communication Network Within the Crowd. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 1293–1303, New York, New York, USA, 2016. ACM Press.

[99] H. P. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, jun 1985.

[100] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. Human computation tasks with global constraints. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 217, New York, New York, USA, 2012. ACM Press.

[101] Sharon Zhou, Melissa Valentine, and Michael S Bernstein. In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–13, New York, New York, USA, 2018. ACM Press.

[102] Yair Zick, Kobi Gal, Yoram Bachrach, and Moshe Mash. How to form winning coalitions in mixed human-computer settings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 465–471. International Joint Conferences on Artificial Intelligence Organization, aug 2017.

Appendices

Appendix A

Experiment Details

In each of our four experiments on Mechanical Turk, we gave workers instructions through an interactive tutorial. These tutorials included motivation for the experiment, information about the interface, and details about the experimental procedure. We also gave post-questionnaires at the end of each experiment. In this section, we give precise details about the content in each of these tutorials and post-questionnaires.

A.1 Study 1: Performance-Based Bonuses

A.1.1 Tutorial

- During this study, you will transcribe 15 audio clips (30 to 40 seconds per clip).
- This is a real-time transcription task: you will not be able to pause or replay the audio. It's okay if you miss words or make mistakes; we don't expect your transcriptions to be perfect.
- We will remove all punctuation and convert your transcript to lowercase.
- Transcribe the first audio clip now. We will use this first round to measure your initial skill level. Click the "Start Clip" button to begin.
- *(Workers transcribed the first audio clip)*
- At end of each of each audio clip, we will compare your transcript with 2 other previous workers. (In this example, we're showing 3 past workers.)

- For each worker, we will show you how many words they typed and how many were correct.
  - How many words did Worker 1 type?
  - How many words did Worker 1 type correctly?
- We will also show you a detailed view of their transcripts. Black words were typed correctly, red words were typed incorrectly, and grey words were not typed.
  - In Worker 2's transcript, what is the status of the word 'northeast'?
  - In Worker 2's transcript, what is the status of the word 'gone'?
  - In Worker 2's transcript, what is the status of the word 'come'?
- We will count how many words were typed by at least one worker. Then. we will give the team a total bonus of 5 cents for every 10 words.
  - How many words did the entire team type?
  - How many cents of bonus did the entire team earn?
- We will split this bonus between the three workers. (Note: if you transcribe every audio clip, we can use your transcriptions in future HITs, and we will award you bonuses as well.)
  - How many cents of bonus did Worker 3 earn in this round?
- Finally, we will ask whether you think these bonuses are fair to the three workers. You may answer 'Fair', 'Neutral', or 'Unfair' by clicking one of the buttons. Do this now to continue to the next audio clip. Thank you for participating!

A.1.2 Post-Questionnaire

Likert-type questions in this questionnaire gave options from 1 to 5 with anchors at 1 (strongly disagree) and 5 (strongly agree).

- Age
- Gender
- My bonus payments reflected the effort I put into this task. (1-5)

- My bonus payments were appropriate for the work I completed. (1-5)
- My bonus payments were justified, given my performance. (1-5)
- My bonus payments were acceptable. (1-5)
- I was satisfied with my bonus payments. (1-5)
- What factors did you consider when rating your bonus payments?
- Did you enjoy this task? Why or why not?
- Would you like to perform tasks in a team with other Turkers? Why or why not?
- Any other feedback:

A.2 Study 2: External Ratings

A.2.1 Tutorial

- In a previous study, we hired workers to transcribe 30 to 40 second audio clips.
- This was a real-time transcription task: workers were not able to pause or replay the audio. This is a dificult task, so we didn't expect their transcripts to be perfect.
- During this study, we will ask you to evaluate the finished transcripts from 16 different teams of 3 workers each.
- On this screen, we're showing you the 3 workers' transcripts. We removed all punctuation and converted the transcripts to lowercase.
- For each worker, we will show you how many words they typed and how many were correct.
  - How many words did Worker 1 type?
  - How many words did Worker 1 type correctly?
- We will also show you a detailed view of their transcripts. Black words were typed correctly, red words were typed incorrectly, and grey words were not typed.
  - In Worker 2's transcript, what is the status of the word 'northeast'?

  - In Worker 2's transcript, what is the status of the word 'gone'?
  - In Worker 2's transcript, what is the status of the word 'come'?
- After the workers finished the task, we counted how many words were typed by at least one worker. Then. we gave the team a total bonus of 5 cents for every 10 words.
  - How many words did the entire team type correctly?
  - How many cents of bonus did the entire team earn?
- Finally, we split this bonus between the three workers.
  - How many cents of bonus did Worker 3 earn for this audio clip?
- In this task, we will ask whether you think these bonuses are fair to the three workers. You may answer 'Fair', 'Neutral', or 'Unfair' by clicking one of the buttons. Do this now to continue to the next team. Thank you for participating!

A.2.2 Post-Questionnaire

- Age
- Gender
- What factors did you consider when rating the bonus payments?
- Would like to rate Turkers' work for other tasks? Why or why not?
- Would you like Turkers to rate your work? Why or why not?
- Any other feedback:

A.3 Experiment 1 and 2: Cooperative Games

Note that both of these experiments used the same interface: the only differences between the experiments were the games. Thus, we used the same tutorial and post-questionnaire for both experiments.

A.3.1 Tutorial

- In this experiment, you will be presented with a number of fictional scenarios where several people must decide how to divide a reward. In each of the scenarios, three people named Alice, Bob, and Charlie are playing a game together.
- We will show you information about how much reward every possible group of players could earn by working as a team. These rewards will be different in each scenario.
  - How many gold pieces would Bob earn alone?
  - How many gold pieces would Alice and Charlie earn by working together?
  - How many gold pieces would all three players earn by working as a team?
- Then, we will ask how you would split the team's reward if all three players worked together. You can chose your answer by dragging the sliders or clicking the buttons below.
- To continue, submit the following answer: 16 gold pieces for Alice, 20 for Bob, and 24 for Charlie. Thank you for participating!

A.3.2 Post-Questionnaire

- Age
- Gender
- What factors did you consider when splitting the rewards?
- Which was more important: the rewards the players could earn alone, or in pairs? Why?
- Do you think other participants would split the rewards differently? Why or why not?
- Any other comments or feedback: