

Peeking into the On-Demand Economy: Crowd Behavior and Incentive Design

a dissertation presented

by

Ming Yin

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

June 2017

© 2017 Ming Yin

All rights reserved.

Peeking into the On-Demand Economy: Crowd Behavior and Incentive Design

Abstract

An increasing number of digital and mobile technologies have emerged today to match customers, in almost real time, with a potentially global pool of self-employed labor, leading to the rise of the on-demand economy, which has brought about dramatic changes in our society. It creates new business models and new dynamics of labor allocation. It enables new models of computation, that is, human-in-the-loop computing. And it leads to new forms of knowledge creation—people all over the world are contributing to scientific studies in dozens of fields, either by making scientific observations as amateur scientists or by participating in online experiments as subjects. Despite its already significant impact, the on-demand economy has still been considered as a black-box approach to soliciting labor from a crowd of on-demand workers. Little is known about how the on-demand economy works and how it can work better.

In this dissertation, using one of the leading on-demand crowdsourcing platforms—Amazon Mechanical Turk—as an example, I present my findings in opening up the black box of on-demand economy. I investigate two lines of problems in this dissertation: first, I focus on understanding who the crowd of on-demand workers are and how they behave in on-demand work; second, I explore how effective incentives can be designed for on-demand work. Through a set of experimental studies, I provide a more precise picture of the on-demand workers, showing that they display significant temporal variations, value social interactions, and desire more flexibility and autonomy. Furthermore, based on a combination of experimental, computational and design methods, I also show the effectiveness of extrinsic financial incentives

in influencing on-demand workers, the feasibility of algorithmically controlling the provision of monetary rewards in a session of on-demand tasks in a cost-efficient way, as well as the potential of incorporating intrinsic motivator like curiosity in on-demand work through clever designs of task interfaces.

Contents

1	Introduction	1
1.1	Understanding Crowd Behavior	6
1.2	Designing Extrinsic and Intrinsic Incentives	9
1.3	Limitations	11
1.4	Contributions and Thesis Overview	12
2	The Temporal Dynamics of the Crowd	15
2.1	The Temporal Variations in Crowd Demographics	20
2.1.1	Experimental Design	20
2.1.2	Identifying the Time-Varying Dimensions	29
2.1.3	Capturing the Key Characteristics for Workers at Different Times	33
2.2	Scientific Studies with the Crowd: How Timing Influences Results	37
2.2.1	Experimental Design	38
2.2.2	Revisiting Worker Demographics	47
2.2.3	Influences on Studies Involving Incentivized Economic Decisions	49
2.2.4	Influences on Studies Examining Cognitive Abilities and Styles	54
2.2.5	Examining Differences in Worker Personality	56
2.3	Discussion	57
2.4	Acknowledgements	59

3	The Communication Network Within the Crowd	60
3.1	Related Work	63
3.2	Experimental Design	64
3.2.1	The Network Mapping HIT	66
3.2.2	Experimental Procedure	70
3.3	Results	70
3.3.1	A Network Enabled by Forums	72
3.3.2	Differences Between Subcommunities	75
3.3.3	The Role of One-on-One Communication	81
3.3.4	Homophily in the Network	82
3.3.5	Correlates of Network Position	84
3.3.6	U.S. vs. International Workers	86
3.4	Discussion	88
3.5	Acknowledgements	90
4	Managing the Flexibility of Crowdwork	91
4.1	Related Work	94
4.2	Experimental Design	96
4.2.1	The Sentiment Analysis Tasks	97
4.2.2	A 3×2 Factorial Design	97
4.2.3	Experimental Procedure	99
4.3	Results	103
4.3.1	Impact on Worker Engagement	104
4.3.2	Impact on Worker Performance	107
4.3.3	Impact on Working Behavior	108
4.4	Towards Measuring the Value of Flexibility	112

4.5	Discussion	115
4.6	Acknowledgements	118
5	Understanding the Effects of Financial Incentives	119
5.1	Placing Financial Incentives in a Task Sequence	122
5.1.1	Related Work	123
5.1.2	Experimental Design	126
5.1.3	Effects of the Magnitude of Rewards Alone	132
5.1.4	Improving the Effectiveness of Rewards in a Sequence	136
5.2	Providing Monetary Interventions in Task Switching	144
5.2.1	Related Work	145
5.2.2	Experimental Design	147
5.2.3	Effects on Intervened Tasks	151
5.2.4	Effects on Non-intervened Tasks	155
5.2.5	More Effective Interventions	157
5.3	Discussion	160
5.4	Acknowledgements	162
6	Monetary Intervention Design: An Algorithmic Perspective	163
6.1	Related Work	166
6.2	Predicting Work Quality under Monetary Interventions	168
6.2.1	Prediction Models	169
6.2.2	Evaluation Datasets	176
6.2.3	An Empirical Comparison of Model Performance	178
6.3	Controlling the Provision of Monetary Interventions Dynamically	186
6.3.1	Making Bonus Decisions with IOHMM	187
6.3.2	Experimental Evaluation with Real Crowd Workers	191

6.3.3	Examining the Robustness through Simulation	197
6.4	Discussion	201
6.5	Acknowledgements	203
7	Designing for Intrinsic Motivation: A Case Study on Curiosity	204
7.1	Related Work	206
7.2	Experimental Design	209
7.2.1	Operationalizing Curiosity	210
7.2.2	Task and Procedure	211
7.2.3	Experimental Conditions: Control and Treatments	212
7.2.4	Research Questions and Hypotheses	213
7.2.5	Choice of Article	215
7.2.6	Analysis Methods	216
7.3	Results	219
7.3.1	Effects of Curiosity Interventions on the Crowd	219
7.3.2	Individual Differences in Reaction to Curiosity Interventions	225
7.3.3	Interactions between Task Characteristics and Curiosity Interventions	227
7.4	Discussion	229
7.5	Acknowledgements	233
8	Conclusion	234
8.1	Summary of Contributions	235
8.2	Connections between Chapters	237
8.3	Future Directions	241
	Bibliography	247

Acknowledgments

I am very fortunate to go through my Ph.D. journey with the help, guidance and support from so many people, for whom I am very grateful.

First and foremost, I would like to express my deepest gratitude to my advisor Yiling Chen. Yiling, you are the best advisor a graduate student can ever ask for. Thank you for giving me the freedom to explore my research interests and patiently guiding me to develop my own taste of research. You taught me some of the most valuable lessons in conducting interdisciplinary research—to examine a problem from multiple angles, to think out of the box, and to always keep the high-level research questions in mind. Thank you for your willingness to listen, for always believing in me, and for all the care and support throughout this journey. I’m extremely fortunate to have you as my advisor.

I would like to thank the other three members on my committee: Barbara Grosz, Krzysztof Gajos and Siddharth Suri. Barbara, thanks for encouraging me to address challenging and critical problems and to pursue an academic career, and thanks for inspiring me with your passion not just for research, but also for influencing the next generation. I can only pay you back by passing it on. Krzysztof, I really appreciate all your inputs and suggestions to my research, and thank you for all the interesting cognitive experiments you shared with me. And Sid, I am so lucky to be able to work with some pioneering researchers in on-demand economy like you. Our collaboration since my internship at Microsoft Research NYC has broaden my view of interdisciplinary research and largely shaped my philosophy of research.

I also have the privilege to have learned from many other mentors and collaborators throughout my Ph.D. journey. David Parkes gave me many insightful suggestions on research and future career, and he has devoted to creating a nurturing environment for the EconCS group as well as an amazing department for the entire Harvard CS community. Yu-An Sun accepted me as an intern at Xerox Research, which was my first internship experience in

U.S., and such experience greatly helped me to carry out independent research from real-world problems. Edith Law introduced to me the HCI perspective of doing crowdsourcing research, and together we explored the potential of using curiosity as an intrinsic motivator in crowdwork and wrote a paper that we are very proud of. Working with Jennifer Wortman Vaughan during my internship at Microsoft Research NYC allowed me to quickly grow as a critical thinker, and I learned to always support empirical observations in research with mathematically rigorous evidence. Mary Gray mentored me during my internship at Microsoft Research New England last year, and she taught me to always think of socio-technical systems in their cultural, political and economic contexts as well as to consider the ethical and policy challenges they raise.

Being part of the Harvard EconCS group has been a really wonderful memory. I thank everyone in the group, including professors, students, postdocs, alumni, visitors and staff, for numerous stimulating conversations and for the great friendship. In particular, I would like to thank all my officemates over the years—Eric Balkanski, Emma Heikensten, Siri Isaksson, Dimitrios Kalimeris, Malvika Rao, Greg Stoddard, Bo Waggoner and Jens Witkowski. You guys made days in the office so much fun!

Importantly, I am thankful for each of the 31,339 on-demand workers who have participated in my study or experiment (this is probably an incomplete statistics though). This dissertation will not be possible without you, literally.

Outside of research, my thanks also go to many friends that I met during my time at Harvard, such as Ran Li, Hongyao Ma, Lan Wang, Di You and Qiaoying Zhang, to name a few. We are spreading at different corners of the world now, but thanks for accompanying me on this journey. I will always remember the beautiful days we shared together.

Yexiang, thank you for always being there to support me, intellectually and emotionally.

Mom and dad, as always, thank you so much, for your unconditional and endless love, for cheering each of my accomplishments, for picking me up when I fall. Thanks for everything!

Chapter 1

Introduction

It is a late Friday night. A group of Harvard alumni has just had a wonderful reunion night at their used to be favorite restaurant at Harvard Square and now needs a ride back to their places. Ten years ago, there was not much they could do other than standing on a corner hoping that some cab drivers would pass by. Nowadays, they can open an app on their mobile, press a few buttons, and within a few minutes, an Uber driver will show up to pick them up.

Meanwhile, in the Maxwell Dworkin building, a computer science graduate student is making a poster in preparation for presenting a paper of hers at a top-tier conference soon, but she is not completely satisfied with the current graphic designs yet. The best solution for her not long ago would be asking a friend—if she happens to know someone with good aesthetic—or hiring a local professional for help. Today, she can easily visit a website like Upwork, on which she can seek for advice from an excellent freelancer who is savvy at design and can be located at anywhere in the world.

Like Uber and Upwork, today, an increasing number of digital and mobile technologies have emerged to match customers, *in almost real time*, with *a potentially global pool* of self-employed labor, leading to the rise of *on-demand economy*. This technology-driven

on-demand economy has created disruptive new paradigms of transaction and production. On the one hand, the *efficient* matching between demand and supply allows customers to get access to whatever they want whenever they want it as easy as clicking a button; on the other hand, the *direct* matching between individual demand and individual supply also empowers more people to embrace a different way of working—they no longer work for a company or an organization, but instead, work for the “demand” and for themselves. Such profound impact of the on-demand economy has spread across various sectors of our daily life, from transportation to grocery delivery, to home cleaning, to legal services.

In addition to creating new business models and new dynamics of labor allocation, the on-demand economy has also led to new models of computation—it has enabled *human-in-the-loop computing*, which is one of the building blocks for the recent progress in artificial intelligence. For example, the renowned ImageNet project has made extensive use of Amazon Mechanical Turk, an on-demand crowdsourcing platform, to obtain accurate human annotations for over ten millions of images [Russakovsky et al., 2015], which makes it possible for computer vision researchers to train machine learning algorithms that surpass human-level performance in object recognition for the first time [He et al., 2015]. The on-demand labor has also been included in the feedback loop of computational processes or been asked to perform tasks that software can’t do, for a wide range of purposes such as improving search result relevance, filtering out inappropriate web content, or providing personal assistance in everyday life [Bridgwater, 2016]. In other words, the artificial intelligence technologies of today still need a degree of human intelligence in them [Gray and Suri, 2017], and the on-demand economy provides just the right kind of convenience for the exchange of such human intelligence.

More broadly, for the entire community of scientific researchers, the on-demand economy has also largely changed how knowledge is created today. Researchers are increasingly relying on on-demand platforms like Amazon Mechanical Turk to conduct inexpensive surveys and experiments with human subjects [Horton et al., 2011, Mason and Suri, 2012]. It is

estimated that in 2015 alone, more than 800 studies were published using data collected from Amazon Mechanical Turk, and these studies span dozens of fields from biomedicine to social sciences [Hitlin, 2016]. Various citizen science projects like Zooniverse and Foldit have also been set up, which attract a huge number of people all over the world to contributing to science, either by making scientific observations as amateur scientists or completing scientific tasks as needed. Zooniverse, for example, has more than 850,000 members¹ across the globe who volunteer to help with research projects in climate, space, literature, etc.

In many senses, the on-demand economy has opened up numerous exciting possibilities in different areas including business, computing and science. Despite all its already significant impact, the on-demand economy has still been considered as a *black-box* approach to soliciting labor from a crowd of on-demand workers. In general, people only have some vague ideas, if not misconceptions, on how the on-demand economy works and how it can work better. In this dissertation, using one of the leading on-demand crowdsourcing platforms—Amazon Mechanical Turk—as an example, I demonstrate my effort in opening up the black box of on-demand economy. In particular, I present a number of studies which provide a more fine-grained picture of the on-demand economy and resolve some misconceptions about it. I choose to conduct my investigation on the on-demand economy using Amazon Mechanical Turk as a starting point because it is one of the major on-demand platforms in the United States with a large worker pool. Moreover, the use of Amazon Mechanical Turk is widespread in both the industry (especially by IT and Internet companies) and the research community; therefore, study results on Amazon Mechanical Turk can be relevant for a wide range of users of the on-demand economy with different purposes, ranging from eliciting human intelligence to enhance artificial intelligence to conducting surveys and experiments with human subjects.

More specifically, to better understand today’s on-demand economy, I focus on obtaining

¹This statistics is retrieved from <http://edutechwiki.unige.ch/en/Zooniverse>.

in-depth knowledge on who the on-demand workers are and how they behave in the on-demand work, both individually and collectively. Such knowledge is very valuable for people to understand the commonalities and differences of on-demand workers in comparison with employees in the traditional economy, and to uncover the ways that the on-demand work gets done. In addition, it further provides insights to both practitioners and researchers on the opportunities and challenges in better leveraging the current on-demand economy given on-demand worker’s characteristics, and potentially improving the on-demand economy in the future to address worker’s needs and wants. Obtaining this knowledge can be particularly challenging though, because it is not practical for people to interview or observe on-demand workers on a large scale given that they can be physically apart from these globally-distributed workers. Although such distance to workers has been virtually eliminated by the digital communication protocols provided by on-demand platforms (e.g., the APIs), these protocols come with certain problems—for instance, they may only provide very limited information on the personal attributes (e.g., demographic information like age, gender, education, etc.) and social characteristics (e.g., whether a worker has friends who also do on-demand work) of an on-demand worker, let alone any detailed information on the procedure for a worker to complete her work (e.g., how a worker schedules her tasks). It is thus important to use innovative methods to collect such data, in order to answer a variety of questions in respect to the *behavior* of the crowd of on-demand workers, including how stable or varying the crowd is over time, whether there is any social interaction among them, and how they work in different tasks with different levels of temporal flexibility.

Of equal importance to understanding the on-demand economy of today is exploring possible ways to improve it in the future. In particular, as the on-demand economy presents a new model of work that differs from the traditional job, it is straightforward to consider multiple elements in work design and search for effective designs for the on-demand work to enhance its efficiency and sustainability. One of the central design elements here is

the design of *incentives* (or motivations), as incentives can largely direct one’s behavior. From the designer’s point of view, the key challenge in incentive design is how incentives should be structured and managed to induce desirable behavior from on-demand workers, such as encouraging active participation and maintaining the quality of work from workers. Psychological theories typically divide motivation into two types—*extrinsic motivation*, which is the desire to do something because it leads to a separate outcome, and *intrinsic motivation*, which is the desire to do something because it is inherently enjoyable [Ryan and Deci, 2000a]. Hence, I approach the problem of incentive design in the on-demand work from both directions and explore methods to motivate the on-demand workers extrinsically and intrinsically.

Corresponding to the above two lines of problems of on-demand economy that I am studying, this dissertation is composed of two parts. In the first part of the dissertation, I present a number of experimental studies to understand the behavior of crowd workers in the on-demand economy. Study results provide a more precise picture of the on-demand workers, showing that they display significant *temporal variations*, value *social interactions*, and desire more *flexibility and autonomy*. In the second part of the dissertation, I design effective incentives for the on-demand work. I empirically show the *effectiveness* of extrinsic financial incentives in influencing work quality and worker effort in on-demand work. Based on *quantitative models* on worker’s reaction to financial incentives, I illustrate the feasibility of *algorithmically controlling* the provision of monetary rewards in a session of on-demand tasks in a cost-efficient way. I also provide design principles for task interfaces of the on-demand work, which can be adopted to initiate intrinsic motivation, such as stimulating the *curiosity* of on-demand workers to engage and incentivize high performance from them.

At the core of this dissertation lies the application of an interdisciplinary, mixed-methods methodology. A fundamental aspect of this methodology is the design and deployment of *large-scale online experiments*. Such large-scale online experimentation allows me to collect rich behavioral datasets that enable the discoveries of some “*unknown unknowns*” about

human behavior, which may not only improve the understanding of on-demand workers but also help to answer important social science questions about humans in general. Moreover, through carefully designed randomized online experiments, I can also explore the “*known unknowns*”, for example, by making *casual inference* about human behavior, which guides me to employ effective interventions and advance the designs of on-demand work from an engineering perspective. On top of the experimentation approach, I add in the application of computational methods to further improve the on-demand economy in an algorithmic way, and the computational approach becomes especially powerful when integrated with the experimental approach—with the combination of these two approaches, models and algorithms can be designed to be aware of human behavior that is observed in the experimental data, and computational solutions can also be quickly evaluated, improved and iterated through experimentation.

1.1 Understanding Crowd Behavior

Understanding the ways the on-demand economy works requires us to understand who the crowd of on-demand workers are and how they behave, so that we can get a sense of how work gets done in the on-demand economy. The traditional black-box view of the on-demand economy has led to limited or even inaccurate perception of the crowd. For example, the promise of the on-demand economy to immediately provide *some* labor to fulfill a customer’s demand upon request (but without a detailed description of various features of the worker) has, to some degree, made it easy for people to neglect individual differences among on-demand workers. As a result, it is unclear, for example, how the population of on-demand workers *varies over time* in terms of their demographics. For researchers who utilize the on-demand economy to facilitate scientific discoveries, the lack of knowledge on the crowd dynamics is even more concerning—there is little evidence supporting or refuting

the existence of *temporal differences* within the population of on-demand workers regarding their economic behavior, cognitive abilities and styles, and personality, yet such evidence is crucial for researchers to interpret how robust their crowd-based findings are. In addition, our impression of on-demand workers has largely been shaped by how they have been “advertised,” which may actually lead to some misconceptions. For example, on-demand platforms typically attract workers by allowing them to complete the work whenever and wherever they want. Thus, it is not surprising that people have viewed the crowd as a group of independent workers who have enjoyed sufficient amount of flexibility to control their own work, and little attention is paid on *connections and social interactions* among workers or whether workers actually have *enough flexibility* in the on-demand work.

To obtain a more comprehensive and accurate understanding of crowd behavior in the on-demand economy, using workers on Amazon Mechanical Turk as an example, I have conducted a set of experimental studies to examine various aspects of the crowd.

First, I investigated the *temporal dynamics* within the population of crowd workers. Results suggested that on-demand workers who are available at different times in a day display significant variations in a number of dimensions in their demographics, and some distinctive features were extracted to characterize workers of different times of day. For example, for the population of U.S. on-demand workers recruited from Amazon Mechanical Turk, there are more inexperienced workers and West coast workers around 2am EST while workers available at 8am EST are significantly older, more experienced, more likely to be white and reside in the Northeast. To see whether conducting scientific studies on on-demand platforms at different times in a day may lead to different results, I further looked into whether temporal differences also exist in the ways that on-demand workers make incentivized decisions in economic games, the levels of cognitive abilities and types of cognitive styles they have, and the kind of personality they show. Experimental results indicated that while the crowd population has a rather stable composition in their cognitive abilities/styles

and personality over time, they exhibit different economic behavior within a day, including fluctuations in their tendency to cooperate as shown in the public goods game and variations in their risk attitudes as evident in the lottery choice game.

Next, I dispelled the notion of the crowd as a collection of independent workers by mapping the entire *communication network* of workers on Amazon Mechanical Turk. In particular, I designed and executed a task in which over 10,000 on-demand workers from across the globe self-reported their communication links to other workers; hence the communication network within the crowd was revealed for the first time. Experimental results showed that there is a rich network topology over a subset of roughly 13% of all the workers who took our task. I conducted further analysis to understand with *whom* workers communicate (e.g., is there any homophily in the communication network?), *how* workers communicate with one another (e.g., does communication happen primarily through online forums or one-on-one channels?), *what* workers communicate about, and the potential *influences* that communication exerts on workers. These findings implied that many on-demand workers value social interactions and have the needs to connect with others either virtually or in person. In other words, behind the scenes of the on-demand economy, there is a substantial amount of organic collaboration developed among on-demand workers.

Finally, I conducted a study on how on-demand workers react to *temporal flexibility* in their work. It was found that while many workers value the flexibility provided in the on-demand work in determining when and how long to work on on-demand platforms, they also find themselves to be much more *constrained within a task* due to the short amount of time allotted to the task. For example, after an Amazon Mechanical Turk worker accepts a task, she must complete the task within a pre-specified time limit in order to get paid. Such time constraint makes completing a task almost like taking an exam, and naturally leads to workers' desire for *more flexibility within tasks*, which is supported by our experimental results. It was observed that granting more in-task flexibility significantly improves the

engagement and performance of on-demand workers, and workers also behave differently in different tasks by, for example, leveraging the flexibility within a task to work at their own pace and schedule their workload in an efficient way. I further explicitly measured the economic values that on-demand workers associate with in-task flexibility and confirmed that about 65%–70% of the workers attach a positive value to it—in fact, it was estimated that on average, on-demand workers are willing to forego a financial compensation of \$0.82/hour or more for the ability to control their own time in the on-demand work.

1.2 Designing Extrinsic and Intrinsic Incentives

Designing effective incentives for the on-demand work is a key step for exploring how the on-demand economy can work better. One of the most widely used and studied ways to incentivize on-demand workers is to provide *extrinsic motivation* like *financial incentives*, yet the folk knowledge and early research suggests that the quality of work does not seem to be affected by how much a worker is paid [Mason and Watts, 2010, Rogstadius et al., 2011]. This is in stark contrast to the common belief of “people respond to incentives” and leads one to suspect whether financial incentives can be effective in encouraging higher effort and better work in the on-demand economy. Even if the answer is yes, for requesters who hire on-demand workers, using monetary rewards to motivate workers is not free. Hence, one critical challenge to address is how to trade off financial cost for work quality and offer monetary rewards in a way to maximize the requester’s overall utility. In addition, there have been numerous attempts to incorporate *intrinsic motivation* in the design of on-demand work, such as highlighting the meaningfulness of work [Chandler and Kapelner, 2013] and applying gamification techniques [von Ahn and Dabbish, 2008]. However, the efficacy of these approaches ranges quite a bit, making it necessary for us to further investigate the potential for using other types of intrinsic motivators in the on-demand work environment, and ideally

provide a principled way to do so.

I first resolved the concern about the effectiveness of financial incentives in the on-demand work by focusing on *performance-contingent financial incentives*, and carefully examining their impact on worker effort and work quality in a *sequence* of on-demand tasks (i.e., an on-demand work session) through randomized experiments. Our experimental results demonstrated that performance-contingent financial incentives *can* effectively motivate better worker performance. More specifically, in a session of tasks of the *same* type, although a higher payment level in each individual task doesn't necessary impact workers, increasing (or decreasing) the magnitude of financial incentives over the subsequent two tasks leads to increased (or decreased) levels of worker effort and work quality. In other words, workers respond more to the "relative magnitude" of incentives rather than the "absolute magnitude." This phenomenon can be related to an *anchoring effect* in worker's perception of the fair payment level—workers may compare the incentive in a task with the reference of fair payment in their minds before deciding their effort levels in tasks, yet their perceptions of fair payment can be largely influenced by the first payment level they receive in a task session. On the other hand, it was also showed that in a session of tasks of *different* types, performance-contingent rewards are most effective in improving worker performance when the task-switching frequency is low and rewards are placed at the *switching points*, where task type has just changed from one to another.

These experimental studies informed us of the importance of characterizing the effects of financial incentives on workers in the context of a task workflow, rather than for individual tasks. In light of this, I proposed a wide range of quantitative models, including supervised learning models, autoregressive models and Markov models, to capture the impact of financial incentives on on-demand workers in a sequence of tasks. I also conducted an empirical comparison across these models to better understand how well they can predict on-demand work quality under monetary interventions, especially in realistic scenarios where the size

of training data or the access to ground truth information is limited. Furthermore, using a particular type of Markov model, that is, the first-order input-output hidden Markov model, I developed an algorithmic approach that enables *dynamic provision* of monetary rewards to workers in an on-demand work session. This algorithmic approach solved the optimization problem for requester utility, considering model predictions on worker’s future performance as well as the requester’s tradeoff between quality and cost. Empirical evaluations on Amazon Mechanical Turk confirmed that the dynamic bonus policies designed using this algorithmic approach can increase requester utility in real on-demand work sessions by 27%–64%.

Finally, to complement studies on the design of extrinsic incentives, I also explored the possible application of *intrinsic motivation* in on-demand work contexts and demonstrated the potential of using *curiosity* as an intrinsic motivator for the on-demand work. In particular, based on the *information gap theory* of curiosity, the concept of curiosity has been operationalized into the task interface design of the on-demand work to create synergy between working on tasks and satisfying one’s curiosity. Examples are provided on designing curiosity interventions in the on-demand work with three design elements—information goal, gap salience, and incremental information reveal. Experimental results suggested that curiosity can be an effective intrinsic incentive to use in future on-demand work designs, as it may significantly improve worker engagement without degrading the worker performance, while the magnitude of its effect is influenced by both personal characteristics of the worker and the nature of the task.

1.3 Limitations

In this dissertation, the investigation on the on-demand economy is conducted on a particular on-demand platform (i.e., Amazon Mechanical Turk). Therefore, specific results may only be directly generalizable to similar platforms (e.g., other micro-task based, online

crowdsourcing platforms like Crowdfunder and ClickWorker), and may not be valid for other significantly different ones. For example, in examining the behavior of on-demand workers, I found that providing more temporal flexibility to Amazon Mechanical Turk workers by allotting more time in a task leads to higher levels of worker engagement and performance. But in other settings (e.g., for on-demand mobile apps like Uber), it may not be practical to provide flexibility in a similar way as it may hurt the core interest of customers (e.g., get a car as fast as possible). Obtaining a more comprehensive view of the on-demand economy for a more diverse set of on-demand platforms is an important next step to deepen the knowledge of on-demand economy.

To do so, it is necessary to both examine the external validity of results in this dissertation, and perhaps more crucially, to understand the unique challenges for different types of platforms that may stem from the specific industry that a platform serves (e.g., transportation). To that end, while specific findings may not generalize, the interdisciplinary, mixed-methods methodology described in this dissertation is generalizable. For example, large-scale online experimentation can be adopted to investigate the effectiveness of different communication messages in nudging Uber drivers to stay on the road [Scheiber, 2017]. Moreover, computational methods can be designed to model drivers' behavior (e.g., whether choose to stay on road, whether accept a passenger request) based on various factors (e.g., time, location, traffic, sensitivity to communication messages, etc.) as well as deciding on possible interventions (e.g., whether and to whom to send a communication message, and what to send) to influence drivers.

1.4 Contributions and Thesis Overview

This dissertation consists of two major components, in which I demonstrate my effort in opening up the black box of on-demand economy, using one of the leading on-demand crowd-

sourcing platforms, Amazon Mechanical Turk, as an example. Through experimental studies, the first component provides a quantitative account of workers in the on-demand economy: who they are and how they behave in the work. Contributions related to understanding crowd behavior include:

- Revealed the temporal dynamics of on-demand workers at different times of day in terms of demographics, economic behavior, cognitive abilities and styles, and personality. (Chapter 2)
- Mapped the communication network among more than 10,000 on-demand workers for the first time, and analyzed the scale and topological structure of this communication network, as well as the ways on-demand workers utilize and be influenced by this network. (Chapter 3)
- Characterized the impact of providing on-demand workers with more control over their own time during work on workers' engagement, performance and ways of completing the work, and further estimated the economic values on-demand workers attach to the temporal flexibility in their work. (Chapter 4)

The second component focuses on designing effective incentives for on-demand work to improve its efficiency and sustainability. Contributions related to incentive design in the on-demand economy include:

- An in-depth, empirical understanding on whether and when financial incentives are effective in influencing worker performance, for both on-demand work sessions of a single type of task and sessions with mixed types of tasks, as well as psychological explanations for the observed phenomenon. (Chapter 5)
- An algorithmic approach to predicting work quality under monetary interventions and dynamically controlling the provision of financial incentives in an on-demand work session to elicit high-quality work in a cost-efficient way. (Chapter 6)

- Design elements and principles that enable the incorporation of curiosity as an intrinsic motivator for the on-demand work through task interfaces. (Chapter 7)

Finally, Chapter 8 concludes and presents discussion for future research directions.

The work presented in Chapters 3, 5, 6 and 7 was produced in collaboration with Yiling Chen, Yu-An Sun, Mary Gray, Siddharth Suri, Jennifer Wortman Vaughan, Edith Law, Joslin Goh, Kevin Chen, Michael Terry and Krzysztof Z. Gajos, and was previously published as conference papers. Pointers to specific papers are provided at the end of each chapter. The research in Chapter 2 (in collaboration with Yiling Chen, Emma Heikensten, and Anna Dreber) and Chapter 4 (in collaboration with Mary Gray and Siddharth Suri) are unpublished working papers at the time of writing this dissertation.

Chapter 2

The Temporal Dynamics of the Crowd

The on-demand economy has created an efficient way to connect the supply and demand of labor, no matter it is for one-off services like driving, or small piece of information work like image annotation, or scientific studies like short surveys and experiments. For example, on Amazon Mechanical Turk (MTurk)¹, an on-demand crowdsourcing platform, requesters of the labor can use an API provided by the platform to post small jobs, referred to as *human intelligence tasks* or *HITs*, along with specified time limits and payments for completing each HIT. A typical HIT might involve translating a paragraph of text, labeling an image, or completing a survey. Workers can then browse available HITs and choose HITs to work on in exchange for the pre-specified payments. With a global, on-demand workforce, requesters can easily get thousands of HITs done in a very short amount of time.

While enjoying the convenience of quick access to the supply of labor, requesters often get very limited information on who the workers that they interact with are through the digital communication protocols, defined by the API of the on-demand platforms—for example, on MTurk, the personal attributes of workers, such as age, gender, and ethnicity, are all hidden from requesters. To address this problem, a large number of studies have been conducted to

¹<https://www.mturk.com/>

understand the demographic, geographical, political, occupational and motivational information for on-demand workers from various platforms [Ipeirotis, 2010, Paolacci et al., 2010, Huff and Tingley, 2015, Avery et al., 2016, Hitlin, 2016, Intuit, 2017]. In particular, as on-demand platforms like Amazon Mechanical Turk has become increasingly popular among researchers as a source of data collection, especially for recruiting human subjects for scientific studies, there have been considerable effort in examining how “representative” the samples of workers on on-demand platforms are. It is found that U.S. subjects recruited from MTurk are more demographically diverse and thus more representative of the U.S. population compared to various convenience samples like the typical American college samples [Buhrmester et al., 2011, Berinsky et al., 2012]. However, it still exhibits notable differences from national probability samples obtained through face-to-face interviews, suggesting that it is not representative of the population as a whole [Berinsky et al., 2012].

In addition, researchers and practitioners have also noticed that the composition of the on-demand worker population may change from time to time. For example, it is reported that from March 2008 to February 2010, the crowd worker population on Amazon Mechanical Turk has largely shifted from a primarily moderate-income, U.S.-based workforce towards an increasingly international group with a large number of young, well-educated Indian workers [Ross et al., 2010]. More recently, Stewart et al. [2015] has estimated that the time taken for half of the workers to leave the MTurk pool and be replaced is about 7 months. These observations raise an important issue for understanding who the crowd of on-demand workers are that deserves in-depth research, that is, the *temporal dynamics* of the crowd.

Arguably, the variations in the on-demand worker population observed in [Ross et al., 2010, Stewart et al., 2015] all represent a “*macro-level*” temporal dynamics that span over a period of months or years. However, the crowd may also exhibit some “*micro-level*” temporal dynamics within individual days. Indeed, the high mobility of workers in the on-demand economy implies that the worker who picks up a task at a particular time of day is someone

who *happens to be available* at that time, suggesting the temporal differences within the on-demand worker population in a day can be significantly larger compared to employees in the traditional economy. Such micro-level temporal dynamics has received much less attention, perhaps because people tend to treat each individual worker within a short timeframe as, more or less, “interchangeable,” especially concerning what kind of work the worker can complete².

The problem of better understanding the micro-level temporal variations in the on-demand worker population is especially relevant for scientific researchers who conduct studies with the crowd, though. In particular, with on-demand platforms like Amazon Mechanical Turk, researchers can essentially collect experimental data at any time that they want. As such, one may wonder whether the subject samples that researchers obtain at different times (e.g., in the morning or at night) differ from each other. Moreover, from an experimenter’s point of view, it is also natural to ask whether conducting an experiment on these platforms at different times can possibly lead to different experimental results.

Recently, a few studies have been conducted to provide some initial knowledge on the micro-level temporal dynamics within the population of on-demand workers. For example, [Difallah et al. \[2015\]](#) have developed a website called “Mechanical Turk Tracker”³, which keeps track of various activities on Amazon Mechanical Turk, including monitoring the fluctuations in its worker demographics through periodical surveys. Researchers have examined whether the *demographic composition* of experiment participants that they recruit through MTurk varies by time of day, day of week and the serial position in the experiment (i.e., whether a subject participates in the experiment in the early stage or the late stage) [[Casey et al., 2017](#), [Arechar et al., 2016](#)]. Moreover, [Arechar et al. \[2016\]](#) further looked into the differences

²For example, each worker on on-demand crowdsourcing platforms may be viewed as near-identical CPU that perform computations for pay [[Suri, 2016](#)].

³<http://www.mturk-tracker.com/>

in *worker behavior* at different times, especially in terms of workers’ incentivized decisions involving prosociality, punishment and discounting (e.g., worker’s decisions in prisoner’s dilemma, dictator games, charitable giving, time discounting, etc.). Results in [Casey et al., 2017, Arechar et al., 2016] indeed point out a few inter-temporal differences in terms of the subject demographics—for instance, participants at different times of day differ from each other in dimensions like the time zones they come from, marital status and experience levels. However, it is also reported that participants at different times of day don’t seem to exhibit significantly different behavior in various economic games.

Interestingly, the findings in both [Casey et al., 2017] and [Arechar et al., 2016] are based on data from experiments in which workers are *not* allowed to participate for more than once. In other words, the reported temporal differences for workers at different times of day in these studies actually represent the variations of experiment participants in one experiment, when the experimenter *leaves that experiment continuously open* for a very long time while sampling participants from the underlying worker pool *without replacement*⁴. However, as workers are naturally presented on on-demand platforms at different times of day, it can be biased to use these results to interpret whether the *available workers* at different times of day differ in their demographic characteristics, as well as whether an experimenter will get the same experimental results if, *hypothetically*, he launched his experiment on the platform in different time slots (e.g., whether launching an experiment during 8am–9am leads to the same experimental results as that in the scenario when the experiment is launched at 5pm–6pm). A solid understanding on these questions is critical for researchers who leverage the on-demand economy as a way to facilitate scientific studies, as it will give them a sense of how robust the findings that they derive from crowd-based online studies are. Moreover, such understanding

⁴In [Arechar et al., 2016], the authors also calculated each participant’s “experienced time,” that is, the time that subjects participate in the experiment according to their *local* time zones, and analyzed the temporal differences using subjects’ experienced times.

will also inform researchers on how to conduct their surveys or experiments on on-demand platforms and communicate their discoveries in an appropriate way.

Thus, in this chapter, we attempt to understand the micro-level temporal dynamics of the crowd by analyzing the differences across on-demand workers that are available at different times of day, and we further examine whether the timing of an experiment on on-demand platforms like MTurk has any influence on experimental results. An ideal approach to answering these questions would be to artificially create a few “parallel universes” of the on-demand platforms—for example, we may build a number of different versions of the MTurk website, and each MTurk worker is randomly assigned to use one of them. These different versions of the MTurk website are identical in all aspects except that the experiment that we are interested in studying will be posted in each version of the website at a different time (all other tasks will be posted on all versions of website in exactly the same way). Such design would allow us to construct a few plausible *counterfactual worlds*, and we can therefore answer our research questions by comparing the worker demographics and experimental results for the experiments that we conduct in each “world.”

This ideal approach is hardly practical, though, given that we are not the provider of the platform. Therefore, in this chapter, we propose a few innovative experimental designs to address our research questions without actually creating the parallel universes (at least not at a full scale), and we limit our attention to examine the temporal dynamics of crowd workers on MTurk in individual days. In Section 2.1, we create an experiment to understand the temporal variations in demographics for available workers on MTurk throughout a day, by collecting worker samples every 3 hours through different tasks posted from different requester accounts. Based on the collected data, we identify a few dimensions of demographics for which significant differences are observed across the available MTurk workers at different times, and we further extract a few distinctive features to characterize workers at different times. These results reflect that there is *indeed* certain fluctuation in the composition of

workers on MTurk throughout a day. Furthermore, in Section 2.2, we examine whether the results of scientific studies conducted on MTurk can differ by time, or more specifically, whether the available crowd workers at different times of day display significantly different economic behavior, cognitive abilities and styles, and personality. We answer these questions using a two-phase experiment which is designed to approximate the idea of “parallel universes” with a representative sample of the entire MTurk worker population. In our experiment, no significant difference is observed in terms of the cognitive abilities and styles or the personality of workers at different times of day. However, we do find that for studies that involve some incentivized decisions from the workers (e.g., classical behavioral economic experiments like the public goods game or the lottery choice game), it is possible that the timing of the study will change its result to some degree. In other words, for some scientific experiments that are conducted with the crowd, the timing of the experiment may in fact exert notable influences on its result. We finally discuss the implications of our findings in Section 2.3.

2.1 The Temporal Variations in Crowd Demographics

In this section, we aim at thoroughly understanding temporal variations in the demographic characteristics of workers who are available on on-demand platforms throughout a day—what are the key differing dimensions and what kind of characteristics workers at different times have? To answer these questions, we conducted an empirical investigation on Amazon Mechanical Turk.

2.1.1 Experimental Design

Existing work that examines the temporal variations of the crowd for different times in a day is based on experiments that prevent worker from participating more than once [Casey et al., 2017, Arechar et al., 2016]. Hence, they can not be used to accurately interpret the

differences among the *available workers* of the platform in different time slots. The key limitation here is that by leaving an experiment open for a long time and restricting each worker to participate in the experiment only once, the obtained worker sample in the later stage of data collection can be biased. This is because at the later stage of data collection, a fraction of available workers then may be forbidden from participating in the experiment again if they have already participated in the experiment earlier, but in fact, they should have been allowed to participate as they are actually presented on the platform at that time.

An alternative design is to leave the experiment open for a long time while allowing each worker to participate for multiple times⁵. However, this design may only attenuate the bias in worker sample in the later stage of data collection at best, because workers may find it boring to complete the same task more than once and therefore choose not to do so. To solve this problem, we may consider another design, in which the experimenter posts a different HIT at a different time in the day, but content-wise, these HITs are similar enough to attract the same pool of workers on the platform. Workers are allowed to participate in as many of these different HITs as they wish, as long as they are available when the HIT is posted. One issue raised by this design is that workers may start to “follow” the experimenter’s requester account (possibly using some scripts) if they find that various kind of tasks provided by this requester are all interesting and well-paid. As a result, these workers will be immediately notified and may come back to work once the experimenter posts new tasks on the platform, even though the tasks are not posted during the time that they typically work on MTurk. In other words, we may create new bias with this design, and such concern is especially serious if, for example, we use a requester account to post different tasks at a fixed interval (e.g., post a different kind of task every 3 hours), and workers manage to figure out this interval.

⁵For example, researchers have adopted this kind of approach to capture the time variability in the demographics of MTurk workers by posting one 5-question demographic survey every 15 minutes, and each worker is allowed to take this survey once per month [Ipeirotis, 2015]. See <http://demographics.mturk-tracker.com/> for the results.

A third design is to create multiple requester accounts and post the same HIT at different times using each of the requester accounts in turn. Yet, from the worker’s point of view, this design may seem to be a bit suspicious to them as they will find that many requesters are posting the exact same task in a day.

Given the limitations on all possible designs that we have discussed above, in this study, we present a fourth, innovative experimental design that we believe to be *natural* while *minimizing the possible bias* in collecting worker samples at different times as much as possible. Specifically, we create 8 requester accounts in total, and each requester account is associated with a HIT that contains a *pre-task demographic survey*, a unique *cognitive task*, as well as a *post-task exit survey*. Every 3 hours, we use a *different* requester account to post its HIT. We thus can collect $24 \div 3 = 8$ worker samples from different times in a day, and each worker sample roughly contains 100–200 workers who participate in the same HIT that is posted from one requester account at a particular time slot. In other words, we collect data on worker demographics at different times in a day through posting different HITs from different requester accounts. In addition, to see whether any possible demographic difference that we observe in the collected worker samples is limited to a particular day, we repeat this process for 5 consecutive workdays. Therefore, in total, we get $8 \times 5 = 40$ worker samples from different days and times, with 5 samples in each particular time slot (e.g., for the 8am slot, we obtain 5 worker samples, each roughly of size 100–200, from Monday, Tuesday, Wednesday, Thursday and Friday, respectively). We provide more detailed information about this experimental design below, especially on a number of design decisions we make to ensure its validity—on the one hand, we try to make sure that the 8 HITs associated with the 8 requester accounts actually look like (ideally irrelevant) tasks from different requesters, so workers will not find it unnatural to see similar tasks from different requesters; on the other hand, we also try to keep all 8 HITs as similar as possible *at the preview stage*, so that we can use them to approach to the same pool of subjects on MTurk and thus minimize the

chances that differences in worker demographics across different samples, if any, are a result of each HIT attracting a different subpopulation of workers.

Pre-task demographic survey. In the pre-task survey of each HIT, we ask workers seven questions regarding their demographic background, including their gender, location (i.e., the state that they currently live in), age, highest level of education, ethnicity, race and religion. To make the 8 HITs look more different, for each question, we randomly select one way to state it from three alternatives (e.g., for the question on gender, the three alternatives are “Gender”, “What is your gender?” and “Please indicate your gender.”). Furthermore, the order of questions in the survey is also randomized.

Cognitive tasks. Including a *unique* type of cognitive task in each HIT is the key step for making the 8 HITs different. More specifically, we consider the following eight types of cognitive tasks:

- *Social intelligence* (Requester account 1): In each task, the worker is shown a pair of eyes with four emotion labels around it and asked to select the word that best describes the emotion that the eyes are showing. Such ability is observed to be related to worker performance on team-based problem solving tasks [Baron-Cohen et al., 2001].
- *Nutrition intelligence* (Requester account 2): In each task, the worker is presented with a pair of photographs of meals and asked to answer a nutrition-related question, such as “which meal has more fat.”
- *Sleight of hand* (Requester account 3): In each task, the worker is shown a picture of a hand and asked to guess whether the hand on the screen is a left or right hand. In addition, in some tasks, we inform workers that the picture is presented as a mirror image and thus workers should reverse their answers (e.g., if a worker sees a left hand in a mirror image, it is actually a right hand).

- *Reaction time* (Requester account 4): In each task, the worker sees a red square on the screen, which may change into another shape at any time. The worker is asked to click on the shape as quickly as she can after the change happens.
- *Thinking style* (Requester account 5): In each task, the worker is shown three words (e.g., “seagull”, “sky” and “dog”) and asked to click on the two words that she feels go together best. This task, initially designed by Ji et al. [2004], determines whether an individual tends to group information holistically or analytically.
- *Spatial intelligence* (Requester account 6): In each task, the worker is asked to answer a question that tests her ability to comprehend 3D images and shapes (e.g., select an unfolded object that can be folded into a target cube).
- *Face recognition* (Requester account 7): In each task, the worker is shown 25 faces and has 1 minute to remember them. When the time is up, the worker gets a sequence of faces and for each of them, she is asked to decide whether that face is among the 25 faces that she previously sees.
- *Memory test* (Requester account 8): In each task, the worker first sees a screen showing anywhere between 1 and 6 symbols (each symbol is a numeric digit) and she is asked to memorize these symbols in 6 seconds. Then, the worker sees a series of screens, each showing one symbol at a time and asking her to decide whether or not that symbol is present in the set that she has just memorized.

Four of these tasks (social intelligence, nutrition intelligence, thinking style and memory test) are adapted from experiments on the online experimental platform LabintheWild⁶ [Reinecke and Gajos, 2015], and the other four tasks are designed by us. As the cognitive task is the main body for each HIT, workers may naturally treat the HITs we post at different times from the corresponding requester accounts as *different tasks from different requesters*.

⁶<http://labinthewild.org/>

Therefore, it’s likely that they will be willing to participate in multiple tasks at different times as long as they are online and interested in the tasks.

Post-task exit survey. In the post-task survey of each HIT, we ask a few more questions regarding workers’ demographics, including:

- *working experience on MTurk* (4 questions): the number of years using MTurk, the number of HITs completed, the number of HITs completed in the last month, approval rate on MTurk;
- *levels of communication related to MTurk work* (2 questions): usage of online forums related to MTurk work, the number of workers communicated with in the last week about MTurk work;
- *income/household information* (6 questions): personal income, household income, the number of income earners in the household, the number of children under 18 in the household, whether MTurk is primary source of income, whether hold other jobs outside MTurk;
- *experience in participating scientific experiments on MTurk* (2 questions): whether participated in scientific experiments on MTurk in the last month, the number of scientific experiments participated in the last month.

Similar to the pre-task survey, to further differentiate the 8 HITs, we also state each question in the post-task survey randomly from three possible options, and the order of questions are randomized whenever appropriate⁷.

Since in this study, we are mainly interested in understanding the temporal variations in worker demographics at different times of day, the main purpose for posting the 8 HITs

⁷We keep the order of the questions fixed for some categories of questions like the income/household information, because for example, it may make more sense to ask about personal and household income subsequently.

from different requester accounts every three hours is to collect worker responses to the demographic questions in the pre-task and post-task surveys in the HITs. Dividing the 21 demographic survey questions into two parts (i.e., the pre-task survey and the post-task survey) is also a part of our design choice. Ideally, because we aim at analyzing the variations across *available workers* at different times, we would put all survey questions at the beginning of the HITs. In this way, we can collect the demographic information for as many workers who have ever *accepted* our tasks as possible, even though some of them may eventually drop out. However, if we do so, workers may find that a number of different HITs from different requester accounts share a long and somewhat similar survey at the start, which can possibly lead them to suspect whether these HITs are “actually” from different requesters. To address this concern, we decide to only keep the 7 basic demographic questions before the cognitive task and leave the other 14 questions after the cognitive task. With this design, we can at least collect as much information as possible on the basic demographic without incurring unnecessary suspicions among workers, as it is very common for requesters to ask workers about these basic demographic information before workers entering the actual task.

Experimental procedure. We conducted our experiment from August 1, 2016 (Monday) to August 5, 2016 (Friday). Each day, HITs were posted at eight time slots (i.e., 2am, 5am, 8am, 11am, 2pm, 5pm, 8pm and 11pm; all according to the Eastern Standard Time), and at each time slot, we used a *different* requester account to post the HIT associated with it. Besides, to minimize the chance for workers to follow requester accounts, each account was used at different time slots across different days. Table 2.1 provides a detailed schedule on how we posted HITs in our experiment.

As we posted HITs 8 times a day for 5 days, we can think of our experiment as an aggregation of 40 *sub-experiments*. Hence, we collected 40 worker samples in total, one for each sub-experiment. Notice that the expiration time limits for all HITs were set to be 1

	August 1 (Mon.)	August 2 (Tue.)	August 3 (Wed.)	August 4 (Thu.)	August 5 (Fri.)
2am	Account 1	Account 6	Account 5	Account 4	Account 2
5am	Account 2	Account 8	Account 7	Account 1	Account 6
8am	Account 3	Account 1	Account 4	Account 6	Account 5
11am	Account 4	Account 3	Account 2	Account 7	Account 8
2pm	Account 5	Account 7	Account 8	Account 3	Account 4
5pm	Account 6	Account 4	Account 1	Account 8	Account 3
8pm	Account 7	Account 5	Account 6	Account 2	Account 1
11pm	Account 8	Account 2	Account 3	Account 5	Account 7

Table 2.1: Schedule for posting HITs in the experiment that was conducted from August 1 to August 5, 2016. For example, at 2am EST, August 1, 2016, we used requester account 1 to post the HIT associated with it (i.e., the HIT with a social intelligence task as its cognitive task).

hour, which means that the worker sample we collected for each time slot was composed of workers who were available in an interval of 1 hour (e.g., the 2am worker sample were made of workers who were available between 2am and 3am). Each worker was restricted to take part in each sub-experiment only once. However, workers can participate in as many sub-experiments as long as they are available. That is, workers can both take multiple HITs from different requester accounts within a day and take HITs from the same requester account across different days. In fact, to make sure that a worker would be willing to take HITs from the same requester account across different days (if the worker happens to be available at those times when these HITs are posted), for a given requester account, we also used different contents for the cognitive task in the HIT on different days (e.g., the set of eye pictures used in the HIT of requester account 1 on Monday was different from that used on Tuesday).

Importantly, we adopted a few approaches to ensure that all 8 HITs in our experiment look similar at the preview stage so that they attracted the same type of workers on MTurk. For example, all HITs were advertised on MTurk as short cognitive experiments, though the wordings were slightly different. Furthermore, the first page of all HITs, which is the only page workers would be able to see at the preview stage, had the same layout (it is the layout for HITs that were created through templates on MTurk) and contained only very general

information about the HIT without any detailed instructions on the specific task. These controls give us the confidence that if we observe any difference across worker samples for different time slots, such observation is not likely to be an artifact of that HITs posted at different times attracted different types of workers.

After the preview stage, once a worker decided to take a HIT and proceeded on to the second page, the worker would see a detailed instruction about the cognitive task that was used in that HIT. Starting from the second page of each HIT, we used different fonts, background images, and color schemes for different HITs, which can possibly help to reinforce the perception that these HITs were from different requesters. The worker then needed to complete the pre-task demographic survey, the cognitive task as well as the post-task exit survey in the HIT. At the end of the HIT, the worker also got a personalized feedback on their performance in the cognitive task and would be paid with a fixed amount of 50 cents after she submitted the HIT. Our experiment was open to U.S. workers only.

Experimental data. 3,998 unique workers participated in at least one sub-experiment and in total, they completed 9,132 pre-task surveys. Among these 3,998 workers, 1,937 (48.4%) workers participated in at least two sub-experiments, yet we found 478 of them were inconsistent with themselves in answering questions in the pre-task demographic surveys (i.e., the worker provided different answers for at least one of the 7 demographic survey questions among all the sub-experiments that she participated in). We therefore excluded all 1,969 pre-task survey responses from these 478 workers from further analysis. As a result, our analysis on the temporal differences in worker’s basic demographic information is conducted on 7,163 pre-task survey responses collected from 3,520 unique workers.

Regarding post-task surveys, we found that 3,405 out of the 3,520 workers who were preserved after the previous data cleaning process actually completed the post-task surveys, and they generated 6,711 responses in total. Again, among these 3,405 workers, 254 of them

were identified as inconsistent with themselves when responding to a few selected questions in the post-task survey⁸. After removing the 982 responses from these 254 workers, we conducted our analysis on demographics in the post-task survey based on 5,729 responses from 3,151 unique workers.

Notice that on average, for each sub-experiment, we collected the basic demographics from a sample of 179 workers through the pre-task survey, as well as some more detailed demographics from a sample of 143 workers using the post-task survey. In practice, it’s common for scientific researchers to conduct an experiment on MTurk to recruit 100–200 workers for each treatment, which suggests that the average size of worker samples that we collected in each of our sub-experiment is representative of that for a typical crowd-based experiment. In other words, if we can observe significant differences in the demographics for the 8 worker samples we collect within one day, it may imply that when an experimenter conducts an experiment of normal size at different times of day, he may approach to collections of subjects with different demographic backgrounds.

2.1.2 Identifying the Time-Varying Dimensions

Our first goal is to examine whether available workers at different times of day have different demographics, and if yes, what are the key dimensions of demographics that workers of different times differ from each other.

We first answer these questions from an *aggregated* level. That is, given a particular time slot (e.g., 2am), we combined the data we get in that time slot from each of the 5 days together. In this way, we created a set of 8 *aggregated worker samples*, one for each time slot. Comparing the worker demographics across the 8 aggregated samples then helps us to

⁸These questions are the number of income earners in the household, the number of children under 18, whether MTurk income is the primary source of income and whether the worker holds a job outside MTurk. These 4 questions were selected to check worker’s self-consistency as answers to these questions are not likely to change in a short period of time.

understand the temporal dynamics of the crowd from a population point of view.

We examined the temporal variations of the worker demographics in terms of each of the 21 survey questions that we asked. For each survey question, we coded a few dependent variables according to possible responses to that question, and each dependent variable represents a certain aspect of worker demographics. For example, for the question on the highest level of education, we created 3 dependent variables—the percentage of workers whose highest education is high school or lower in a particular worker sample, the percentage of workers whose highest education is some college or equivalent, and the percentage of workers whose highest education is bachelor degree or higher. In total, we created 55 dependent variables based on all survey questions. Next, for each dependent variable, we attempted to examine whether there is any difference in terms of the value of this variable across the 8 aggregated samples, with the null hypotheses being that the values are all equal across different samples. For continuous dependent variables (e.g., the age of workers), one-way analysis of variance (ANOVA) [Wasserman, 2003] or Kurskal-Wallis tests [Kruskal and Wallis, 1952] was used for the tests depending on the distributions of the data, and for proportions (e.g., the percentage of workers living in California), proportion test [Wasserman, 2003] was used for the hypothesis testing.

Given that in total, we conducted 55 hypothesis testings, to control the family-wise error rate to be at the level of $\alpha = 0.05$, that is, to ensure the probability of rejecting at least one true null hypothesis (i.e., making at least one type I error among all hypothesis testings) to be at most 0.05, we apply the Bonferroni correction [Frank Bretz and Westfall, 2011] and only report statistically significant results if the unadjusted p-value is at most 9.09×10^{-4} . Out of the 55 dependent variables, we found statistically significant difference across the aggregated worker samples for 31 of them, and Table 2.2 provides a list of them. Results in Table 2.2 clearly suggests that for the population of crowd workers who were available on MTurk from August 1, 2016 to August 5, 2016, there are significant differences across

Question	Dependent variable	unadjusted p-value
Location	Percentage of worker from California	$< 2.2 \times 10^{-16}$
	Percentage of workers from Florida	2.40×10^{-5}
	Percentage of worker from the Pacific division	$< 2.2 \times 10^{-16}$
	Percentage of worker from the South Atlantic division	3.10×10^{-8}
	Percentage of worker from the West region	$< 2.2 \times 10^{-16}$
	Percentage of worker from the South region	1.22×10^{-5}
	Percentage of worker from the Northeast region	0.0005
	Percentage of worker from the Midwest region	0.0003
Age	Average age of workers	3.16×10^{-6}
	Percentage of workers under 30	4.99×10^{-11}
	Percentage of workers under 35	4.65×10^{-8}
Race	Percentage of white workers	8.47×10^{-7}
# of years using MTurk	Percentage of workers who use MTurk for fewer than 1 month	$< 2.2 \times 10^{-16}$
	Percentage of workers who use MTurk for fewer than 6 months	$< 2.2 \times 10^{-16}$
# of HITs completed (total)	Median number of HITs a worker completed	7.31×10^{-19}
	Percentage of workers who completed 1000+ HITs	$< 2.2 \times 10^{-16}$
	Percentage of workers who completed 5000+ HITs	1.96×10^{-10}
# of HITs completed (last month)	Percentage of workers who completed <25 HITs last month	3.29×10^{-8}
	Percentage of workers who completed <50 HITs last month	4.45×10^{-10}
Approval rate	Percentage of workers with 99% or higher approval rate	8.59×10^{-5}
# of forums used	Median number of forums a worker used	9.98×10^{-6}
	Percentage of workers who don't use forums	1.10×10^{-7}
# of workers communicated with (last month)	Percentage of workers who didn't communicate with anyone	0.0009
	Percentage of workers who communicated with 10+ people	0.0008
Household income	Percentage of workers with household income in [\$45K, \$115K]	0.0006
# of income earners in household	Percentage of households with 1 income earner	0.0003
	Percentage of households with 2 income earners	6.05×10^{-9}
	Percentage of households with 3 or more income earners	4.51×10^{-5}
# of children under 18	Percentage of workers who has no child under 18	2.86×10^{-5}
# of experiments participated (last month)	Percentage of workers who participated in <5 experiments	3.63×10^{-8}
	Percentage of workers who participated in 90+ experiments	1.24×10^{-7}

Table 2.2: A list of aspects of worker demographics for which statistically significant differences are identified across the eight aggregated worker samples collected at different times in a day.

available workers at different times, in terms of a few key dimensions of their demographic backgrounds, including location, age, race, working experience with MTurk, etc. On the other hand, we also find that workers at different times don't exhibit much variations in terms of their religion, personal income, whether MTurk is their primary source of income and whether they hold other jobs outside MTurk, as the unadjusted p-values of hypothesis tests for all dependent variables associated with these dimensions are larger than 0.05.

In addition to analyzing the aggregated, population-level temporal differences, we are also interested in understanding the differences among available workers at different times on the level of *individual days*, when the average size of each worker sample is of the same order of magnitude as the sample size for a typical experiment. Understanding such temporal

Question	Dependent variable	# of days with $p < 0.05$
Location	Percentage of worker from California	5
	Percentage of worker from the Pacific division	5
	Percentage of worker from the West U.S. region	5
# of years using MTurk	Percentage of workers who use MTurk for fewer than 1 month	5
	Percentage of workers who use MTurk for fewer than 6 months	5
# of HITs completed (total)	Median number of HITs a worker completed	5
	Percentage of workers who completed 1000+ HITs	5
	Percentage of workers who completed 5000+ HITs	5
# of HITs completed (last month)	Percentage of workers who completed <50 HITs last month	4
# of forums used	Median number of forums a worker used	5
	Percentage of workers who don't use forums	5
# of experiments participated (last month)	Percentage of workers who participated in <5 experiments	5
	Percentage of workers who participated in 90+ experiments	5

Table 2.3: A list of aspects of worker demographics for which statistically significant differences are identified across worker samples collected at different times on the level of individual days.

differences on the day-level is particularly relevant for experimenters, because it gives an experimenter a sense of whether his experiment (where 100–200 subjects are recruited for each treatment) can possibly approach to subpopulations with significantly different demographics from the entire worker pool if, hypothetically, he launches the same experiment at different times within a day. To address this question, for each of the 5 days in our experiment, we conducted the 55 hypothesis testings across the 8 worker samples that we got from that day. We claim the temporal difference in a dependent variable to be statistically significant on the day-level, if we obtain a p-value of 0.05 or smaller in the hypothesis tests on that variable for at least 4 out of 5 days.

Table 2.3 shows the set of dimensions that we have identified significant temporal differences on the day-level. As the table indicates, launching an experiment at different times in a day may result in subject samples that significantly differ from each other in terms of where the subjects come from, how experienced the subject is with MTurk and/or scientific experiments, and how they complete the work on MTurk (e.g., work more independently or tend to communicate with other workers on forums). In other words, these results imply that experimenters may need to use cautions when deciding the timing of their experiments on on-demand platforms like MTurk, especially if any of these differing dimensions of worker

demographics across different times can possibly influence the experimental results⁹.

2.1.3 Capturing the Key Characteristics for Workers at Different Times

In the previous subsection, we have showed that significant temporal variations *do* exist in the demographics for available workers on MTurk at different times, and we further identified a few differing dimensions. To accurately describe the features for workers at different times of day, we next move on to have a more detailed examination on what the key characteristics for workers in each time slot are.

Again, we start with extracting key demographic characteristics for workers at different times from a population point of view, using the aggregated worker samples. To get an intuitive idea, we first plot the worker compositions across all eight aggregated samples with respect to each of the 21 demographic survey questions, and Figure 2.1 shows a subset of the plots. For example, in Figure 2.1a, for each time slot, we present the percentage of workers in the aggregated sample of that slot who comes from the northeast, midwest, south or west regions of the U.S., and further compare such breakdown across all 8 time slots. From a visual inspection on these figures, it's easy to see that, for instance, workers who are available at 2am are featured by a large portion of west U.S. workers (about 40% of them are from west U.S.), while workers who are available at 8am are characterized by a rather small percentage of west U.S. workers (about 10% of them are from west U.S.). In addition, it's also visually apparent that the 2am workers are likely to be younger, have completed fewer HITs on MTurk, and live in a household with a single income earner, while the 8am workers

⁹For example, researchers found that when participants are not naive to experimental materials, the effect sizes observed in the experiments can be significantly reduced [Chandler et al., 2015]. As we find that the available MTurk workers at different times in a day have different levels of experience with scientific experiments, researchers may need to carefully consider whether such temporal difference would influence the effect size of their experiments.

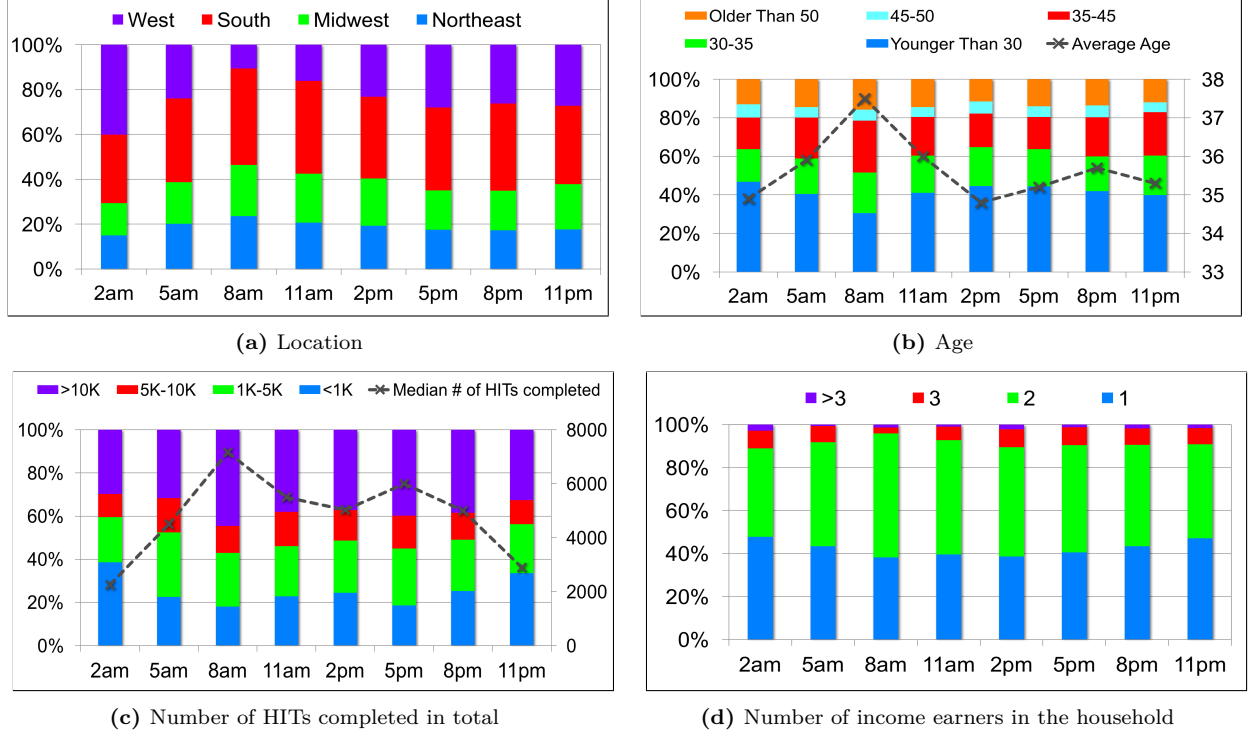


Figure 2.1: Temporal differences in the aggregated worker examples in terms of worker location, age, number of HITs completed in total and number of income earners in the household. For Figures 2.1b and 2.1c, the dashed gray lines are associated with the right y-axes.

tend to be older, have completed more HITs, and live in a household with 2 income earners.

To see whether these characteristics for workers in different time slots that we conclude from our visual inspections are statistically significant, we conduct further statistical tests. In particular, given a particular dependent variable, after using a statistical test (i.e., one-way ANOVA, one-way Kruskal-Wallis ANOVA, or proportion test, depending on the data type) to examine whether the values of this dependent variable are statistically the same across all aggregated worker samples as we have described in the previous subsection, we further conduct a post-hoc test to understand which pairs of samples have significantly different values on this variable from each other. Since we have 8 aggregated samples in total, for any particular sample, we can get a set of 7 comparison results, indicating whether the value of the dependent variable in the given sample is significantly different from those in each of the

other 7 samples. We define a dependent variable X to be the “key characteristic” for workers in time slot Y , if a significant difference has been detected (at the level of $p = 0.05$ with the Bonferroni-Holm correction) between the value of X for the worker sample in time slot Y and those for *more than half* of the other 7 worker samples (i.e., at least 4 out of the other 7 slots), and the direction of comparisons are *consistent* (e.g., it’s always the case that the value of X for workers at time Y is smaller than that for workers at other time slots in the pairwise comparisons that are statistically significant). Roughly speaking, identifying a key characteristic X for workers at time Y means that in terms of X , the available workers at time Y is consistently different from other available workers for at least half of the time in a day. For example, if we find the percentage of workers from west U.S. at 2am is significantly higher than that percentage for workers in at least 4 other time slots, we will label the 2am workers as “more likely to come from west U.S.”

Following this criteria, we summarize the key characteristics that we have identified for available workers in each of the 8 time slots in Table 2.4. We first notice that workers in the 2pm, 5pm and 8pm slots don’t have any key characteristic identified, implying that available workers in these time periods are “*average workers*” who are similar to workers in other time slots for at least half of the time in a day. On the other hand, we also find that workers in the 2am and 8am slots have many characteristics that distinguish them from the average workers, and these characteristics span a wide range of dimensions in worker demographics including worker location, age, race, working experience on MTurk, levels of communication related to MTurk work and household information. In addition, workers in the 5am, 11am, and 11pm time slots are observed to be different from the average workers on some particular dimensions as well—for example, there are significantly more workers who are inexperienced with scientific experiments at 5am, fewer workers from west U.S. at 11am, and more workers who have limited work experience on MTurk at 11pm.

Finally, to understand whether there is any consistently significant key characteristic for

Time slots	Key characteristics
2am	% of worker from California/Pacific division/West region is higher (7)
	% of worker from Florida/South Atlantic division is lower (6)
	% of worker who used MTurk for 1 month or fewer is higher (5)
	% of worker who used MTurk for 6 months or fewer is higher (6)
	Median number of HITs completed in total is smaller (6)
	% of worker who completed 1000+ HITs in total is lower (6)
	% of worker who completed 5000+ HITs in total is lower (5)
	% of worker who completed <25 HITs last month is higher (5)
	% of worker who completed <50 HITs last month is higher (5)
5am	% of worker who communicated with 10+ workers last month is lower (4)
	% of workers participated in <5 experiments last month is higher (5)
8am	% of worker from California is lower (6)
	% of worker from the Pacific division/West region is lower (7)
	% of worker from the Northeast region is higher (4)
	Average age is higher (5)
	% of worker under 30 is lower (7)
	% of worker under 35 is lower (6)
	% of white worker is higher (4)
	Median number of HITs completed in total is larger (5)
	% of worker who completed 1000+ HITs in total is higher (4)
11am	% of worker who don't use forums is lower (5)
	% of worker living in a household with 2 income earners is higher (5)
2pm	% of worker living in a household with 3 or more income earners is lower (5)
	% of worker from California is lower (6)
5pm	% of worker from Pacific division/West region is lower (6)
8pm	N/A
11pm	N/A
	% of worker who used MTurk for 1 month or fewer is higher (5)
	% of worker who used MTurk for 6 months or fewer is higher (6)
	Median number of HITs completed in total is smaller (5)
	% of workers who completed 1000+ HITs is lower (6)
	% of workers who completed 5000+ HITs is lower (4)

Table 2.4: A list of key characteristics for available workers at different times of day on the aggregated level. Numbers in the parentheses for each key characteristic represent the number of statistically significant differences detected among a total of 7 pairwise comparisons between the aggregated worker sample of the given time slot and that for another time slot.

workers at different times on the level of individual days, we repeat our analyses on the set of 8 worker samples for each of the 5 days. We denote a dependent variable X to be the key characteristic for workers in time slot Y on the day-level, if for at least 4 out of 5 days, we find a significant difference in the value of X between workers in time slot Y and workers in at least *two* other time slots (at the level of $p = 0.05$). Table 2.5 reports all such key characteristics. These results imply that for an experimenter, if he launches an experiment of typical size at 2am EST, he will almost certainly get a collection of subjects who are significantly more likely to be from west U.S. and with significantly lower levels of work

Time slots	Key characteristics
2am	% of worker from California is higher (4)
	% of worker from Pacific division/West region is higher (5)
	Median number of HITs completed in total is smaller (4)
	% of worker who completed 1000+ HITs in total is lower (4)
5am	N/A
8am	% of worker from the Pacific division is lower (5)
	% of worker from the West region is lower (4)
	Median number of HITs completed in total is larger (4)
11am	N/A
2pm	N/A
5pm	N/A
8pm	N/A
11pm	% of worker who used MTurk for 1 month or fewer is higher (4)

Table 2.5: A list of key characteristics for available workers at different times of day that are consistently observed on the level of individual days. Numbers in the parentheses for each key characteristic represent the number of days (out of 5) where a statistically significant difference is detected in at least 2 pairwise comparisons (out of 7) between the worker sample of the given time slot on one day and that for another time slot on the same day.

experience on MTurk compared to the case if he launches the experiment at other times, no matter which day in the week the experiment is launched. Similarly, an experimenter will get a sample of workers who are less likely coming from west U.S. and more experienced with MTurk if he decides to launch an experiment at 8am EST, and he will approach to a sub-population of inexperienced workers if the experiment is conducted at 11pm EST.

2.2 Scientific Studies with the Crowd: How Timing Influences Results

In the previous section, we have experimentally showed that there are temporal variations in worker demographics throughout a day for available workers on on-demand platforms like MTurk. For researchers who leverage on-demand platforms to conduct scientific studies, observing these temporal variations naturally leads to the question of whether the timing of a study will influence its results. More specifically, imagine a researcher who conducts an experiment on MTurk to understand certain aspects of human behavior—he decides to

launch his experiment at 8am EST in a day, collects enough amount of data according to his experimental plan and then draws conclusions from the data that he collects. But what if in a counterfactual world, the researcher actually launches his experiment at 5pm EST? Will he obtain the same conclusion in this 5pm experiment as what he gets from the 8am experiment? In this section, we aim to answer this question.

2.2.1 Experimental Design

Different from our experiment in the previous section, to examine whether the timing of a study (e.g., a survey or a behavior experiment) has any impact on study results, we have to conduct the *same* study at multiple different times within a day, so that we can have a direct comparison on the study results. Similar to what we have discussed in Section 2.1.1, the first possible experimental design—posting the same study at multiple different times without allowing each worker to participate for more than once—can not give us an accurate answer to our research question, because workers are naturally presented on MTurk at different time periods. More specifically, consider our previous example that a researcher launches an experiment at 8am, or counterfactually, at 5pm. If a worker i typically works on MTurk from 8am to 6pm, she should have been permitted to participate in the experiment no matter it is posted at 8am or 5pm. If we actually post the experiment on MTurk twice in a day at 8am and 5pm, respectively, and further forbidden a worker from participating in the experiment for more than once, we can possibly exclude significant amount of data from participants like worker i in experiments that are conducted later (i.e., the 5pm experiment), which may result in certain bias in the results of those experiments. On the other hand, if we post the same study at different times and allow each worker to take the study for multiple times, we may only be able to attenuate the data bias problem at best (as workers may find it boring to take the same study more than once), and it is also hard for us to determine whether the differences in experimental results, if any, are because of the experimental timing or

worker’s experience with the study (i.e., workers may behave differently in the study after participating in it many times).

Without finding a satisfying solution from our existing options of experimental designs, we return to the ideal approach for answering our research question, that is, to create “parallel universes” of the on-demand platform by randomly assigning each worker to one version of the MTurk website and posting the study of interest on each version of the website at a different time. In practice, this is very challenging as we don’t have the access to the *entire population* of MTurk workers and therefore there is no way for us to conduct a randomization on all the workers. However, we realize that although it’s impossible to conduct randomization on the entire population of MTurk workers, it’s feasible to do so on a *representative sample* of all MTurk workers. Inspired by this idea, we propose a two-phase experimental design as follows.

A two-phase experiment. We design our experiment as a two-phase experiment. In particular, in the first phase, we continuously post a recruiting HIT on MTurk for a *long* period of time. The purpose of the first phase is to get a *representative sample* of the entire MTurk worker population for the time period when we conduct our experiment. In this recruiting HIT, workers are told that we will run an experiment to study people’s cognitive skills and decision-making behavior in the near future, and they can sign up to that experiment by answering a set of simple demographic questions (i.e., all the 7 basic demographic questions that we used in the pre-task survey of the experiment in Section 2.1). Each worker is allowed to participate in the recruiting HIT *only once*.

All workers who have signed up in the first phase are then eligible for participating in the second phase which contains the actual experiment that we are interested in studying, and we will describe more detail about the experiment content later. Before the second phase starts, we randomly assign each signed-up worker to one of the 4 time slots (i.e., 2am EST, 8am EST, 2pm EST, 5pm EST). Importantly, a worker’s time slot assignment decides when

the worker can find out our second phase experiment HIT—we *only* open the experiment HIT to each worker during the time slot that the worker is assigned to¹⁰, and each worker can only take part in the experiment *once*. For example, if a worker is assigned to the 2am slot, she will only be able to find out our second phase experiment HIT on MTurk around 2am, but not 8am, 2pm or 5pm. We again set the expiration time for each HIT in the second phase experiment to be 1 hour, so subjects we recruit in a particular time slot actually arrive at the experiment within a time period of 1 hour (e.g., subjects in the 2am experiment accept the HIT between 2am and 3am). Furthermore, we do *not* use separate emails to communicate with workers about when our second phase experiment HITs are launched. In other words, a worker takes our second phase experiment HIT only if she *happens to be available* on the platform around the time period that she is assigned to. Such experimental design effectively allows us to build 4 “parallel universes” for the representative sample of the MTurk worker population that we collect in the first phase. Hence, comparing the results obtained from experiment HITs that are posted at different times in a day can help us to understand the potential impact of experimental timing on experimental results.

As we have found in Section 2.1, among the 4 selected time slots, the available workers at 2am EST and 8am EST have a number of distinctive demographic characteristics, while workers at 2pm EST and 5pm EST tend to be average workers. This diversity in worker demographics for the 4 selected time slots can potentially improve the chance for us to observe different experimental results at different times.

Experiment content. We include a wide range of different tasks in our second phase experiment HIT. These tasks cover a variety of studies that economists, psychologists, social scientists and computer scientists may be interested in conducting on MTurk. In particular,

¹⁰This is realized by creating a new qualification type on MTurk which corresponds to the group assignment of workers, and the second phase experiment HITs that are posted at different times is only visible to workers with a particular value (i.e., group number) for the newly created qualification.

we consider 3 tasks that examine people’s decision-making behavior in classical economic games, and comparing decisions in these games for available workers across different time slots can help us to understand that for experimental and behavior economists, whether they need to concern about the possible influences of experimental timing on the results of their crowd-based studies:

- *Dictator game*: In this task, a worker is told that she will be randomly paired with another participant in our experiment to play a game—one of them will be randomly selected as Player A (i.e., the “dictator”) and the other will be Player B (i.e., the recipient). If the worker is Player A, she will be given 1 dollar and need to decide how much of the money she is willing to give to Player B. If the worker is Player B, she will get whatever amount of money that the Player A in her pair decides to give to her. The dictator game is frequently studied in the area of experimental economics, and it is partly used to interpret people’s social preferences such as altruism and inequality aversion [[Kahneman et al., 1986](#)].
- *Public goods game*: In this task, a worker is told that she will be randomly matched with two other participants in our experiment to play a game—Each of them will first get 60 cents, and then the worker is asked to divide the money between a private account and a public account. The worker can keep all the money that she puts in her private account for herself. Meanwhile, for each cent that the worker decides to put in her public account, it will be multiplied by 1.5 and will become a public fund that is owned by all three participants in the group. At the end of the game, we will divide the public fund equally into three shares, so the worker can get one share of the public fund in addition to the money in her private account. The public goods game and its variants are often used in experimental economics to understand the cooperative behavior of people [[Fehr and Schmidt, 1999](#), [Fischbacher et al., 2001](#)].

- *Lottery choice*: In this task, a worker is presented with a series of 11 pairs of lottery choices. In each pair of lotteries, one lottery option is fixed as “earn \$0 with 50% of the chance and earn \$2 with 50% of the chance”, while the other option is “earn \$ x for sure” where $x \in \{0.4, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.6\}$. The worker is asked to pick one lottery that she prefers for each pair of lottery options. The order for presenting lottery pairs in the task is randomized. The lottery choice game has been adopted in a large number of studies to understand the risk preferences of individuals [Holt et al., 2002].

To elicit worker’s actual behavior when facing *real* economic decisions, we provide monetary rewards as bonus payments to workers in the above three tasks. Specifically, in both the dictator game and the public goods game, the amount of bonus payment a worker will get in each game is *actually* decided by the outcomes of the game. For example, if a worker is assigned as Player A in the dictator game and she decides to give Player B \$0.30, she will get the rest \$0.70 as her bonus in this game. For the lottery choice game, we will randomly choose one lottery that the worker prefers (among the total 11 lotteries that the worker has picked) and realize it. The amount of bonus payment the worker will earn from the lottery choice game is then decided by the realization of the chosen lottery. To make sure that each worker in our experiment understands the rules of the dictator game and the public goods game, we further add two qualification questions in each of these two games to test worker’s understandings on the game. A worker can only earn bonus payments in a game if she answers all qualification questions for that game correctly.

Besides the 3 economic games, we also include 3 cognitive tasks that we have introduced in Section 2.1—a social intelligence task (i.e., reading emotions from eyes; 20 questions in total), a nutrition intelligence task (i.e., answering a nutrition-related question after examining a pair of photographs of meals; 20 questions in total) and a thinking style task (i.e., deciding which two words go together best among three words; 30 questions in total)—in the HIT.

While the results of these cognitive tasks provide valuable insights for psychologists and social scientists to better understand the cognitive abilities or styles of individuals, some of these tasks also resemble a variety of data collection tasks that computer scientists typically conduct on platforms like MTurk, for either academic or commercial purposes (e.g., the social intelligence task can be thought of as a special image annotation/classification task). Therefore, including these 3 cognitive tasks as a part of our second phase experiment HIT enables us to provide some initial answers on whether psychologists and social scientists can draw similar conclusions from their cognitive studies if they conduct the same studies at different times in a day, as well as on how much the collected data quality varies by the task posting times for computer scientists who leverage the crowd for data collection.

In addition, we further add a personality survey in the HIT to get a sense of whether launching an experiment on MTurk at different times in a day implies any difference in the personality for subjects that are recruited. Such knowledge is of great value to scientific researchers as individual’s behavior is highly related to her personality [Ajzen, 2005, Snyder, 1983, Colvin, 1993]. Since we have found in Section 2.1 that available workers in different time slots come from quite different locations, and there is evidence suggesting distinctive personality profiles associated with various geographical regions of U.S. [Rentfrow et al., 2013], it’s reasonable to conjecture that we may observe a temporal difference in worker personality on MTurk through the personality survey. More specifically, personality is measured in this survey through the big-five inventory [John and Srivastava, 1999], which contains a set of 44 statements and a worker is asked to indicate how much she agrees with each statement on a 5-point scale, from 1 (disagree strongly) to 5 (agree strongly). The worker’s responses are then used to compute the scores for five factors of her personality, including extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience.

Finally, at the end of second phase experiment HIT, we again ask the worker a set of 14 extra demographic survey questions, which are the same as those questions that we have

used in the post-task exit survey of the experiment in Section 2.1.

Experimental data. For each worker who completes the recruiting HIT in our first phase, we record her responses to each of the 7 basic demographic questions (e.g., age, gender, location). For a worker who completes our second phase experiment HIT, we collect a wide range of data on worker’s economic behavior, cognitive abilities or styles and personality using worker’s responses in each task of the HIT.

First, for each of the three economic games in the HIT, we record all the decisions a worker makes in these games, including the amount of money she transfers to Player B if she is assigned as the Player A in the dictator game, the amount of money she decides to put in her public account, and her preferred lotteries for each of the 11 lottery pairs. In particular, for the lottery choice game, we denote the option of “earn $\$x$ for sure” as the “safe choice,” and we further use the total number of safe choices a worker selects in all 11 lottery pairs as a summary statistic for the worker’s behavior in the lottery choice game.

Regarding the 3 cognitive tasks, as we have access to the ground truth for both the social intelligence and the nutrition intelligence tasks, we use the number of questions that a worker answers correctly for each type of task as a measure of the worker’s cognitive ability in that task. Meanwhile, for the thinking style task, given a particular group of three words that are presented in one of the 30 questions, we denote some pairs of words in it as an “analytic combination” or a “holistic combination” according to [Ji et al., 2004]. For example, in a group of three words “seagull”, “sky” and “dog”, “seagull” and “dog” is an analytic combination because they both belong to the same abstract category (i.e., they are all animals), while “seagull” and “sky” is a holistic combination grouped together by their function—seagulls fly in the sky. We then count the number of word pairs picked by a worker in all 30 questions of the thinking style task that belong to analytical combinations (alternatively, holistic combinations), and use that value to represent the cognitive style of the

worker, that is, the degree to which the worker tends to reason in an analytical (alternatively, holistic) way.

As for the personality survey, we summarize a worker’s personality by computing the scores for each of the five key dimensions in personality using the worker’s responses in the survey, following the instructions in [John and Srivastava, 1999]. And finally, we also keep a copy of worker’s answers to all the 14 detailed demographic questions that they provide to us at the end of the HIT, regarding aspects like their working experience with MTurk and income/household information.

Experimental procedure. As we have discussed early, we design the experiment into two phases where workers recruited in the first phase are randomly assigned to one of the 4 time slots, which determines the time period when a worker will be able to find out the second phase experiment. We also do *not* conduct separate email communication with workers about the second phase experiment to ensure that workers who take part in the second phase experiment are the ones who are naturally available in the time slots that they are assigned to. These design decisions, yet, suggest that the retention rate for our second phase experiment can be inevitably low¹¹.

To get a sense of the retention rate, we first launched our experiment as a small-scale pilot from April 18, 2017 (Tuesday) to April 21, 2017 (Friday). More specifically, we conducted the first phase of our experiment during April 18–19, 2017. The recruiting HIT was posted for the first time at 12am EST, April 18, 2017, and it was re-posted every 2 minutes before being deleted at 11:59pm, April 19, 2017. Each worker got a fixed payment of 15 cents from completing the recruiting HIT, and in total, 2,508 workers signed up to our experiment in these two days. We then *uniformly randomly* assigned each of these 2,508 workers to one of

¹¹For example, it is possible that although a large number of workers are assigned to a particular time slot Y, very few workers actually show up in the experiment that is conducted at time Y because most workers are not available on the platform during that time.

the 4 time slots, and further conducted the second phase of our experiment within all these workers on April 20 and April 21 at 2am EST, 8am EST, 2pm EST and 5pm EST, as we have described earlier. Each worker got a fixed payment of \$1.5 by completing the second phase experiment HIT, and she might also earn extra bonuses depending on her decisions in the economic games in the HIT. During the period of April 20–21, we got experimental data from 21, 71, 64, and 59 workers for the 2am, 8am, 2pm, 5pm slots respectively, implying a daily retention rate of 1.65%–5.70% for each time slot. Not surprisingly, the retention rate for the 2am slot is the lowest, suggesting the number of available workers at that time is likely to be very low.

As our goal is to understand whether the experimental timing of a crowd-based study on MTurk can possibly influence the experimental results, for each time slot in our second phase experiment, we aim at collecting experimental data from a set of 100–200 workers at that time so that the sample size is similar to that for a treatment in a typical crowd-based experiment. The observed low retention rate in the pilot leads us to relaunch our first phase experiment again to get a larger representative sample of MTurk workers. In particular, we posted the recruiting HIT again between 12am EST, April 24, 2017 (Monday), and 11:59pm EST, April 28, 2017 (Friday), with the HIT being re-posted every 2 minutes. We increased the payment of the recruiting HIT from 15 cents to 20 cents at 5pm EST, April 26 (Wednesday) to accelerate the worker recruiting process. During this period, an additional 2,094 workers signed up to our experiment. That is, combining all the workers who completed our recruiting HIT during April 18–19 and April 24–28 together, in total, we got a representative sample of 4,602 unique MTurk workers.

We then conducted a *uniformly random* assignment of workers to time slots for the 2,094 workers that we recruited during April 24–28, and relaunched our second phase experiment HITs from May 1 (Monday) to May 5 (Friday), 2017. While the HIT was open to all 4,602 workers who were signed up to our experiment (at the time slot that corresponds to each

worker’s assignment), each worker was only allowed to take this HIT once. Thus, if a worker had completed the HIT during April 20–21, she could not take the HIT again during May 1–5. Combining all participants of the second phase experiment during April 20–21 and May 1–5 together, we eventually collected the experimental data from 87, 192, 231, and 215 workers for the 2am, 8am, 2pm, 5pm experiments, respectively. In other words, the size of recruited subjects sample for each time slot is roughly on the same magnitude of that for the sample size of a treatment in a typical crowd-based experiment.

All the HITs in our experiment was open to U.S. workers only. Moreover, as scientific researchers typically post their studies on MTurk on weekdays, to understand the realistic effect of experimental timing on experimental results, we further restricted ourselves to conduct experiments on weekdays (according to EST) only.

2.2.2 Revisiting Worker Demographics

First of all, since we have collected the demographic information for each worker in both the first and the second phase of our experiment, it is interesting to examine whether we can observe the same kind of significant differences in demographics for available workers at different times in a day, as we have reported in Section 2.1. Given the sample size in each time slot of our experiment is on the level of 100–200, which is similar to the average worker sample size for each sub-experiment of Section 2.1, we are essentially interested in checking whether the results in Table 2.3 (i.e., dimensions of worker demographics for which significant temporal differences exist on the level of individual days) still hold for a set of worker demographics data that is collected through a different experimental design.

The results of our check are reported in Table 2.6. As the table suggests, for many dimensions of worker demographics that we have observed significant temporal variations in Section 2.1, we again make a similar observation when examining the temporal differences in demographics for workers who participated in our second phase experiment in different

Question	Dependent variable	p-values
Location	Percentage of worker from California	0.0075**
	Percentage of worker from the Pacific division	0.0002***
	Percentage of worker from the West U.S. region	0.0013**
# of years using MTurk	Percentage of workers who use MTurk for fewer than 1 month	0.2571
	Percentage of workers who use MTurk for fewer than 6 months	0.1252
# of HITs completed (total)	Median number of HITs a worker completed	0.0239*
	Percentage of workers who completed 1000+ HITs	0.4203
	Percentage of workers who completed 5000+ HITs	0.2350
# of HITs completed (last month)	Percentage of workers who completed <50 HITs last month	0.8494
# of forums used	Median number of forums a worker used	0.0253*
	Percentage of workers who don't use forums	0.0697†
# of experiments participated (last month)	Percentage of workers who participated in <5 experiments	0.2527
	Percentage of workers who participated in 90+ experiments	0.0206*

Table 2.6: Statistical test results for examining whether worker demographics significantly vary over time using the data that we collect through the two-phase experiment, with †, *, **, and *** representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

time slots. These dimensions include worker location, the level of experience with MTurk (in terms of the number of HITs completed in total), the number of forums used that are related to MTurk work, as well as the level of experience in participating scientific experiments on MTurk.

On the other hand, in our check, we find that values for the two dependent variables that are related to the number of years a worker uses MTurk are not significantly different for available workers in different time slots, which seems to be inconsistent with our previous conclusion. However, as we conducted the experiment in Section 2.1 during August 2016 and conducted the two-phase experiment that we describe above mostly during May 2017, worker’s answers to the question of “the number of years using MTurk” also changed between these two experiments. For example, in May 2017, a worker who indicated to use MTurk for fewer than 1 month in August 2016 would have used MTurk for 8–9 months, and thus she would choose the option of “half to one year” for this question. In other words, the statistical tests on the percentage of workers who use MTurk for fewer than 1 month (or 6 months) across different times in the two experiments are not directly comparable. To see that in our two-phase experiment, whether the available workers at different time slots exhibit any

difference in terms of the number of years they use MTurk, we conduct a series of extra statistical tests on the other dependent variables that are coded from worker’s answer to this question (e.g., the percentage of workers who use MTurk for more than 2 years). Our test results show a significant (or marginally significant) difference in the percentage of workers who use MTurk for more than 2 years (or 3 years) across different time slots with $p = 0.0266$ (or $p = 0.0883$), which again supports the conclusion that the available workers at different times can vary a lot in terms of how long they have used MTurk. Similarly, regarding the demographic dimension on the number of HITs a worker completed in the last month, we also found that the percentage of workers who completed fewer than 250 HITs (or more than 750 HITs) last month is marginally different across different time slots with $p = 0.0596$ (or $p = 0.0566$).

In sum, our validity check largely confirms our observations on the variations of worker demographics across different times throughout a day, as reported in Table 2.3. As we are able to reach the same conclusions on the temporal differences using two sets of data that are collected from different experimental designs in different years, we believe that these detected temporal differences in worker demographics are very *robust*.

2.2.3 Influences on Studies Involving Incentivized Economic Decisions

Next, we move on to examine whether and how the timing of a crowd-based study may affect the experimental results in classical behavioral economic games. In particular, we consider the incentivized decisions workers make in these games, such as the amount of money a worker is willing to transfer to another worker in the dictator game, the amount of money a worker is willing to put in her public account in a public goods game, and the number of safe choices a worker makes in a series of lottery choice games. As no significant

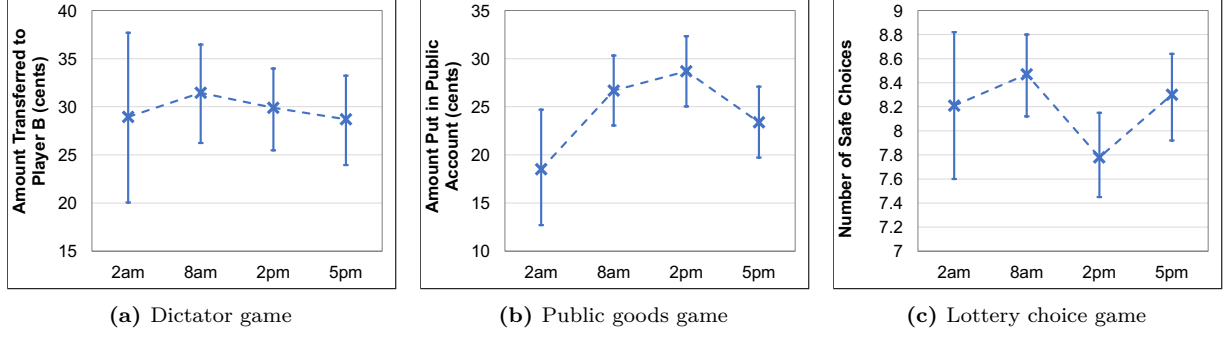


Figure 2.2: The incentivized decisions that workers in different time slots make for each of the 3 economic games. Mean values are reported and error bars represent bootstrapped 95% confidence intervals of the mean value.

difference is observed in worker’s incentivized decisions in different days, we aggregate the data for each of the 4 time slots across all days. Figure 2.2 compares the mean values of these economic decisions across the 4 time slots for each of the 3 economic games in our experiment HIT¹². Moreover, to obtain a range within which each of the mean values is likely to lie, we further attach a bootstrapped 95% confidence interval for each mean value in the figure. For example, given the set of data on the number of safe choices a worker at 2am EST made in her lottery choice game, we can estimate a 95% confidence level for its mean value as follows: We first obtain a *bootstrap resample* by sampling with replacement from the set of data while maintaining the same sample size as the original data, and compute the mean value for this bootstrap resample. This process is repeated for 1,000 times and thus we get a bootstrap distribution of the mean value. The 95% bootstrapped confidence interval of the mean value is then determined by the 2.5-percentile and the 97.5-percentile of the empirical distribution.

We make a few interesting observations by inspecting Figure 2.2 visually. On the one hand, it seems that workers in different time slots make similar decisions in terms of how much they are willing to transfer to another worker in the dictator game, as the confidence intervals for the mean values of transferred money across the 4 time slots largely overlap in

¹²For the dictator game and the public goods game, we consider only those data points if the worker answers the qualification questions correctly.

Figure 2.2a. On the other hand, the stories for the public goods game and lottery choice game are quite different—in both Figures 2.2b and 2.2c, we find that the confidence intervals for some pairs of time slots hardly overlap at all, suggesting that launching the experiment at different times in a day may actually lead to significantly different experimental results for these two games. For example, it seems that on average, workers at 2am put less money in their public accounts compared to workers at 2pm in a public good game, while workers at 8am are more risk averse (i.e., make more safe choices) than workers at 2pm.

To confirm our visual intuition, for each of the 3 economic games, we conduct an one-way ANOVA test for the values of the corresponding incentivized decision in different time slots to examine whether the distributions of these values are statistically the same at different times. The test results are reported in Table 2.7. Consistent with our previous intuition, we find that workers at different times don’t exhibit significant differences on their decisions in dictator games, yet they do behave significantly differently in the public goods game and the lottery choice game. This means that if an experimental or behavioral economist is interested in understanding how collaborative people tend to be through conducting a public goods game on MTurk, or interested in obtaining insights on the risk preferences of individuals by running a sequence of lottery choice games on MTurk, he may end up with different conclusions if he launched the experiment at different times in a day. More specifically, the post-hoc pairwise comparisons suggest that conducting a public goods game at 2am EST will lead the experimenter to believe people to be significantly less cooperative compared to that in the case if, hypothetically, he conducted the game at 2pm EST ($p = 0.0248$). Similarly, if an experimenter decides to run the lottery choice games at 8am EST, he would conclude the crowd to be significantly more risk averse than that in the scenario when he ran the experiment at 2pm EST ($p = 0.0353$).

To further understand whether the significant behavioral differences that we observe for economic games conducted in different time slots are simply a result of the variations in

Economic decision	p-values
Dictator game: amount transferred to Player B	0.8906
Public goods game: amount put in public account	0.0248*
Lottery choice game: number of safe choices	0.0427*

Table 2.7: p-values of the one-way ANOVA tests on economic decisions that workers at different times make, with * representing significance level of 0.05.

worker demographics over time, we fit each worker’s decision in the public goods game or the lottery choice game into linear regression models while controlling for her demographic information, and results are reported in Table 2.8. For both games, we consider two linear regression models—one using only the basic demographics as covariates (i.e., Model 1 for both games, columns 2 and 4 in Table 2.8), while the other controlling for both the basic and a few more detailed demographic information (i.e., Model 2 for both games, columns 3 and 5 in Table 2.8). For simplicity, we do not consider interaction terms in all the models.

More specifically, in the regression models of the public goods game, the dependent variable is the amount of money a worker puts in her public account, and we set workers in the 2am slot as our reference. Results in Table 2.8 show that workers in the 2pm slot puts significantly more money in their public accounts compared to workers in the 2am slot, even after the worker demographics is controlled. This implies that the influences of experimental timing on the results of the public goods game can not be fully explained by the differences in worker demographics across different times—in fact, according to our regression results, none of the worker demographics is actually observed to be significantly correlated with the worker’s decision in the public goods game.

As for the lottery choice game, we use the number of safe choices a worker selects as the dependent variable and workers in the 2pm slot are set as the reference. Again, we find the significant temporal differences in worker’s decisions in the lottery choice game is robust to demographic controls—workers who are available on MTurk at 8am and 5pm are still significantly more risk averse than available workers at 2pm when demographic information

	Public goods (Model 1)	Public goods (Model 2)	Lottery choice (Model 1)	Lottery choice (Model 2)
Intercept	19.82^{***} (5.32)	21.24^{***} (5.54)	7.12^{***} (0.45)	7.06^{***} (0.46)
2am			0.32 (0.33)	0.23 (0.33)
8am	6.18 (3.81)	6.03 (3.84)	0.61[*] (0.26)	0.57[*] (0.26)
2pm	8.39[*] (3.73)	8.41[*] (3.76)		
5pm	3.78 (3.71)	3.60 (3.73)	0.51[*] (0.25)	0.45[†] (0.25)
Female	1.16 (2.18)	1.28 (2.24)	-0.04 (0.20)	0.01 (0.21)
Age	-0.03 (0.09)	-0.04 (0.10)	0.02^{**} (0.01)	0.02^{**} (0.01)
Northeast	2.38 (3.33)	2.21 (3.35)	0.55[†] (0.31)	0.56[†] (0.31)
South	-1.65 (2.91)	-1.89 (2.95)	0.07 (0.27)	0.02 (0.27)
West	-1.12 (3.30)	-1.71 (3.36)	0.29 (0.31)	0.38 (0.31)
Bachelor	-0.79 (2.11)	-0.79 (2.18)	-0.27 (0.20)	-0.06 (0.20)
Hispanic/Latino	0.90 (4.63)	1.15 (4.65)	-0.73[†] (0.44)	-0.79[†] (0.43)
White	2.38 (2.83)	1.92 (2.91)	-0.24 (0.26)	-0.18 (0.26)
Christian	-1.34 (2.16)	-1.17 (2.18)	-0.07 (0.20)	-0.05 (0.20)
HITs (total)		0.00 (0.00)		0.00 (0.00)
Personal income > 37.5K		-0.13 (2.80)		0.10 (0.25)
Household income > 70K		-1.94 (2.63)		-0.71^{**} (0.35)
Half income from MTurk		-2.82 (2.67)		0.67^{**} (0.25)
Completed > 90 experiments (last month)		2.88 (2.44)		-0.36 (0.24)

Table 2.8: Linear regressions for the decisions workers made in the public good game and the lottery choice game. Coefficients and standard errors are reported. The statistical significance of the estimated coefficient is marked as a superscript, with [†], ^{*}, ^{**}, and ^{***} representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

is used as covariates in the models. Interestingly, we also note a few significant correlations between an individual’s risk attitude and her demographics. For example, workers who are older, live in Northeast U.S., and rely on MTurk for at least half of their income tend to be more risk averse, while Hispanic or Latino workers and workers whose household income is higher than \$70,000 are more risk-seeking. To see whether the temporal differences of worker behavior in the lottery choice game can be partly attributed to these significantly correlated demographics, we then compare the worker compositions across different time slots

on these dimensions of demographics. We find that compared to the 2pm workers, workers who are available at 8am and 5pm are indeed older, more MTurk-dependent, less likely to be Hispanic/Latino or come from households with an income of \$70,000 or higher, though the differences are not statistically significant. For the demographic compositions in terms of worker location, it is observed that the percentage of Northeast U.S. workers indeed differs significantly across the 4 time slots ($p = 0.0007$). However, this is mostly due to a very low fraction of Northeast workers in the 2am slot, and for the other three time slots (i.e., 8am, 2pm, 5pm) where significant differences in worker’s risk preferences are observed, there is no statistically significant difference in terms of whether workers in these time slots come from Northeast U.S.

To summarize, through our analyses on the incentivized decisions that workers at different times make in a number of classical behavioral economic games, we find that the timing of a crowd-based behavioral economic experiment *may* change the economic behavior that workers display in the experiment. Importantly, the change in the experimental results is not just due to the variations of worker demographics across different times, although these variations may also play a role.

2.2.4 Influences on Studies Examining Cognitive Abilities and Styles

Our previous analyses have examined whether different experimental timing for crowd-based behavioral economics studies can lead to different results, and we give a positive answer to that question. Now, we look into whether similar results can be found for a variety of cognitive experiments that psychologists or social scientists may be interested in conducting with the crowd to better understand the cognitive ability and style of people. We consider worker’s performance in each of the 3 cognitive tasks in our HIT, that is, the number of

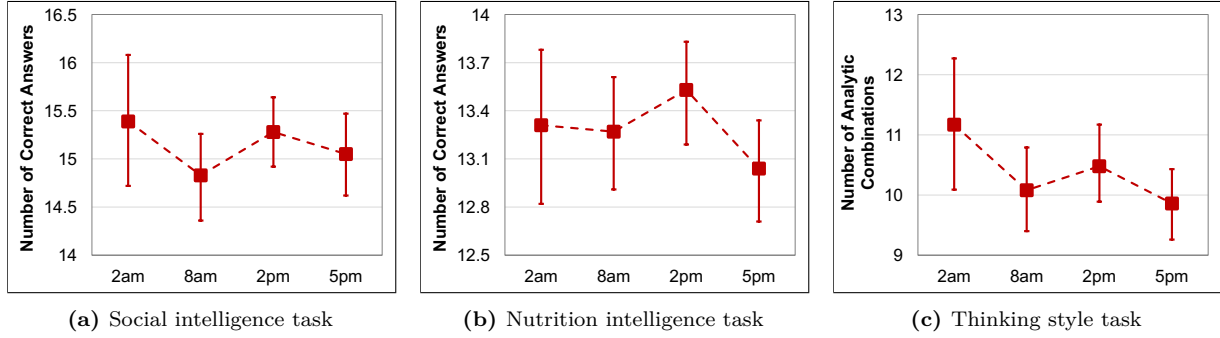


Figure 2.3: Worker performance in different time slots for each of the 3 cognitive tasks. Mean values are reported and error bars represent bootstrapped 95% confidence intervals of the mean value.

Worker performance metrics	p-values
Social intelligence: number of correct answers	0.3846
Nutrition intelligence: number of correct answers	0.1996
Thinking style: number of analytic combinations	0.1604

Table 2.9: p-values of the one-way ANOVA on worker performance in cognitive tasks at different times in a day.

questions in the social intelligence task that a worker answers correctly, the number of questions in the nutrition intelligence task that a worker answers correctly, and the number of analytical combinations of words that a worker selects in the thinking style task. Again, we don't find significant differences in worker performance across different days, which allows us to combine the data in different days together for further analyses. Figures 2.3a, 2.3b and 2.3c show the average worker performance with the bootstrapped 95% confidence interval across all 4 time slots for the social intelligence, nutrition intelligence and thinking style tasks, respectively.

As we can see in the figures, there is no obvious difference in worker performance across different times in a day for all the 3 cognitive experiments that we conduct. We further report the one-way ANOVA test results on worker performance for each of the 3 cognitive tasks in Table 2.9—worker performance in none of the three cognitive tasks is significantly different across time, implying that for psychologists and social scientists, the specific timing that

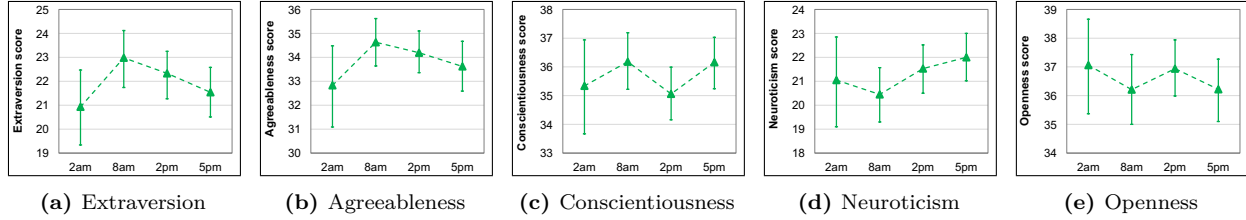


Figure 2.4: Worker personality in different time slots for each of the 5 factors. Mean values are reported and error bars represent bootstrapped 95% confidence intervals of the mean value.

they decide to launch cognitive experiments on MTurk has limited influence on the cognitive abilities or styles that they will be able to observe from the subjects (at least for cognitive abilities and styles that are studied in our experiments). Furthermore, given that some of these cognitive tasks are quite similar to many data collection tasks that computer scientists often post on MTurk (e.g., a computer vision researcher may ask MTurk workers to help him annotate human emotions in a set of pictures, which is similar to the social intelligence task in our HIT), our results also indicate that computer scientists may not need to worry too much about the fluctuation of the crowdsourced data quality over time. In other words, there is no such a time period, at least among the 4 time slots that we have examined, that if a data collection task is posted during that period, the crowd will return the requester with a batch of data of significantly higher (or lower) quality.

2.2.5 Examining Differences in Worker Personality

Finally, we examine whether launching the experiment at different times in a day will result in samples of subjects who display significant differences in their personality. Figure 2.4 compares the worker personality across 2am, 8am, 2pm and 5pm for each of the five major factors in personality, i.e., extraversion, agreeableness, conscientiousness, neuroticism and openness, and Table 2.10 reports the one-way ANOVA test results on each factor. It is observed that the available workers at different time periods don't have significant differences

Worker personality	p-values
Extraversion	0.1356
Agreeableness	0.2102
Conscientiousness	0.2528
Neuroticism	0.2774
Openness	0.6653

Table 2.10: p-values of the one-way ANOVA on personality for available workers at different times in a day.

in their worker personality on any of the five factors. In other words, experimenters don't need to worry too much about approaching to pools of workers with different personality if they launch their studies at different times in a day.

2.3 Discussion

In this chapter, we adopt two innovative experimental designs to understand the temporal dynamics of the crowd. In particular, we have showed that available workers on on-demand platforms like Amazon Mechanical Turk at different times in a day may exhibit *significant differences* in terms of their demographic backgrounds, such as their location and experience levels with MTurk. In addition, for researchers who conduct scientific studies with the crowd, we also find that it is possible that the specific timing that they decide to launch their studies on the platform may *change* the results that they will be able to obtain from the crowd.

These findings have very important implications for scientific researchers. First of all, given that experimental results can be influenced by the timing of the experiments, it is necessary for researchers to carefully record and report their experimental procedure, especially in terms of the experimental timings, in the communication of their scientific discoveries to improve the replicability of the findings. Besides, it's also worthwhile for researchers to consider conducting their crowd-based studies multiple times at different times in a day to better understand the robustness of their results. Finally, as the varying worker demographics over

time may also partly contribute to the differences in experimental results, researchers should collect the demographic information from all subjects of their studies and control for the demographics in their analyses whenever possible.

There are many interesting future directions to extend this work. For example, while we have answered in this chapter *whether* experimental timing can affect experimental results for crowd-based studies, we do not have a comprehensive understanding on *why* or *why not* yet. For example, as we have showed in Section 2.2.3, worker’s incentivized decisions in the public goods game are significantly different across different time slots, but it seems that such differences are not resulted from the temporal variations in worker demographics or worker personality. So what makes workers behave differently? Could it be worker’s prior knowledge about other workers in that time slot? In addition, why significant temporal differences are observed for worker behavior in the public goods game and the lottery choice game, but not the dictator game? What kind of experiments display higher levels of “robustness” against experimental timing? These are all interesting research questions that deserve in-depth research.

We have examined the temporal differences in worker’s demographics, economic behavior, cognitive abilities and styles, and personality in this chapter. There is another important question that researchers may care about regarding to launching experiments on on-demand platforms at different times. That is, if researchers conduct *randomized experiments* (i.e., randomly assign subjects into control and treatment groups), whether the results of these experiments can be influenced by the experimental timing. We conjecture that it is possible. For example, as we have identified the demographic composition of the worker population changes over time, if the treatment effect in a randomized experiment is highly correlated to one significantly temporally-varying worker demographic, researchers are likely to obtain different effect size when they launch the experiment at different times. Further research is needed to verify whether the results of randomized experiments can indeed be influenced by

experimental timing and to further understand why if the answer is yes.

Moreover, in this chapter, we focus on micro-level temporal dynamics of the on-demand workers in terms of their variations across *different times within a day*. As previous studies have noted, the crowd of on-demand workers also experiences evolvement over time and thus exhibits a significant temporal dynamics on the macro-level of months, years, or decades. Understanding the interplay between the macro-level and micro-level of temporal dynamics is thus an interesting future topic. In particular, given the rapid growth of the on-demand economy in recent years as well as in a foreseeable future, it is important for us to keep track of the demographics and behavior of the crowd in a long term. These data will then allow us to conduct various longitudinal studies on the crowd, including revisiting the topics that we have discussed in this chapter from time to time. With these studies, we can both get a more accurate and updated knowledge about the crowd of the moment, and possibly understand the development of on-demand economy in a broader context of the economic, political and cultural movement of the entire society.

2.4 Acknowledgements

The work in this chapter was produced in collaboration with Yiling Chen, Emma Heikensten, and Anna Dreber. We are grateful to Krzysztof Z. Gajos for providing the data for several cognitive tasks in the experiments.

Chapter 3

The Communication Network Within the Crowd

The traditional black-box view of the on-demand economy has not only made it difficult for us to understand who the on-demand workers are, but also how these workers perform the work. Unlike employees in traditional companies or organizations who are likely to share the same working space with each other, the crowd of on-demand workers can possibly be dispersed all over the world. It is, therefore, not uncommon for people to view the “crowd” as a group of *independent* workers, who do not, and do not need to, talk to or work with one another. For the requesters of the on-demand labor, this perception of on-demand workers being independent has been further strengthened by the digital communication protocol between them and the workers. For example, on a typical on-demand crowdsourcing platform like Amazon Mechanical Turk, the platform’s API hides from requesters personal attributes of workers (e.g., demographics), as well as *social* characteristics of workers, such as how many friends they have who also do crowdwork or if they are currently working on a task with other workers. Without this information, it is not surprising that requesters come to view the crowd of on-demand workers as independent from one another, with little attention paid

to the connections between them.

This notion of crowds as independent workers was recently dispelled by [Gray et al. \[2016\]](#), who opened up the black box and showed that workers are *not* independent but rather connected through social ties. Through a mix of ethnographic fieldwork, in-person interviews, surveys, and large scale data analyses of four different crowdsourcing platforms, they showed that workers collaborate with one another to meet social and technical needs left wanting by the crowdsourcing platforms studied. More specifically, they showed that workers collaborate on three fronts: 1) helping each other get through the administrative overhead involved in doing crowdwork (e.g., signing up for an account and getting paid, which can be especially challenging outside of the United States), 2) sharing information about lucrative tasks and reputable (or irreputable) requesters, and 3) completing work together. Thus, [Gray et al. \[2016\]](#) showed the crowd is not a collection of independent workers, but that there exist edges between the workers.

While prior work showed that communication exists, it left open the problem of understanding the scale, structure and impact of this communication. How widespread is the communication? What is the topology of the communication network? And how does participation in this communication network relate to the lives of crowd workers?

In this chapter, we set out to thoroughly understand the connectivity between the crowd of on-demand workers and answer all the above questions by mapping the entire communication network of workers on a leading crowdsourcing platform, Amazon Mechanical Turk (MTurk). We aim to understand the network’s properties and the implications that communication across the network has on all parties in the on-demand economy. To do so, we designed a task that encouraged workers to self-report their connections to other workers in a privacy-preserving way. The task was designed to provide value back to workers by allowing them to explore the network and learn about the workers they connect to as well as the greater network of crowd workers. The edges that workers provide are self-reported and thus not

perfectly accurate. However, they give us a close approximation of the true communication network underlying MTurk, and a sense of how widespread communication among workers is.

We analyze the structural features of the MTurk communication network. While a large segment of the population does, in fact, appear to be made up of isolated nodes, we show that there is a rich network topology over the subset of workers who report connections. That is, there is a substantial network *within* the crowd.

We show that online forums dedicated to working on MTurk play a key role in allowing workers to communicate across the network. Forums create overlapping subcommunities among workers. Forums differ from each other in terms of the topological structure of their subcommunities, the temporal nature of communication, and the content of discussions. Meanwhile, one-on-one channels are also used by some workers to communicate, yet they play a different role in fostering communication when compared to online forums. We also observe various types of homophily between workers. That is, we observe that workers are more likely to communicate with other workers who live in the same country, have worked on MTurk for a similar amount of time, and prefer the same types of MTurk tasks (e.g., classification or scientific experiments).

By correlating topological features of the network with a number of worker properties, we find that workers' positions in the network are related to various aspects of their MTurk experiences, such as how long they have stayed on MTurk, whether they make use of online forums, how successful they are as MTurk workers, and how fast they can find interesting tasks on MTurk. And as a final case study of how workers with different properties participate in the network differently, we provide a comparison between workers who live in and out of the United States and show that these two populations hold different positions in the network, adopt different channels for communication, and focus on different topics in their communication.

3.1 Related Work

The results of [Gray et al. \[2016\]](#) are based on data gathered by a team of ethnographers who spent roughly 19 months in India interviewing over 100 crowd workers, conducting repeat interviews with many of them over time to understand the longitudinal effects of crowdwork. [Gray et al. \[2016\]](#) augmented their interviews with large scale surveys of the crowd worker population in both the U.S. and India and an analysis of a HIT designed to understand where MTurk workers are located and what resources they use to find HITs. Their key finding is that workers collaborate with each other, often to make up for technical or social shortcomings in the platform. The notion that some workers talk and collaborate with one another is also supported by the 35 interviews of Indian workers that [Gupta et al. \[2014\]](#) conducted, mostly via Skype. Both studies indicate that workers collaborate to share tasks, aid each other in doing tasks, and provide social interaction that is often missing in online labor. This notion inspired our goal of mapping the worker network. Our contribution above and beyond these studies is to scale up their findings and dig deeper into the structure of communication. While they find communication between 35 to over 100 interview subjects, we measure and analyze the communication network of over 10,000 MTurk workers.

One theme that appears prominently in our findings is the importance of online forums to the structure of the communication network. Prior research has shown the importance of these forums in the work and lives of MTurk workers. [Martin et al. \[2014\]](#) spent hundreds of hours reading posts on TurkerNation, a popular online forum for MTurk workers, to understand this online community. They showed that workers primarily work on MTurk to augment their pay and that workers spend a lot of time talking about requesters and tasks in search of requesters with good reputations and tasks with high pay. Similarly, [Zyskowski and Milland \[2015\]](#) conducted an ethnographic study of TurkerNation. They observed participants in chat rooms and interviewed them. They state that on TurkerNation, “common topics of

discussion include the best jobs of the day, how to build one’s reputation, how to earn more money, and how to make working more fun.” Thus workers are using forums not just to find lucrative tasks but also to provide each other with social support. These qualitative studies inform our work. Our goal is to scale these studies up and see how big the communication network between MTurk workers is, what topology it has, and how workers use it.

Researchers have built at least two platforms that facilitate worker communication. First, TurkOpticon is a system developed by [Irani and Silberman \[2013\]](#) used by workers to rate requesters in terms of their communicativity, fairness, generosity, and promptness. Second, Dynamo [\[Salehi et al., 2015\]](#) is a community platform designed to aid MTurk workers with collective action problems such as “reining in problematic academic research practices” and gathering support for a letter-writing campaign. These works facilitate worker communication for focused goals. The purpose of our work is different in that we seek to understand the structure and scale of the overall communication network that has organically grown among the workers themselves.

3.2 Experimental Design

Amazon Mechanical Turk is an on-demand crowdsourcing platform in which requesters can post small tasks (i.e., HITs) with pre-specified payments for workers to complete, while workers can browse available tasks and choose HITs to work on. Once a worker has submitted her work for a given HIT, the HIT’s requester can review this work, accepting it if it is high quality and rejecting it if not. If work is rejected, the worker receives no payment. The rejection is also reflected in the worker’s *approval rate*, which is simply the fraction of HITs the worker has done that have been accepted. The approval rate serves as part of a *de facto* reputation system, and requesters often make HITs available only to workers with a high approval rate. Amazon additionally designates some workers as *Masters*. While Amazon

Forum	URL	Registered Users	Posts	Start Date
Reddit HWTF	https://www.reddit.com/r/HITsWorthTurkingFor/	32,297	unknown	February, 2012
MTurkGrind	http://www.mturkgrind.com/	6,743	748,983	October, 2013
TurkerNation	http://turkernation.com/	15,411	311,816	August, 2011
MTurkForum	http://www.mturkforum.com/	53,883 (932 active)	1,354,249	January, 2009
CloudMeBaby	http://www.cloudmebaby.com/	4,180	32,072	July, 2012
Facebook (groups)	http://facebook.com/	unknown	unknown	unknown

Table 3.1: Statistics as of October 4, 2015 on the six online MTurk forums listed as options for the question on forums.

does not disclose the criteria used to grant the Masters qualification, it is viewed as a sign of high quality, and requesters may choose to make HITs available only to workers who have received this qualification.

Amazon does not provide a platform for workers to interact with each other. However, MTurk workers have created a variety of forums focused on navigating MTurk. A brief overview of the most popular forums is given in Table 3.1. These forums differ somewhat in functionality. Reddit’s HITsWorthTurkingFor (HWTF) is a highly active subreddit primarily used by workers to share links to, and information on, good HITs. MTurkGrind, TurkerNation, and MTurkForum are post-driven discussion boards organized around a range of themes, much like USENET newsgroups. They each offer moderated areas and distinct but comparable conversation modes organized by discussion threads. Each of these forums has tens of thousands of threads dedicated to a wide range of topics. Registered members of these forums can participate in the discussion in any threads they are interested in, and they may also interact with each other in chat rooms or through private messaging systems provided by the forums. CloudMeBaby is a site devoted to helping navigate and improve cloud based workplaces including MTurk. In addition to these public forums, there are a number of both private and public MTurk-related Facebook groups, varying in size from tens of users to thousands of users.

Since the communication network among workers is not accessible from the API provided by MTurk—in fact, the network exists outside of and separate from the MTurk platform—we cannot simply download, crawl, or scrape this network. In this section, we describe a HIT

that we designed to give workers incentive to self-report their connections with each other in the communication network underlying MTurk. We believe the approach described here is novel and could be of independent interest.

3.2.1 The Network Mapping HIT

We designed a five-step HIT to gather information from workers and allow them to self-report other workers with whom they communicate. In the first step of our HIT, each worker was asked to create a unique nickname for herself. This nickname had several intended purposes. First, it was used as a unique identifier for the worker, preserving the worker's privacy since it was not based on the worker's MTurk ID or other identifying information. (Workers were encouraged not to use their real name, though we had no way to enforce this.) Additionally, it was used as a way for other workers with whom this worker communicates to add edges to this worker and identify this worker in the network. This is described in more detail below.

In the second step, workers were asked nine survey questions about their demographics and MTurk usage:

- **Location:** Which country do you currently live in?
- **Age:** Which year were you born in?
- **Gender:** What is your gender?
- **Education:** What is the highest degree or level of school you have completed?
- **Master:** Are you a Mechanical Turk Master?
- **Approval Rate:** What's your approval rate on Mechanical Turk?
- **Experience:** How long have you been Turking?
- **Tasks:** What types of MTurk tasks do you typically do?
- **Forums:** What online MTurk forums do you regularly use?

For the question on tasks, we provided a list of eleven common types of MTurk tasks such as data entry, survey, and scientific experiments, and allowed workers to choose one or more. For the question on forums, we enumerated the six popular MTurk forums in Table 3.1. Workers could choose any number of these forums, specify other forums they use, or say that they do not use any forums.

We allowed workers to set privacy preferences individually for each of the nine questions. For each question, a worker could choose whether to share her answer with all other workers who completed our HIT, to share her answer with only those workers connected to her in the communication network, or to keep her answer private.

In the third step, workers were asked to answer two free-form questions related to their experience on MTurk:

- Why did you start Turking?
- What motivates you to keep Turking?

These questions were carefully chosen to obtain information that other workers would find valuable and interesting as a way of providing value back to workers who completed our HIT. We ran a pilot survey in which we asked workers what they would most like to know about other workers in the MTurk community and extracted the most popular questions. Our hope was that presenting information that workers found valuable would encourage workers to explore the communication network through the visualization and to truthfully report their connections. Workers were informed that their answers to these two questions would be shared with all other workers who completed our HIT as part of the network visualization which they would view in Step 5.

In Step 4, each worker was asked to pause and take a moment to exchange nicknames with other MTurk workers that she knows. Workers were told that they could do this in any way they wanted and given several examples including exchanging nicknames in person, over the phone, through instant messaging, or through text messaging.

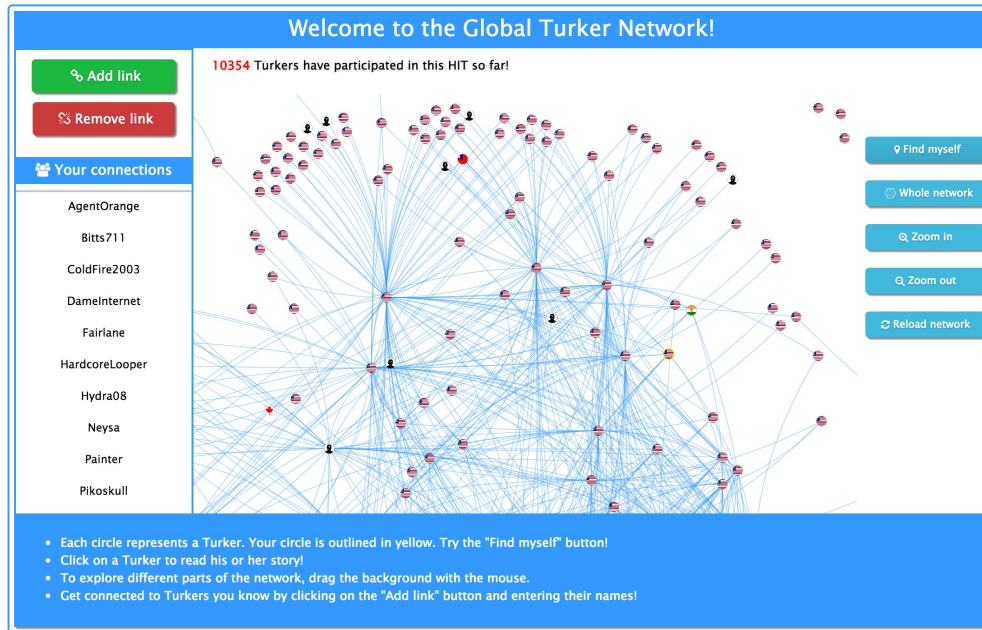


Figure 3.1: Screenshot of the user interface for the visualization of the MTurk communication network.

In the final step, workers were shown a visualization of the current state of the communication network (Figure 3.1). Each worker in the network was represented by a node displaying the national flag of her country (if her privacy settings allowed), and edges were shown between pairs of connected workers. The worker was able to locate herself, zoom in and out of the network, and click on any worker in the network to view his information. When a worker clicked on another worker to whom she was connected, she would see his nickname as well as all information that he had chosen to share with his connections. Crucially, when a worker clicked on a worker to whom she was *not* connected, she would *not* see his nickname and would see only information he chose to share with all workers. Thus such workers were effectively anonymous.

At this point, workers could add an edge to any other worker by providing his nickname. When adding an edge, the worker was asked an additional two questions:

- How do you usually talk to this worker?
- What do you usually talk about with this worker?

For the first question, the worker was provided with a list of communication channels such as forums, phone calls, email, and instant messaging, and allowed to choose one or more. For the second, the worker was given a list of topics such as sharing HITs, discussing requesters, sharing Turing tools/scripts, and chatting about day-to-day life, and could choose one or more. After entering this information, an undirected edge between the two workers was immediately added to the network. Workers were also able to remove edges to other workers.

Before submitting the HIT, the worker was given a unique URL that would allow her to return to the visualization to add or remove additional edges and continue to explore.

Note that by design, an edge between two workers could only be added if one of the workers knew the other’s nickname, which could only occur if the workers had communicated¹. Thus we believe that the vast majority of the edges in the network represent a true exchange of information, or in other words, a communication between workers. Of course there are likely pairs of workers who communicate but did not choose to exchange nicknames. However, exchanging nicknames allowed workers to learn interesting information about each other and better understand their own place in the MTurk community. We believe this design nudged workers towards reporting many of their true connections, though the true communication network is perhaps even more dense and vast than we show here.

We cannot rule out the possibility that the very existence of our HIT caused communication between pairs of workers who had not previously communicated with each other. This is unavoidable; in general, every new HIT has the potential to provoke new communication and the communication network is always evolving. We attempted to minimize this effect by intentionally deciding not to pay workers per edge added, as this would result in workers adding edges to those they do not regularly communicate with.

¹A worker could potentially guess another worker’s nickname, but we do not believe this frequently occurred. If it did, the second worker could remove the unwanted edge.

3.2.2 Experimental Procedure

We posted our HIT to MTurk. Workers who accepted the HIT read through a description of the task, signed a consent form stating that they were voluntarily participating in our experiment, and then completed the HIT as described above. The payment for the HIT was fixed at \$1 USD and the average completion time was roughly 10 minutes. The HIT was open to all workers on MTurk. Each worker was allowed to complete the HIT only once, but could return to their personalized URL to further explore the network and add or delete edges as often as they liked.

To ensure our HIT was well-functioning and scalable, we intentionally launched our experiment in phases during August and September of 2015. We first launched two small batches on August 11 (60 HITs) and August 12 (200 HITs). We notified workers on TurkerNation ahead of time about these two test launches. Next, to test the scalability we launched two larger batches on August 17 (596 HITs) and August 20–21 (1594 HITs). Satisfied with these initial tests, we finally left our HIT up for 2 weeks straight from August 28 to September 11, with the exception of 2 days (September 3–4) during which our requester account accidentally ran out of money due to the unexpected popularity of our HIT. After our HIT was taken down, workers continued to update the network via their private URLs. We report on data collected on September 13 once the addition and removal of edges had greatly slowed.

3.3 Results

A total of 10,354 workers completed our HIT. [Stewart et al. \[2015\]](#) estimated that when conducting behavioral research on MTurk, one laboratory is sampling from a pool of roughly 7,300 workers, and that the seven laboratories they studied sampled from an overall pool of roughly 11,800 workers. This suggests that our HIT was approximately a census of the active workers at the time. Of the workers who did our HIT, 1,389 (13.4%) either added

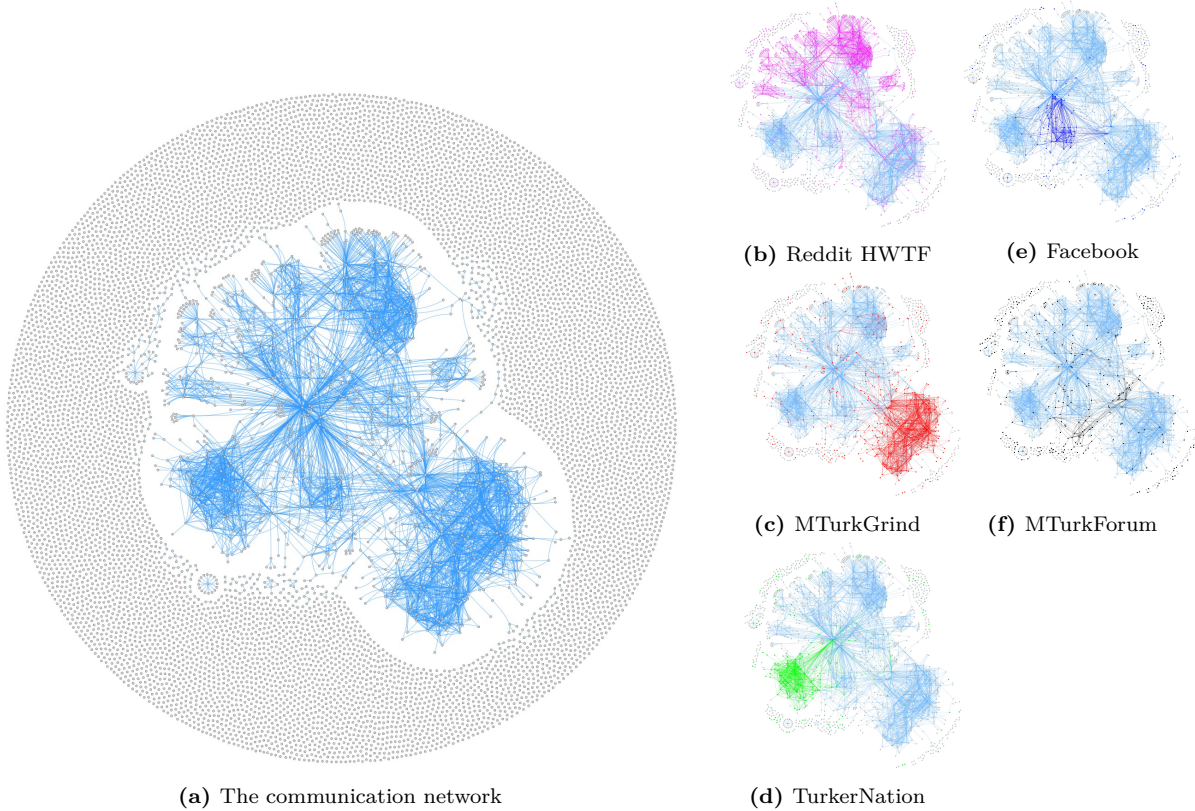


Figure 3.2: 3.2a: The communication network among Amazon Mechanical Turk workers. 3.2b-3.2f: Subnetworks for Reddit HWTF (magenta; 660 workers, 1837 edges), MTurkGrind (red; 392 workers, 1331 edges), TurkerNation (green; 200 workers, 740 edges), Facebook (blue; 133 workers, 357 edges), and MTurkForum (black; 312 workers, 244 edges).

at least one edge to another worker or had an edge added to them by another worker. We refer to these workers as *connected*. Among connected workers, a total of 5,268 edges were added, resulting in a mean degree of 7.6, median degree of 2, and maximum degree of 321. The largest connected component of the communication network consisted of 994 workers (71.6% of all connected workers), while the next largest consisted of just 49 workers (3.5% of all connected workers). Of the remaining connected components 117 were made up of a single edge between a pair of workers.

The communication network is shown in Figure 3.2a. Examining the network visually, it appears that the largest connected component is made up of several densely connected

clusters of workers. Below we show that this structure largely coincides with workers’ use of different online forums dedicated to Mechanical Turk work.

3.3.1 A Network Enabled by Forums

Forum use is extremely widespread among workers who completed our HIT, with 59.1% of all workers and 83.0% of connected workers reporting that they use at least one forum². The overwhelming majority of edges involved communication through a forum as 89.9% of the edges added were between pairs of workers that communicate via forums, and 86.2% between pairs that communicate *exclusively* through forums. Since the vast majority of communication between workers occurs on online forums, we next analyze the structure of the subnetworks defined by each of the forums.

We extract the subnetwork corresponding to each forum by keeping only *connected* workers who use that forum and only edges between pairs of these workers. As a sanity check, pairs of workers reported communicating with each other through forums for the vast majority of these edges (93% averaged over all subnetworks). Figures 3.2b-3.2f illustrate the subnetworks for Reddit HWTF, MTurkGrind, TurkerNation, Facebook, and MTurkForum, respectively. We omit CloudMeBaby as only 0.9% of all workers reported using it. As is visually apparent from the figures, users of different forums make up distinct but overlapping subcommunities, which explains much of the structure in the network.

To quantify our visual intuition, we measure whether or not workers who use the same forum are more likely to connect to each other than to other workers. The sociological phenomenon that contact between similar people occurs at a higher rate than among dissimilar

²The Forums question was added to Step 2 of our HIT on August 20 when we first realized the prevalence of forum usage. We asked the 856 workers who completed our mapping HIT before August 20 which forums they regularly use in a separate, one-question follow-up HIT and 659 responded. As a result, 98.1% of all workers answered the question. Whenever we report statistics related to forum usage, we restrict attention to workers who answered the Forums question.

Forum Name	ECGR	ACGR	p-value	User Fraction (q)	R	H
Reddit HWTF	0.50	0.30	<0.001	0.48	0.69	0.70
MTurkGrind	0.41	0.23	<0.001	0.28	0.53	0.65
TurkerNation	0.25	0.18	0.005	0.14	0.56	0.62
Facebook	0.18	0.17	0.362	0.10	0.53	0.45
MTurkForum	0.36	0.25	<0.001	0.23	0.39	0.28

Table 3.2: Left section: Expected cross-group ratio and actual cross-group ratio for the usage of each forum. Right section: One-sided homophily measures for each forum.

people is called *homophily* [McPherson et al., 2001]. Thus, we are interested in understanding the extent to which homophily exists with respect to forum use.

One standard approach to quantifying homophily is the *homophily test* described in [Easley and Kleinberg, 2010]. Consider a binary property \mathcal{C} that a node may or may not satisfy. In our case, satisfying \mathcal{C} might mean using a particular forum like MTurkGrind. Let q denote the fraction of the population who satisfy \mathcal{C} , S denote the set of all nodes that satisfy \mathcal{C} , and T denote the set of all nodes that do not. If there is no homophily with respect to \mathcal{C} , edges would be equally likely to form between all pairs of nodes in the network independent of whether those nodes satisfy \mathcal{C} . So, in the case of no homophily each node on an edge would satisfy \mathcal{C} independently with probability q , and the probability that any edge would be between one node in S and one node in T would be $2q(1 - q)$. We refer to this quantity as the *expected cross-group ratio (ECGR)* of \mathcal{C} . If, on the other hand, nodes in S were more likely to connect to other nodes in S , and nodes in T to other nodes in T , then the actual fraction of edges that would be between nodes in S and T , or the *actual cross-group ratio (ACGR)*, would be significantly lower. The homophily test compares these ratios.

Table 3.2 (left section) reports the results of homophily tests run separately for each forum³, limited only to connected workers. For each of the five forums, we find that the actual cross-group ratio is lower than the expected cross-group ratio. This provides evidence

³Note that while the test of Easley and Kleinberg [2010] easily extends beyond binary properties, we must run it separately for each forum since workers may select multiple forums.

for homophily with respect to the use of each forum, confirming the visual intuition given by Figure 3.2. To check whether the differences are statistically significant, we keep the network structure fixed and simulate a random assignment of node property values (that is, whether or not a node uses a particular forum) by assigning each node to use the forum with probability equal to the fraction q of users who use the forum in the real worker population. We repeat this process 1,000 times, calculating the cross-group ratio for each of the 1,000 resulting networks. An empirical p-value can then be computed as the fraction of these simulated networks with a cross-group ratio smaller than the ACGR we measure. As reported in Table 3.2 (left section), the differences are significant for almost all forums.

The results of the homophily tests may, in fact, underestimate the amount of homophily in the network. This is because while we might expect workers who use Facebook forums, for example, to be more likely to connect with other workers who use Facebook forums, it is unclear if workers who do not use Facebook forums are much more likely to connect with other workers who do not. To address this, we look at two alternative measures of such “one-sided” homophily. For a given node i , let n_i be the total number of edges incident on i , and $n_{i,S}$ be the number of edges incident on i that connect to nodes in S . Intuitively, it is a sign of homophily if, on average, the fraction of the edges that are incident on some node in S that to connect to other nodes in S is higher than the fraction of nodes in the total population that are in S , i.e., if $R \equiv (1/|S|) \sum_{i \in S} (n_{i,S}/n_i) > q$. The measure R treats all nodes equally. The homophily index of Currarini et al. [2009], defined as $H \equiv \sum_{i \in S} n_{i,S} / \sum_{i \in S} n_i$, is similar but effectively gives more weight to nodes with higher degree. Again, if $H > q$, there is evidence of homophily.

Table 3.2 (right section) shows both R and H for each forum along with the fraction of workers who reported using that forum, again limited to connected workers. As expected, these measures show a clear and striking tendency for workers to connect to other workers who use the same forums.

Given that workers are more likely to communicate with others from the same forums, information should flow easily within subcommunities. One may wonder how information spreads *between* subcommunities. Are there “connectors” in the network who bridge subcommunities [Burt, 2004, 2007]? In fact, 32.4% of connected workers reported using more than one forum regularly, providing ample opportunities for information to flow from one forum to another through these individuals. Furthermore, among all edges connecting a pair of workers that both reported using forums, 71.8% are between pairs in which at least one worker uses a forum that the other does not. This provides another route for information to spread between subcommunities. This observation supports the theoretical prediction of Kleinberg et al. [2008] that if there are informational benefits to bridging communities, many people will take a position in the network to earn, share, and ultimately dilute these benefits.

3.3.2 Differences Between Subcommunities

We next highlight three major differences across these subnetworks in terms of topological structure, temporal communication patterns, and content of communication, and then discuss implications. As before, we extract the subnetwork corresponding to a forum by taking all connected workers who use the forum and all edges between these workers.

Topological Differences

We first examine how tightly connected each subcommunity is using two metrics: *density* and *transitivity*. Given a network with n nodes and m edges, the density of the network is defined as $d \equiv \frac{2m}{n(n-1)}$, which is the ratio between the actual number of edges in the network and the maximum number of edges that could exist in any network with n nodes [Wasserman and Faust, 1994]. Transitivity measures the degree to which triangles in the network are closed. Let $n_{triangle}$ be the number of triangles in a network (i.e., sets of three nodes with

Forum Name	Density (d)	Transitivity (t)	Diameter	Avg. Shortest Distance
Reddit HWTF	0.008	0.30	9	8.36
MTurkGrind	0.017	0.38	13	7.15
TurkerNation	0.037	0.48	5	4.55
Facebook	0.041	0.38	6	4.37
MTurkForum	0.005	0.11	10	7.85

Table 3.3: Density, transitivity and distance metrics for each subcommunity.

edges between each pair) and n_{triple} be the number of connected triples (i.e., nodes x , y , and z with an edge between x and y and another between y and z ; a set of three nodes can create up to three triples). The network’s transitivity is then $t \equiv 3n_{triangle}/n_{triple}$, which measures the ratio between the actual number of triangles and the maximum number of triangles that could occur in any network with n_{triple} triples.

Intuitively, higher density and higher transitivity both imply a more densely connected network. Table 3.3 reports the density and transitivity for each of the five subcommunities. The degree of connectivity varies a lot between subcommunities, with TurkerNation and Facebook being the most tightly connected and MTurkForum the least tightly connected.

To further understand how densely connected the subcommunities are, we measure the *diameter* and the *average shortest distance* between two nodes for the largest connected component of each subcommunity. With the exception of MTurkForum, the largest connected component contains the majority of nodes in the subnetwork for each forum. Table 3.3 summarizes these results. TurkerNation and Facebook have the smallest diameter and average shortest distance respectively, suggesting that workers in the largest connected components of these two subcommunities are closer to each other. This echoes our previous observation that the TurkerNation and Facebook subcommunities are more highly interconnected. Despite the largest connected component in the MTurkForum subnetwork containing only 35.3% of workers who use the forum (110 workers), the diameter is still large, further evidence that the MTurkForum community is not tightly connected.

Individual subcommunities are not uniformly dense, but composed of a mixture of tight-

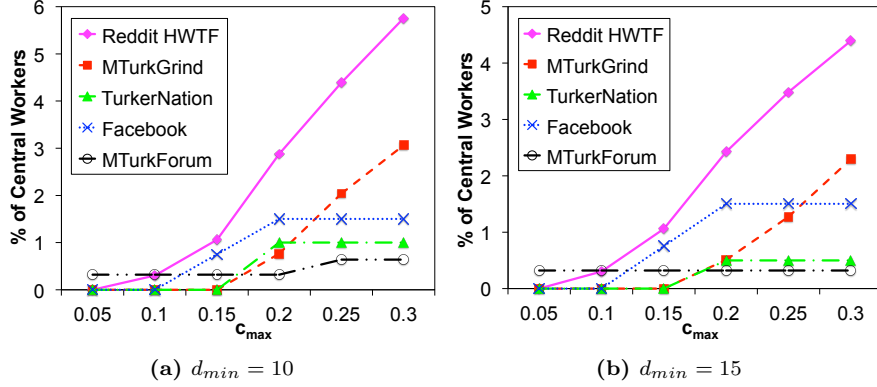


Figure 3.3: The percentage of central workers of star structures in each subcommunity.

knit groups and “star structures,” consistent with the core-periphery structure of social networks [Borgatti and Everett, 1999]. Within tight-knit groups, most workers communicate with each other, forming cliques in the extreme. The sizes of the largest cliques in Reddit HWTF, MTurkGrind, TurkerNation, Facebook, and MTurkForum are 11, 16, 16, 12, and 6, respectively, and these largest cliques account for 1.67%, 4.08%, 8.00%, 9.02% and 1.92% of all workers in each subcommunity. In contrast, *star structures* occur when a large number of workers connect to a common central worker but not much to each other. To identify star structures, we formally define a “central worker” to be any node with degree at least some value d_{min} and clustering coefficient⁴ at most some value c_{max} . We use the number of central workers identified in a network as a proxy for the number of star structures in it. Figure 3.3 shows the fraction of workers who are central workers in each subcommunity when we vary d_{min} and c_{max} . By this measure, there exist many more star structures in the Reddit HWTF subcommunity than in any others, a phenomenon that can be observed by a visual inspection of Figures 3.2b – 3.2f. This suggests that workers may be using Reddit HWTF in a different way than the other forums. We provide more evidence of this below.

⁴The clustering coefficient of a node is $c \equiv 2 \times |\{e_{j,k} : e_{j,k} \in E, j, k \in N\}| / (d(d-1))$, where d is the node’s degree, N is the set of the node’s neighbors, E is the set of edges among nodes in N , and $e_{j,k}$ is the edge connecting nodes j and k [Watts and Strogatz, 1998]. This is the ratio between the number of edges between the node’s neighbors and the maximum number of edges between d nodes.

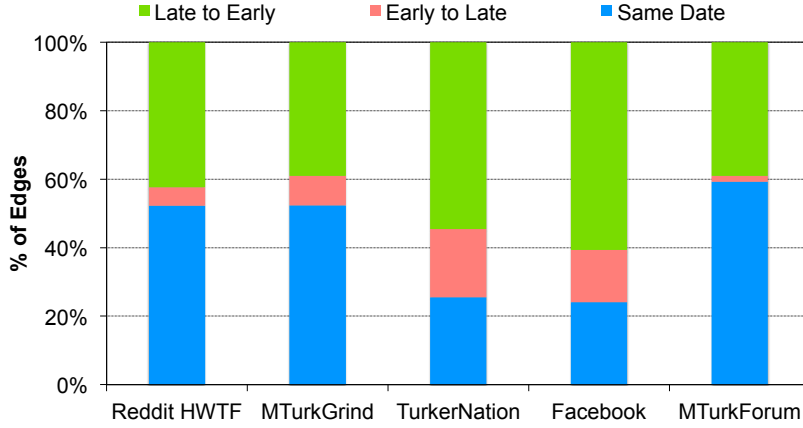


Figure 3.4: Temporal communication in each subcommunity.

Temporal Communication Differences

HIT completion timestamps can be used to understand the temporal nature of communication in each subcommunity. For this analysis, we coarsely divide the edges in the network into three categories: edges between workers who completed the HIT on the same day, edges added by a worker to another worker who completed the HIT on an earlier day, and edges added by a worker to another worker who completed the HIT on a later day. Note that the third type of edge can only occur when a worker returns to the network visualization another day via their private URL.

Figure 3.4 shows the fraction of edges that are of each type for each of the subcommunities. More than half of the edges in the Reddit HWTF, MTurkGrind, and MTurkForum subcommunities are between workers who took the HIT on the same day. On the contrary, workers who use TurkerNation and Facebook are much more likely to communicate with other workers who took the HIT on different days. Strikingly, at least 15%–20% of the edges in the TurkerNation and Facebook subcommunities were created by workers who had submitted the HIT on a previous day, but returned to the network to add additional edges later.

To further understand the temporal nature of communication, we calculate two additional quantities for each subcommunity: the empirical probability of a worker in the subcommunity

Forum Name	Same Day	Different Day
Reddit HWTF	0.049	0.005
MTurkGrind	0.077	0.010
TurkerNation	0.081	0.032
Facebook	0.074	0.035
MTurkForum	0.030	0.002

Table 3.4: Mean probability of connecting to a worker who took the HIT on the same day or a different day.

adding an edge to another worker in the subcommunity conditioned on that worker arriving the same day, and the empirical probability of a worker adding an edge to another worker conditioned on that worker arriving a different day. Specifically, for each worker in a subcommunity we calculate the fraction of all workers who arrived the same day with whom the worker shares an edge and the fraction of all workers who arrived on different days with whom the worker shares an edge, and we average these empirical probabilities across workers. The results, given in Table 3.4, show that an average worker who uses Reddit HWTF or MTurkForum is an order of magnitude more likely to connect to a worker who accepted the HIT on the same day as opposed to a different day. This effect is dramatically smaller for workers using TurkerNation or Facebook.

Taken together, these results suggest that workers might use Reddit HWTF and MTurkForum to broadcast or obtain information that is immediately actionable, communicating primarily with other workers who happen to be online at the same time. This is in contrast with workers on TurkerNation and Facebook, perhaps indicating that workers on the latter forums form longer lasting relationships.

Communication Content Differences

We turn to a comparison of the topics discussed in different subcommunities. Figure 3.5 shows the fraction of connected pairs that report communicating about each of five topics: HITs, requesters, Turking scripts and tools, day-to-day-life, and other things. Consistent with

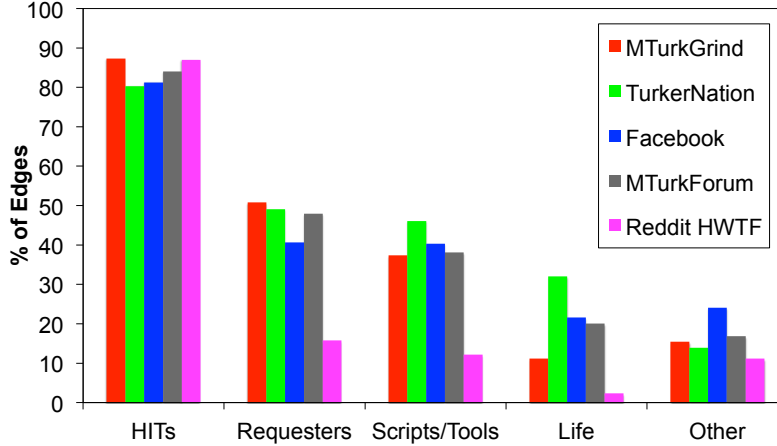


Figure 3.5: Topics discussed in each subcommunity.

the previous literature on forum usage [Gray et al., 2016, Gupta et al., 2014, Zyskowski and Milland, 2015], we find intensive discussion about HITs in all subcommunities. Workers in Reddit HWTF almost exclusively discuss HITs. Workers in other subcommunities are more likely to share information about requesters, provide technical support, and recreate the social environment otherwise missing from online work. TurkerNation has the most communication on day-to-day life and Facebook has the most communication on other topics, suggesting that workers use these forums in a more social manner.

Comparing the Subcommunities

Next we put all the differences we have observed together to help us understand how these subcommunities are similar and how they are different. On the one hand, TurkerNation and Facebook might be more socially oriented than other forums, leading to more tightly connected subcommunities, workers who felt the urge to add edges to other workers they know even if those workers took the HIT on a different day, and more discussions not directly related to MTurk work. In comparison, Reddit HWTF, MTurkGrind, and MTurkForum appear to be mostly dedicated to discussions about details of MTurk work. Reddit HWTF in particular displays a variety of features (e.g., prevalence of star structures and discussions almost

exclusively about HITs) which suggest that workers treat it as a platform for broadcasting good HITs above all else. MTurkGrind appears to be something in between a social community and a broadcasting platform, which may be related to the fact that 51.3% of all connected workers who use MTurkGrind also reported using Reddit HWTF. One might conjecture either that MTurkGrind has developed into an independent, more socialized community partly from a pool of Reddit HWTF users, or that MTurkGrind has started to attract users from Reddit HWTF who seek more social interactions. Finally, as we discuss in Section 3.3.6, MTurkForum accounts for a significant amount of the communication that occurs between workers outside of the United States. This might explain why it seems less connected than other subcommunities.

3.3.3 The Role of One-on-One Communication

While the majority of communication occurs over forums, workers also report communicating one-on-one via in-person discussions, phone calls, emails, text messages, instant messages, video chatting, and other channels. Overall 13.8% of connected pairs communicate at least partially through one-on-one channels, and 10.1% communicate exclusively through one-on-one channels. Among those pairs that communicate at least partially one-on-one, the three most popular communication channels are instant messaging (27.3%), in-person discussion (18.0%), and email (15.8%).

The role of one-on-one communication is different from that of communication via forums. While forum use is responsible for enabling much of the communication within the largest connected component, one-on-one communication is much more common in the smaller components. Inside the largest component, only 10.7% of connected pairs communicate at least partially through one-on-one channels, and 7.29% exclusively so. Outside of this component, the story is very different: 74.0% of pairs communicate at least partially through one-on-one channels and 63.6% exclusively so. Thus, one-on-one communication accounts for

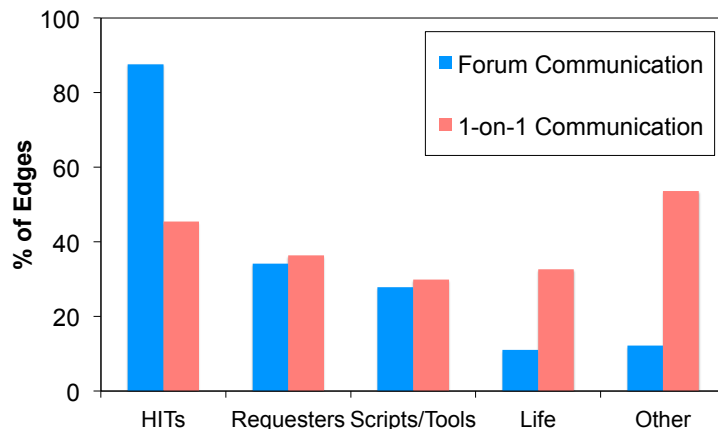


Figure 3.6: Comparison of topics discussed through forums vs. one-on-one communication.

the majority of edges outside of the largest component.

In addition, the distribution of topics discussed by pairs of workers who communicate one-on-one differs substantially from that of workers who communicate over forums. Figure 3.6 compares the amount of discussion for each topic (i.e., the percentage of pairs that communicate on the topic) among pairs who communicate one-on-one vs. in forums. Workers primarily use forums to discuss HITs, while workers who communicate one-on-one communicate much less about HITs and more about day-to-day life and other topics.

3.3.4 Homophily in the Network

We have seen that there is a communication network within the crowd and that workers communicate across the network both via forums and one-on-one channels. It is natural to ask who it is that workers are most likely to communicate with. In Section 3.3.1, we showed that there is homophily in the network in terms of forum usage. We now examine whether there is homophily in the network with respect to other worker characteristics.

To answer this question, we follow the same approach used in Section 3.3.1. First, we apply (generalized, non-binary) homophily tests to examine and compare cross-group ratios. Next, we compare the one-sided homophily measures R and H with the fraction q of workers

who share the same property among all connected workers. Using this approach, we do not see strong, consistent evidence for homophily along characteristics such as worker age, gender, education, approval rate, or if a worker is an MTurk Master.

We did, however, find that there is homophily in the network for two other worker characteristics: location and length of time on MTurk. For a worker’s location (limited to just U.S. and Indian workers, $ECGR = 0.249$, $ACGR = 0.107$, $p < 0.001$), it is observed that U.S. workers are much more likely to connect to other U.S. workers ($q = 0.857$, $R = 0.906$, $H = 0.943$), and the tendency for Indian workers to connect with other Indian workers is even more substantial ($q = 0.130$, $R = 0.781$, $H = 0.580$). For the length of time on MTurk ($ECGR = 0.844$, $ACGR = 0.809$, $p < 0.001$), the values for both one-sided homophily measures are also larger than the fraction of workers for almost all groups (“less than 1 year”, “1-2 years”, “2-3 years”, “more than 4 years”) and close for the remaining “3-4 years” group ($q = 0.0914$, $R = 0.1694 > q$, yet $H = 0.0907$ is just slightly less than q). This implies that experienced workers are likely to connect to experienced workers while inexperienced workers tend to communicate with inexperienced workers.

Finally, we analyze homophily around the types of tasks workers regularly do. We could not conduct a single unified homophily test for task type since the vast majority of workers regularly work on more than one type of task. We also did not conduct homophily tests on the binary property of whether or not a worker does a particular type of task (as we did with forum usage) because two workers who do not have a particular task type in common may still be very likely to connect to each other because of their shared interest on one or more other types of task. This would make interpreting the ACGR difficult. Figure 3.7 shows that both one-sided homophily measures are larger than the corresponding fraction of workers who do that type of task for almost all task types, with only the exception of transcription. This indicates that workers tend to communicate with others who work on similar tasks.

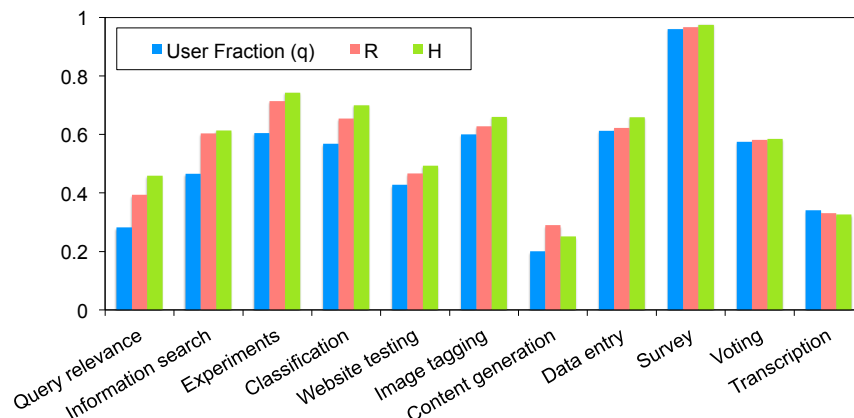


Figure 3.7: Comparisons of worker fraction and one-sided homophily measures for each type of task.

3.3.5 Correlates of Network Position

Next we report our findings on the relationship between network position and various worker properties such as length of time on MTurk, success on MTurk, and access to information. Note that the relationships we report are correlations only. It is impossible to determine whether there is a causal relationship between network position and worker properties from our data.

First, we examine whether workers’ positions in the network have any relationship with how long they have been on MTurk. According to Table 3.5, the percentage of workers that have been on MTurk for more than 1 year is higher among connected workers than unconnected workers. Consistent with our understanding that the network within the crowd is largely conducted over forums, Table 3.5 shows that connected workers are also more likely to use forums than unconnected workers.

Next, we attempt to understand whether workers’ network positions relate to how successful they are. While “success” on MTurk is hard to measure, we can use as a proxy a worker’s approval rate and whether or not the worker has been granted Masters status. These capture how successful a worker has been at getting her own work approved. As Table 3.5 suggests, by both of these measures, connected workers are more successful than

Property	Connected	Unconnected
Be active >1 year	54.9%	45.9%
Use forums	83.0%	55.5%
Have Master status	11.4%	6.9%
Mean approval rate	98.6%	97.4%

Table 3.5: Relationship between whether a worker is connected and various worker properties.

unconnected workers as they are more likely to be MTurk Masters and have higher approval rate on average. At first glance these two effects may seem small, but a 1% increase in approval rate or a Masters qualification allows a worker access to many more HITs which could dramatically affect her income. Thus these are very important outcomes for workers.

Finally, we investigate the connection between workers’ network positions and how fast they learn about HITs. We analyze how the network characteristics of workers who accepted our own network mapping HIT changed over time. Specifically, we sort all workers according to the time that they took our HIT and bin them into groups of 200. Figure 3.8 shows the percentage of connected workers in each bin⁵. There is a clear decreasing trend over time: connected workers were likely to learn about our HIT earlier than unconnected workers. Figure 3.9 shows a box plot of the degrees of connected workers who took our HIT on different days. Since our data was collected two days after we took down the HIT when few new edges were being added, we believe we gave workers ample time to connect to those workers who took our HIT late, reducing the chance that the low degrees of these workers are an artifact of our data. Here we see that workers who found our HIT earlier also seem to have larger degrees. If this phenomenon generalizes across HITs, this dynamic might result in connected workers starving out isolated workers from high paying tasks.

These results suggest that there are potential benefits to crowd workers associated with

⁵As mentioned in Section 3.2.2, we notified TurkerNation workers about our test batches of HITs on August 11–12 before launch. Hence we exclude workers who took the HIT on these 2 days in this analysis to minimize possible bias.

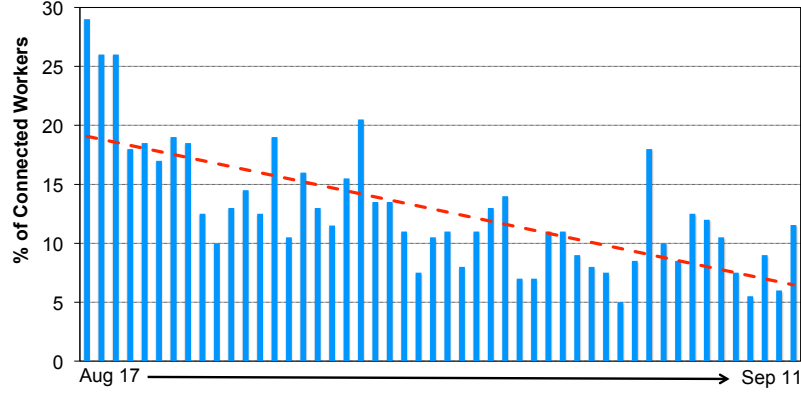


Figure 3.8: The percentage of connected workers in each bin of 200 workers, ordered by time. The red dashed line is a linear regression trend line.

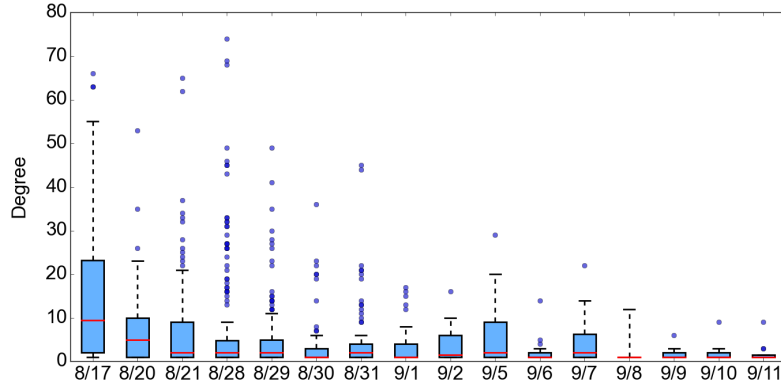


Figure 3.9: Degrees for connected workers who did our HIT by day. We omit four outliers with degree over 100.

their positions in the network. Being connected is correlated with longevity on the site, higher probability of getting work accepted, and the ability to learn about HITs faster than unconnected workers.

3.3.6 U.S. vs. International Workers

Finally, we study the differences between workers who are located inside and outside of the United States. Overall, 9,108 workers (88.0%) reported being located in the U.S., while the remaining 1,246 workers (12.0%) reported being located in other countries. While international workers are more likely to be connected than U.S. workers (13.1% U.S. vs.

16.0% international), connected U.S. workers have higher degree on average than connected international workers (8.19 vs. 3.96). This coincides with the finding that a higher percentage of U.S. workers (59.9%) reported using forums than international workers (53.2%), and this comparison is even sharper when we restrict to connected workers (85.8% U.S. vs. 66.5% international).

U.S. workers rely heavily on forums to communicate with each other (91.1% of connected pairs of U.S. workers communicate with each other on forums, and 88.1% exclusively so). International workers tend to use one-on-one channels dramatically more often (76.7% of connected pairs of international workers communicate through one-on-one channels, and 56.9% exclusively so). Interestingly, the most popular forum among U.S. workers is Reddit HWTF while international workers are most likely to use MTurkForum.

The topics discussed among these workers also differ. In particular, a larger fraction of U.S. pairs communicate about HITs (83.2% U.S. vs. 61.2% international), while international workers are much more likely to chat about day-to-day life (12.5% U.S. vs. 30.2% international). This finding coincides with the finding in Figure 3.6 that forum communication is more likely to focus on HITs while one-on-one communication is more likely to focus on day-to-day life.

In terms of network position, most of the connected U.S. workers (79.8%) are part of the largest connected component, while the majority of connected international workers (77.9%) are in smaller components.

Taken as a whole, this analysis resolves a question left open from Section 3.3.1: Who are the connected workers who lie outside the largest connected component? These are largely international workers who mostly communicate one-on-one on topics not limited to MTurk work only and are most likely to use MTurkForum if they use a forum at all.

3.4 Discussion

We designed and executed a HIT to map the network of workers on Amazon Mechanical Turk, and showed that there is a substantial communication network within the crowd. Put another way, the crowd is not a collection of independent workers. It is a network. The largest connected component of this network is made up mostly of U.S. workers communicating on various online MTurk forums on which discussion is mostly focused on aspects of MTurk work such as sharing HITs. The network additionally contains many smaller components composed largely of international workers talking with each other through one-on-one channels in which conversations focus on topics like the workers' day-to-day lives in addition to MTurk work. Workers who are part of the network tend to communicate with other workers who are similar to themselves in terms of geographic location, worker experience, and the types of tasks they prefer. Being part of the network may confer some informational advantages to workers allowing them to hear about HITs before workers who are not part of the network. Overall, connected workers tend to be experienced and of high quality.

The existence of the network within the crowd has implications for requesters, workers, and platform designers. Requesters should be aware that the workers they recruit are not an independent sample from the community of active workers. Instead, workers are effectively sampled from a network of workers bound together by the online forums they use or the type of tasks they prefer to do. Since there is homophily among workers, if one worker does a HIT she is more likely to recruit a fellow worker who is similar to her to do the HIT next. If a requester who is using Mechanical Turk to conduct behavioral experiments [Rand, 2011, Mason and Suri, 2012, Paolacci et al., 2010] randomly assigns workers to the treatment and control groups, both groups are still statistically equivalent in all aspects. However, such requesters should carefully consider if the treatment itself would be artificially increased or decreased depending on the characteristics of the population sampled. This is especially

true for characteristics like location and experience with MTurk, for which we have shown homophily. Additionally, since workers frequently communicate with one another about HITs, it is natural to ask whether the work that they submit is generated independently or whether they may, for example, share answers with one another. Any discussion among workers of the contents of HITs could bias results.

Our results show that many workers share lucrative tasks and information about reputable requesters with their network connections. With access to this extra information, connected workers might be able to start on high quality tasks before other workers hear about them. In the extreme, this might lead to connected workers using up all of the high paying tasks before isolated workers have had a chance to find them, effectively starving out the isolated workers. Thus, we speculate that being a part of the network may confer an advantage to workers.

All of the forums discussed in this paper were built by workers and exist outside of the Mechanical Turk platform and website. We offer two explanations as to why workers would spend their time building and using these forums. First, it could be the case that participation in forums results in higher pay for workers since they gain access to information about lucrative tasks, as discussed above. Beyond that, workers might inherently value the social interactions that these forums provide. A quote from [Zyskowski and Milland \[2015\]](#) indicates that some workers value online forums for both of these reasons: “If I had not found TurkerNation, I would not have made as much money for sure. And the fun we have when things are slow: priceless.” Platform designers should be aware that some functionality of their site is missing, so much so that workers felt the need to build that functionality on their own, at their own expense. Crowdsourcing platforms should perhaps consider whether there are ways to make their sites more social and provide workers with the interaction they clearly value.

3.5 Acknowledgements

The research in this chapter was produced in collaboration with Mary Gray, Siddharth Suri, and Jennifer Wortman Vaughan, during an internship at Microsoft Research. Portions of this chapter previously appeared in the WWW publication *The Communication Network within the Crowd* [Yin et al., 2016]. We thank Andrew Mao for helpful technical advice, Kati London for helpful design advice, and the anonymous WWW 2016 reviewers for many helpful comments.

Chapter 4

Managing the Flexibility of Crowdwork

On-demand platforms often attract workers by claiming that the on-demand work opportunities on their platforms have a unique advantage compared to the traditional jobs in companies and organizations, that is, the *flexibility*. For example, Amazon Mechanical Turk, a leading on-demand crowdsourcing platform, advertises on its website that a Mechanical Turk worker can “work from home, choose your own work hours, and get paid for doing good work.” Meanwhile, on-demand workers may indeed attach crucial value to the flexibility of the work beyond the immediate price-per-task payment that they earn from their work. For example, it is found that for workers on crowdsourcing platforms, the job flexibility provided by the crowdwork is a major factor that is associated with a favorable preference for workers to pursue a crowdsourcing career [Deng and Joshi, 2013].

At the first glance, the on-demand work is indeed quite flexible as workers seem to be able to choose whatever time, location, and manner to work that is most convenient for them. However, a few recent studies have suggested that the on-demand work may not be as flexible as it has been advertised, especially in terms of the temporal flexibility. For example,

Check if the website has an alternative language or a market			
Requester:	 ET	HIT Expiration Date:	May 10, 2017 (1 week 2 days)
		Time Allotted:	60 seconds
		Reward:	\$0.02
		HITs Available:	633

Figure 4.1: A requester needs to set a parameter of the maximum amount of time allotted for each task on Amazon Mechanical Turk.

in contrast to our impression of on-demand workers typically completing the work during their spare time to earn some extra cash, it is recently found that most participants in the on-demand economy treat it as a secondary source of income to address the income fluctuation in their traditional wage or salaried jobs so that they can cover their living expenses [Farrell and Greig, 2016]. For some workers, the on-demand work is even their sole source of income. As such, instead of completing on-demand work only when they *want* to do so, many workers in fact try to complete as much on-demand work as possible, *whenever it is available*, in order to make a living [Smith, 2016b]. The nature of “on-demand” work then decides that for these workers, they actually need to work whenever there is demand from customers (e.g., work during the rush hours as an on-demand driver) rather than enjoying the flexibility of adjusting their own working schedule.

Moreover, even if a worker can fully decide when she would like to participate in the on-demand work, she may still face additional constraints within each individual task that she works on. Take Amazon Mechanical Turk (MTurk) as an example—on MTurk, when a requester posts a task, one important parameter that he needs to set is the maximum amount of time assigned to the task, which is referred to as “*time allotted*” (see Figure 4.1 for an example). A worker needs to complete the task within this time limit in order to get paid; otherwise, the task will be expired and she may not be able to accept the task again. As most tasks on MTurk are “micro-tasks” that typically can be completed within a few minutes, it is not uncommon for a requester to set the time limit to be rather short. In fact, the default value of “time allotted” for a task on MTurk is 1 hour, and for about 85% of the

tasks on MTurk, the requesters choose to set the time allotted to be no longer than 1 hour¹. These short time limits of individual tasks inevitably lead to certain inflexibility for workers: workers may not be able deal with occasional interruptions in the tasks like restroom breaks or picking up a phone call without having the tasks expired; in addition, workers may find themselves quite constrained in scheduling tasks—for example, in a survey on crowd worker’s experience, one stay-home mother suggested that the short time allotted on tasks prevented her from accepting those tasks that she would like to conduct because they conflict with her child care responsibilities [Deng et al., 2016].

In the traditional workplaces, there are a large number of studies clearly suggesting that the flexibility of a job can influence both the work outcomes and workers themselves [Baltes et al., 1999, Joyce et al., 2010, Nijp et al., 2012]. It is, therefore, natural to ask in the new on-demand work settings, whether workers are influenced by the flexibility of the work in a similar way. This is, in fact, the first question we attempt to answer in this chapter. In particular, while it is difficult to control the flexibility of on-demand work by manipulating the timing of demand, it is relatively easy to manage the flexibility *within* a task by allotting different amount of time to each task, which we refer to as the *in-task flexibility*. Granting sufficient in-task flexibility then implies the ability for a worker to control her working time in a task once she decides to take it, including deciding when to actually start to work on the task and whether and when to take breaks within the task.

In this chapter, we first focus on examining whether and how does the in-task flexibility influence the worker engagement, performance and working behavior (e.g., the ways that workers complete a task) in the on-demand crowdwork through an experiment on Amazon Mechanical Turk. Our results suggest that providing more in-task flexibility to crowd workers not only leads to significant improved levels of engagement and performance, but also changes

¹This percentage is arrived by examining the time allotted for all available tasks on MTurk on May 1, 2017, using the author’s own MTurk worker account.

the way that workers work on tasks. More specifically, the working behavior data that we collect in the experiment is consistent with the hypothesis that more in-task flexibility allows workers to work at their own pace (e.g., take breaks as needed within the tasks; therefore take fewer breaks between subsequent tasks) as well as schedule their tasks in an efficient way (e.g., work on urgent tasks with tight time limits first while putting other tasks in the queues). Moreover, it is observed that workers who are more active in participating in the on-demand crowdwork (e.g., workers who spend more hours on MTurk in a week) are more likely to leverage the extra amount of time allotted to them in each task.

The positive association between in-task flexibility and work outcomes such as engagement and performance seems to indicate that workers value the in-task flexibility. To further understand that from worker’s point of view, how important it is to have sufficient flexibility within a task, we conduct a survey to explicitly measure the economic values of in-task flexibility for workers. Based on our survey, we find that about 65%–70% of the workers attach a positive value to the in-task flexibility. In particular, it is estimated that on average, workers are willing to take a pay cut of at least \$0.82/hour to work on tasks which give them more freedom in controlling their time.

4.1 Related Work

The impact of *job flexibility* on workers has been extensively studied within traditional companies and organizations in the organizational behavior and psychology literature. While the broad term of “job flexibility” includes the flexibility in various dimensions like work schedule and work location, the concept of “*temporal flexibility*” or “*work time control*” specifically refers to the flexibility regarding working times. The temporal flexibility can be further divided into a number of sub-dimensions, including the control over when to start and end the workday (i.e., “flextime”), when to take breaks, when to take days off or work

overtime, etc.

A large number of studies have been conducted to understand the relationship between the temporal flexibility in traditional workplaces and many job-related outcomes as well as different aspects of worker’s lives. For example, it was found that an increase of temporal flexibility in terms of flextime was associated with positive effects on worker productivity and job satisfaction [Baltes et al., 1999]. Flexible working arrangement such as self-scheduled shifts was observed to lead to improved health conditions and wellbeing [Joyce et al., 2010]. Research also suggested that organizational interventions that were designed to promote greater employee control over work time not only reduced the perceived stress for employees [Moen et al., 2016a], but also lowered the turnover intentions [Moen et al., 2016b]. In addition, there were further evidence which supported a positive association between temporal flexibility and the work-life balance of workers [Hill et al., 2001, Nijp et al., 2012]. We refer interested readers to a recent systematic review by Nijp et al. [2012] for more information.

Furthermore, two major types of mechanisms are provided to explain why temporal flexibility can significantly influence job-related outcomes and worker’s lives. The *time-regulation mechanism* suggests that the work time control allows workers to better regulate their time demands, such as reduce the level of work-family conflict [Geurts and Demerouti, 2003, Shockley and Allen, 2007]. Meanwhile, the *recovery-regulation mechanism* indicates that the temporal flexibility may give workers the opportunities to lessen the fatigue from work by taking breaks as needed or prevent the work overload at the first place [Costa, 2003, Nijp et al., 2012].

More broadly, job flexibility is a part of *job autonomy*, which refers to the freedom, independence, and discretion to plan out the work and determine the procedures in the work, and job autonomy is one of the five “core” characteristics of a job as suggested in the *job characteristics theory* [Turner and Lawrence, 1965, Hackman and Oldham, 1980]. According to this theory, job autonomy, together with other four core job characteristics (i.e.,

skill variety, task identity, task significance and feedback), can directly and indirectly affect employee’s work related attitudes and behaviors, including worker motivation, satisfaction, performance, absenteeism and turnover [Hackman and Lawler, 1971, Hackman and Oldham, 1975]. In addition, the *self-determination theory* in psychology [Ryan and Deci, 2000b, Deci and Ryan, 2012] also includes autonomy as one of the three basic psychological needs for people to self-motivate (i.e., be intrinsically motivated). For example, it was experimentally showed that when explicit deadlines were imposed on a task and hence the levels of autonomy for workers were limited, workers became less interested in the task [Amabile et al., 1976].

Our work is different from the previous studies in two ways. First, we focus on examining the impact of flexibility on the *on-demand crowdwork*, which is often composed of small-sized tasks and generally believed to be more flexible than traditional jobs. It is thus interesting to see whether the level of flexibility in the work still has a similar effect on job-related outcomes and working behavior for on-demand workers. To the best of our knowledge, our work is the first study to answer this question. Second, in this study, we restrict our attention on understanding the effects and values of the *in-task flexibility*, which is reflected by the amount of time allotted to each task and describes the worker’s freedom in controlling their working time *within* individual tasks. As an analogy, the in-task flexibility in the on-demand work is similar to the flexibility within each project (e.g., whether a tight deadline is imposed on a project or not) for a traditional job rather than the flextime. Our study, therefore, specifically explores how the in-task flexibility affects crowd workers.

4.2 Experimental Design

In this section, we describe an online experiment on Amazon Mechanical Turk that we designed and conducted to understand whether and how granting workers with more in-task flexibility (i.e., allotting extra amount of time to tasks, which allows workers to control their

own working time in the tasks) can affect worker’s engagement, performance as well as their working behavior in the tasks.

4.2.1 The Sentiment Analysis Tasks

The tasks we used in this experiment were sentiment analysis tasks. In particular, each task contained an Amazon customer review for an automobile related product, and workers were recruited to analyze the sentiment in the review. The set of customer reviews in the tasks were taken from [McAuley and Leskovec, 2013]. Each review used in our task had 150–200 words, and workers were asked to indicate whether the review is positive or negative in their opinion. As the ground truth, we got access to the actual customer rating associated with each of the reviews on a 5-point scale, with a higher rating indicating a higher satisfaction level. We classified reviews with a rating of 4 or 5 as positive reviews, and reviews with a rating of 1 or 2 as negative reviews. Reviews with a rating of 3, which we determined as neither positive nor negative, were therefore *not* used in our tasks. Through a pilot study, we found that it took a worker about 30 seconds on average to read one review and determine the sentiment in it. Figure 4.2 shows an example of the sentiment analysis task.

4.2.2 A 3×2 Factorial Design

By controlling how much time we allotted to each sentiment analysis task and whether we provided an estimate of the task completion time in a task, we created a set of six treatments. In particular, each treatment is defined by the following two dimensions:

- *time allotted*: the amount of time allotted in a task, with three possible levels—1 minute, 1 hour, and 1 day;
- *provision of time estimate*: whether to provide an estimate of the task completion time in a task—if time estimate is provided, we will let workers know that we expect that it

Sentiment analysis

We have a large number of customer reviews on automobile related products and we want to know whether these reviews are positive or negative. Please help us classify these customer reviews.

Please carefully read the following customer review and decide whether it is a positive review or a negative review. You can classify as many reviews as you want.

This is an idea whose time has not arrived. Perhaps if one has a teeny tiny vehicle then this may work out. First of all, you have to purchase micro-fiber towels per the instructions. Be sure to get lots of them. You are instructed to do one panel at a time - spray, wipe and buff. Supposedly you will save 80 - 120 gallons of water and eliminate toxic runoff (apparently referring to the soapy water). I HAVE NEVER USED ANYWHERE NEAR THAT MUCH WATER TO WASH MY CAR. With eco touch I am left with a whole slew of micro-fiber towels that would end up in a landfill because I don't know how to recycle dirty towels. I believe that eco touch has a problem with "truth in advertising". I can see using this for minor cleaning such as a deposit left by a bird or an insect meeting its demise on my vehicle. But, to wash the entire car is best done with a bucket, hose, sponge and chamois cloth. Or, take it to a car wash.

In my opinion, this review is:

☐ positive
 ☐ negative

Submit

Figure 4.2: Example of an sentiment analysis task.

takes roughly 30 seconds to complete one sentiment analysis task; if time estimate is not provided, we will not tell workers this information.

With the combination of 3 levels of time allotted and 2 possible values for providing time estimate, we have a 3×2 factorial design which led to a total of six treatments.

As our pilot study suggested that completing one sentiment analysis task took 30 seconds on average, we in fact granted workers enough amount of time to complete the task for all three levels of time allotted. However, compared to workers who were allotted 1 minute for each task, workers who got 1 hour or 1 day for each task had much more in-task flexibility and thus were able to control their working time in the task (e.g., decide when to start the task, whether to take a break in the task) to a much larger degree. Comparing worker engagement, performance and working behavior across treatments with different time allotted can thus help us to understand the impact of in-task flexibility on crowd workers.

On the other hand, one may wonder whether workers would use the amount of time allotted in a task to infer the difficulty of the task—for example, workers may take the time allotted as a proxy for how long it would cost to complete the task. If this is indeed the case,

one can imagine that, for example, workers who were allotted 1 day for each task would not interpret getting extra amount of time for a task as more in-task flexibility, but rather an indicator of the task being complex and time-consuming. To control worker’s perception on the time allotted, we added in the second dimension of “provision of time estimate” and explicitly informed workers about the estimated completion time of a task in those treatments where time estimate was provided. By examining whether there is any difference in worker engagement, performance, and working behavior across treatments with and without time estimate, as well as the interactions between the two factors (i.e., time allotted and the provision of time estimate), we can have a more in-depth understanding on how workers interpret and thus be affected by the amount of time allotted in a task.

4.2.3 Experimental Procedure

A two-phase experiment. We conducted our experiment in two phases. The first phase is the recruitment phase, in which we posted a 20-cent participant recruiting HIT for future sentiment analysis tasks on Amazon Mechanical Turk (MTurk) on September 8, 2016. Workers who were interested in completing the future sentiment analysis tasks can sign up by answering three survey questions about their usage of MTurk (i.e., number of years using MTurk, number of hours working on MTurk in the last week, number of income sources out of MTurk) in the recruiting HIT and submit it. Workers were informed that we would include all workers who submitted the recruiting HIT into the participant pool for the sentiment analysis tasks, hence they were instructed to answer the survey questions honestly. The second phase is the phase for the actual experiment, which was conducted on September 12, 2016. Each worker who signed up through the recruiting HIT was randomly assigned to one of the six treatments and was provided with 100 sentiment analysis HITs, with each HIT containing one sentiment analysis task. We communicated with workers about these sentiment analysis HITs through email once we launched them. Depending on the treatment that the worker

was assigned to, the amount of time allotted to each task can be 1 minute, 1 hour or 1 day, and a completion time estimate may or may not be included in the instruction of the task. Workers were asked to complete as many sentiment analysis HITs as they want, and they can get a fixed payment of 5 cents for each HIT they completed. While workers were working on the tasks, we recorded data on their engagement, performance and working behavior in the tasks, which we will detail later.

Such design of two-phase experiment allowed us to mimic the requester’s management on in-task flexibility in a most natural way. In particular, consider an alternative scenario where we don’t have the recruiting HIT of the first phase and directly assign workers into one of the six treatments upon their arrival in the actual experiment HIT in an *online* fashion. In order to control the time allotted in a task, we will have to set an extra timer of 1 minute, 1 hour or 1 day on the task *within* each HIT in addition to the default timer of the HIT that we have already set when defining the parameter of time allotted for the HIT, which is before the random assignment of workers. From the worker’s perspective, this “embedded timer” can be confusing and can potentially lead to negative discussions about our experiment HITs on online forums. On the contrary, when we have the separate recruiting phase, we can conduct the random assignment of workers *offline*. Thus, in the actual experiment phase, workers who are assigned to a treatment with time allotted being T ($T \in \{1 \text{ minute}, 1 \text{ hour}, 1 \text{ day}\}$) will be exposed to a group of 100 HITs where each HIT has the time allotted parameter being set as T , and there is no need for setting an extra timer within the HIT. Importantly, using the MTurk qualification, we ensure that workers will only be able to see and work on HITs in the treatment that they were assigned to, although we post the sentiment analysis HITs for all treatments at the same time.

Experimental data. We kept track of a wide range of engagement, performance and working behavior data while workers completed our sentiment analysis tasks. More specifically,

we used the number of sentiment analysis tasks that a worker *accepted* and *completed* as the metrics for measuring worker engagement. Given that we only presented 100 sentiment analysis HITs to workers, both the number of accepted tasks and completed tasks had an upper bound of 100. We further compared worker’s answer in each task to the ground truth, determined whether the answer was correct and then calculated the *accuracy* of a worker by averaging over all tasks that she completed. Worker’s accuracy was then used as the metric of worker performance.

In addition, we also kept a detailed log for each worker’s interaction with each task that she completed. In particular, when the worker i accepted a task t , we recorded a timestamp a_i^t as the time for *task acceptance*, which was also the time that the worker entered the task for the first time. Once a worker accepted a task, the task would be automatically added into her HIT *queue* on MTurk. The worker can either immediately start to work on the task or she can continue to search for other tasks, and the tasks that she accepted would stay in her queue until the time allotted for these tasks was reached. Therefore, depending on how worker i interacted with task t after she accepted it, we recorded some additional timestamps—if worker i had task t open in her browser ever since she accepted the task, then the only other timestamp (if any) we collected for the worker on this task is s_i^t , which was the time for *task submission*²; on the contrary, if the worker searched for other tasks after accepting task t (so she didn’t keep task t open in her browser) and then came back to work on task t later, in addition to the task submission timestamp s_i^t , we also kept another sequence of timestamps $r_i^t(j)$, $1 \leq j \leq n_i^t$, with $r_i^t(j)$ representing the time when worker i *re-entered* task t from her HIT queue for the j -th time, and n_i^t was the total number of times that worker i re-entered task t . Naturally, we have $a_i^t < r_i^t(1) < \dots < r_i^t(n_i^t) < s_i^t$. We next

²Note that having a task open in the browser doesn’t imply that the worker is working on the task. For example, the worker can work on one task while keeping other tasks open in her browser, or the worker can take a break within a task. It is not practical for us to monitor when the worker actually works on the task. It is also possible that a worker accepts a task but never submits it; she may return the task, or let it expired.

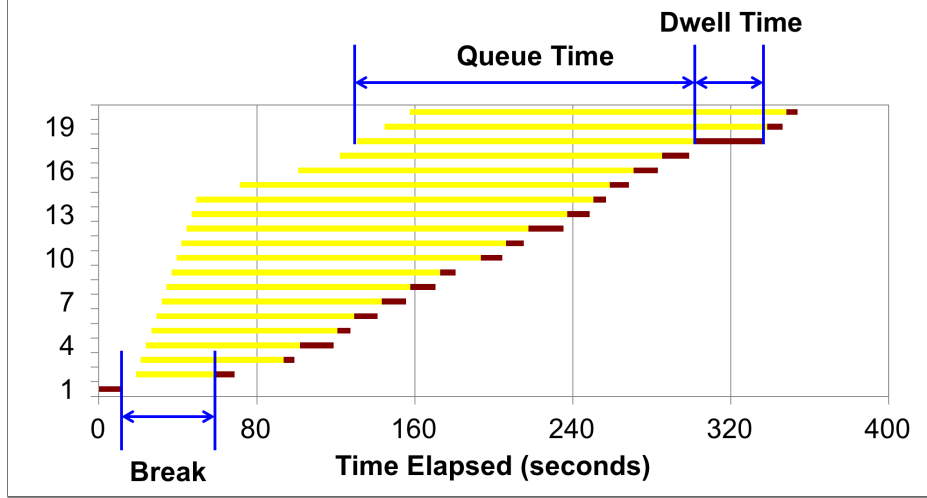


Figure 4.3: An example for defining metrics of working behavior. In this example, the worker completes 20 tasks in total, and the log for each task is represented as a horizontal bar: the leftmost end of each bar represents the task acceptance time; the rightmost end of each bar represents the task submission time. Each bar may be further divided into a yellow part and a dark red part. The transition point from yellow to dark red in each bar (if any) represents the time when the worker re-enters the task for the last time.

sorted all the tasks that worker i completed according to the increasing order of the task acceptance time. A *break* between the subsequent two tasks then refers to the gap between the time that worker i submitted one task and the time that she entered the next task for the last time.

Based on the log of worker's interaction with tasks, we defined the following metrics for measuring worker's working behavior:

- *queue time* (q_i^t): the amount of time elapsed from worker i accepting task t to entering the task for the last time. When $n_i^t = 0$, $q_i^t = 0$; otherwise, $q_i^t = r_i^t(n_i^t) - a_i^t$.
- *dwell time* (d_i^t): the amount of time elapsed from worker i entering the task for the last time to submitting the task. When $n_i^t = 0$, $d_i^t = s_i^t - a_i^t$; otherwise, $d_i^t = s_i^t - r_i^t(n_i^t)$.
- *number of x -minute breaks*: the total number of breaks a worker took that were longer than x minutes.
- *first x -minute break timing*: the total number of tasks that had been completed before

a worker took her first break that was longer than x minutes.

Figure 4.3 gives a visual example on how the above working behavior metrics are defined. Comparing these metrics across different treatments allows us to thoroughly understand whether granting more in-task flexibility affects worker behavior in terms of how they process tasks, including how long they put a task in their queues, how long they dwell on a task, how many times they need to take breaks, and how early they need to take a long break. Our hypothesis is that more in-task flexibility allows workers to complete tasks according to their own pace (e.g., take breaks within tasks if needed and thus increase the dwell time on a task while decrease the needs for breaks between subsequent tasks). Moreover, with more in-task flexibility, workers may be able to schedule the accepted tasks in a more efficient way to minimize possible conflicts (e.g., put tasks with higher levels of in-task flexibility in their queues for a longer period of time in order to cater for more urgent tasks first).

4.3 Results

In total, 1,999 workers signed up to the sentiment analysis tasks through the recruiting HIT in the first phase of our experiment. We then assigned each worker to one of the six treatments uniformly randomly. Among workers who signed up, 1,379 workers actually accepted at least one sentiment analysis task in the second phase of our experiment. No significant difference is observed across workers in different treatments in terms of either their usage of MTurk (i.e., their responses to the survey questions in the recruiting HIT) or the actual experiment participation rate (i.e., the percentage of signed-up workers in each treatment who actually participated in the sentiment analysis tasks).

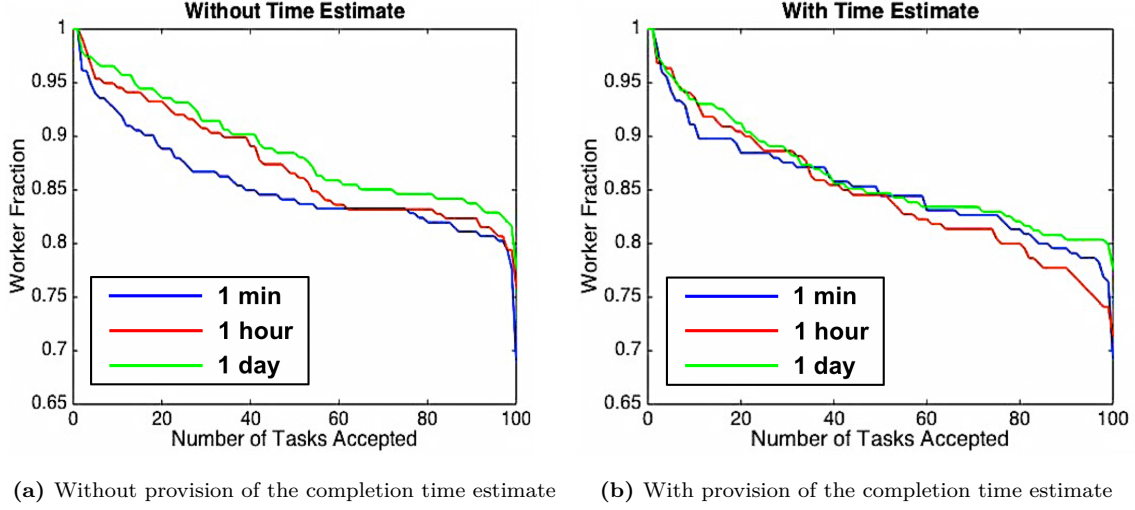


Figure 4.4: Retention curves showing the fraction of workers who continued to accept task after X tasks.

4.3.1 Impact on Worker Engagement

First of all, we examine whether granting workers with more in-task flexibility has any impact on worker engagement. We hypothesized that with more in-task flexibility, workers are more engaged in the tasks and thus are willing to work on a larger number of sentiment analysis tasks (i.e., accept and complete more sentiment analysis tasks). Meanwhile, we also conjecture that without the explicit estimate on how long it takes to complete a task, workers may use the time allotted in a task as an approximate for the time cost before accepting the tasks, which indicates a potential positive interaction effect between time allotted and the provision of time estimate—the improvement in worker engagement with time allotted can be larger when time estimate is provided.

Figures 4.4a and 4.4b show the curves on the fraction of workers who *accepted* at least X ($0 \leq X \leq 100$) tasks, for treatments without or with completion time estimate, respectively. Visually, we find that given a fixed X , the fraction of workers who accepted at least X tasks tend to be always higher in treatments where time allotted for each task is longer, and such difference is especially significant between workers in the 1-minute treatments and workers in

	# of acceptance (Model 1)	# of acceptance (Model 2)	# of submission (Model 1)	# of submission (Model 2)
Intercept	4.456 ^{***} (0.006)	4.446 ^{***} (0.007)	4.388 ^{***} (0.006)	4.394 ^{***} (0.007)
w/ time estimate	-0.021 ^{***} (0.006)	-0.002 (0.010)	0.005 (0.006)	-0.006 (0.011)
1 hour	0.009 (0.007)	0.024 [*] (0.010)	0.013 [†] (0.007)	0.008 (0.010)
1 day	0.026 ^{***} (0.007)	0.039 ^{***} (0.010)	0.014 [†] (0.007)	0.003 (0.010)
w/ estimate \times 1 hour		-0.030 [*] (0.014)		0.010 (0.015)
w/ estimate \times 1 day		-0.028 [*] (0.028)		0.023 (0.015)

Table 4.1: Negative binomial regressions for the number of tasks a worker accepted or submitted. Coefficients and standard errors are reported. The statistical significance of the estimated coefficient is marked as a superscript, with [†], ^{*}, ^{**}, and ^{***} representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

the 1-day treatments. This supports our hypothesis that more in-task flexibility improves worker engagement—with more time allotted in a task, workers tend to increase the number of tasks that they accept.

We next attempt to test whether the impact of in-task flexibility on the number of tasks that workers accept is statistically significant. The data on the number of tasks a worker accepts is highly skewed—as we can see in Figure 4.4, 70%–80% of the workers accepted all 100 sentiment analysis tasks that we offered to them. Therefore, we used negative binomial regressions to properly analyze these over-dispersed count data, and results are reported in Table 4.1. In particular, Model 1 on the number of accepted tasks (i.e., the second column of the table) reports only the *main effects* of the two factors, time allotted and the provision of time estimate. According to the regression results, providing extra amount of time in a task indeed increases the number of tasks a worker accepts, and such increase is statistically significant when allotting an excessively long period of time (i.e., 1 day) to the 30-second task. Interestingly, we also find that the main effect of the provision of time estimate is *negative*, suggesting that workers tend to accept fewer tasks when an estimate of the task completion time is presented in the task instruction.

To have a better understanding on how the provision of completion time estimate affects the task acceptance, we further consider a second model which includes the interaction between the two factors (i.e., Model 2 for the number of accepted tasks, shown in the third column of Table 4.1). Contrary to our conjecture, we find a significant *negative interaction* between time allotted and the provision of time estimate, implying that the improvement in worker engagement (in terms of the number of tasks accepted) is more significant when we don't provide a time cost estimate in the tasks.

We provide a few possible explanations for our observations. First, the sentiment analysis task used in our experiment is quite simple and straightforward, and it is also a common type of task on MTurk. Hence, it is relatively easy for workers to quickly estimate the time cost of the task by themselves, either through working on a few of these tasks or checking the related discussions about these tasks on online forums. As a result, unlike what we have conjectured, for this particular type of task that we used in the experiment, workers may not need to use the time allotted in a task to infer how long it will take to complete the task. Second, in treatments with the completion time estimate, the difference between the actual time cost of the task and the allotted time to the task is made *salient* to the workers, yet we do not explicitly explain *why* we allot extra amount of time in the tasks. Since it is unclear to workers why they were allotted such long periods of time for rather short tasks, workers may start to worry about the potential mismatch between the requester's expectation and their own understandings of the tasks, and thus may become hesitate to accept more tasks in order to minimize their risks. Finally, it is also possible that for some workers, they find the task time estimate we provide is not consistent with their own experiences and thus decide to stop accepting tasks due to psychological factors like mistrust to the requester or low levels of self-efficacy—the latter can be especially true if a worker finds that for her, completing a sentiment analysis task takes significantly longer than 30 seconds.

In addition to the impact of in-task flexibility on the number of tasks a worker accepted,

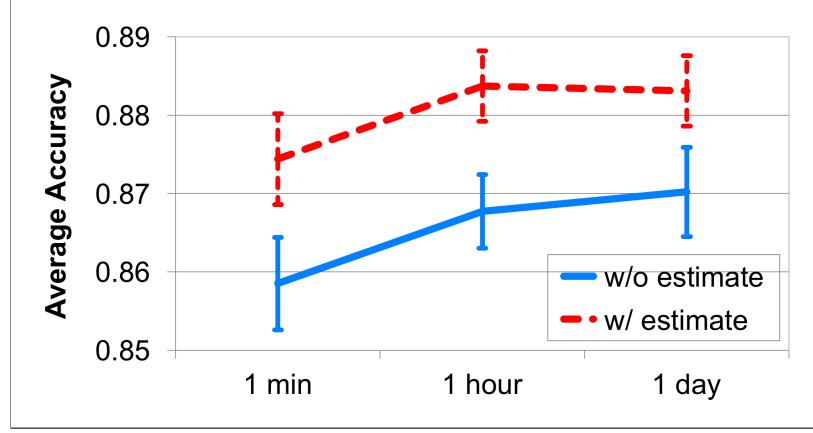


Figure 4.5: Comparison on workers’ average accuracy in different treatments. Error bars represent standard errors of the means.

we also look into the impact on the number of tasks a worker *completed* (i.e., submitted). Again, we conduct negative binomial regressions on the data and results are reported in the fourth and fifth columns of Table 4.1. Similar to the effects on task acceptance, we also find that granting extra amount of time in a task leads to an increase in the number of tasks a worker submits, and such increase is marginally significant. The provision of task time estimate, however, does not affect the task submission much.

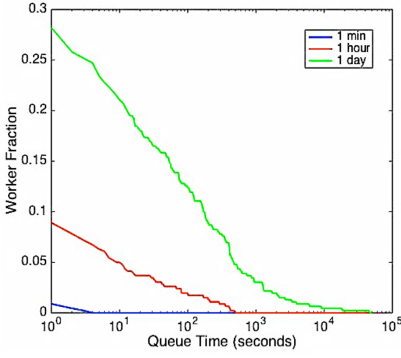
4.3.2 Impact on Worker Performance

Our second goal is to understand the influences of in-task flexibility on worker performance. Figure 4.5 displays the average accuracy for workers in each of the six treatments in our experiment, in which we clearly observe an upward trend as more time is allotted in a task, suggesting that granting more in-task flexibility can also improve worker performance in the tasks. Meanwhile, we also find that providing a completion time estimate in a task is associated with a higher worker accuracy, and no interaction effect between time allotted and the provision of time estimate is observed through our visual inspection. We further confirm the significance of the worker performance improvement through statistical tests—as the worker accuracy data is not normally distributed, a two-way ANOVA is not suitable

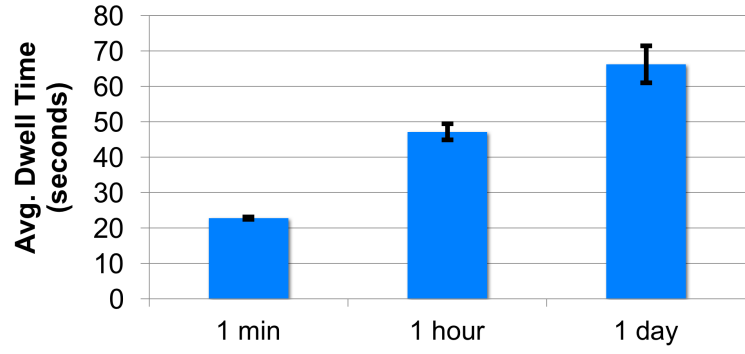
here. Thus, to examine the effect of time allotted on worker performance, given a particular level of time allotted $T \in \{1 \text{ minute}, 1 \text{ hour}, 1 \text{ day}\}$, we combine the worker accuracy data in the two treatments with the time allotted level T (and the task time estimate is either provided or not provided) together. In this way, we get three samples of worker accuracies, one for each level of time allotted, and we then use an one-way Kruskal Wallis ANOVA test to examine whether these three samples originate from the same distribution. The test result suggests that the differences in accuracy are statistically significant across treatments with different time allotted ($p = 0.022$). Pairwise comparisons further indicate that workers in the 1-day treatments (or 1-hour treatments) are significantly (or marginally significantly) more accurate compared to workers in the 1-minute treatments with $p = 0.029$ (or $p = 0.071$). Similarly, we combine the worker accuracy data in the three treatments with the same value on the provision of time estimate together, and a Wilcoxon rank-sum test confirms that the improvement in worker accuracy brought up by the provision of time estimate is statistically significant ($p = 5.080 \times 10^{-5}$). We conjecture that this is because workers interpret the provision of time estimate as a signal of the requester’s familiarity with his tasks and thus workers choose to consciously keep producing high-quality work to satisfy the requester.

4.3.3 Impact on Working Behavior

Finally, we explore the relationship between in-task flexibility and worker’s working behavior in the tasks (i.e., the ways workers interact with and complete tasks). As we don’t observe significant differences in working behavior between treatments with or without the provision of task time estimate, in the following, we focus on studying whether providing extra amount of time in a task will affect the working behavior of workers. Thus, for each level of time allotted, we combine the behavior data in the corresponding two treatments (i.e., one treatment with time estimate and another without), and the analyses are conducted on the three aggregated samples of behavior data.



(a) Average Queue Time



(b) Average Dwell Time

Figure 4.6: Comparison on the average queue time and average dwell time for workers in treatments with different levels of time allotted. 4.6a: the fraction of workers whose average queue time is longer than X ; 4.6b: the mean values for worker's average dwell time on tasks, and error bars represent the standard errors of the mean.

For each worker, we first calculate her *average queue time* and *average dwell time* by taking an average of the queue time and dwell time for all tasks that she completed. Figure 4.6a then compares the distributions of worker's average queue time when the amount of time allotted in a task varies, while Figure 4.6b demonstrates the mean values of average dwell time across different treatments. As the figures show, for any given X , the percentage of workers whose average queue time is longer than X is much higher when the time allotted in a task is longer, suggesting that workers tend to put tasks in their queues for a longer period of time if extra amount of time is allotted in a task. In other words, granting workers with extra time in a task effectively gives workers the flexibility to control when they *start* working on the task, which may further allow workers to optimally schedule tasks so that they can work on more urgent tasks (which can be either other accepted tasks in their queues with shorter time limits, or tasks outside of the on-demand work like taking care of their kids) first. Similarly, we also observe that worker's average dwell time on a task increases with the time allotted in the task, which may be because that with the extra amount of time allotted in a task, workers can take breaks *within* a task if needed. The differences in average queue time and

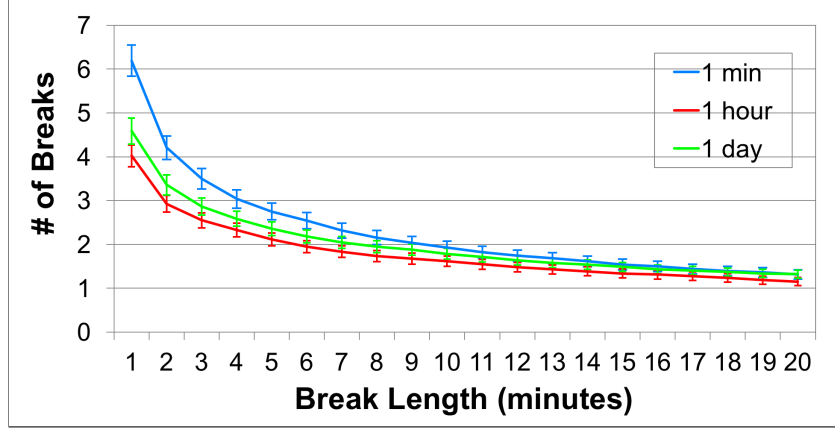


Figure 4.7: The average number of breaks a worker takes that are longer than X minutes in treatments with different levels of time allotted. Error bars represent the standard errors of the mean.

average dwell time across treatments with different levels of time allotted is also statistically significant according to the results of one-way Kruskal Wallis ANOVA tests ($p < 10^{-26}$).

We then move on to examine the impact of in-task flexibility on how workers take breaks *between* tasks. Figure 4.7 compares the average number of breaks of different lengths that a worker takes between subsequent two tasks when the time allotted in a task differs. In general, we find that compared to workers who are assigned to the 1-minute tasks, workers in the 1-hour or 1-day treatments seem to take much fewer breaks, especially when the length of the break is relatively short. Taking a closer look at the data, we further observe that when workers are allotted only 1 minute for the task, they need to take significantly more short breaks (i.e., breaks between 1 to 5 minutes, $p = 5.32 \times 10^{-15}$) and medium breaks (i.e., breaks between 5 to 10 minutes, $p = 0.005$), yet the number of long breaks (i.e., breaks that are longer than 10 minutes) they take is similar to that in treatments with the other two levels of time allotted. This is consistent with our hypothesis—when workers are assigned with relatively short period of time in the tasks, they don’t have much in-task flexibility and may feel quite constrained within the tasks; therefore, it is necessary for them to take more breaks *between* subsequent tasks to, for example, recover from fatigue or deal with interruptions

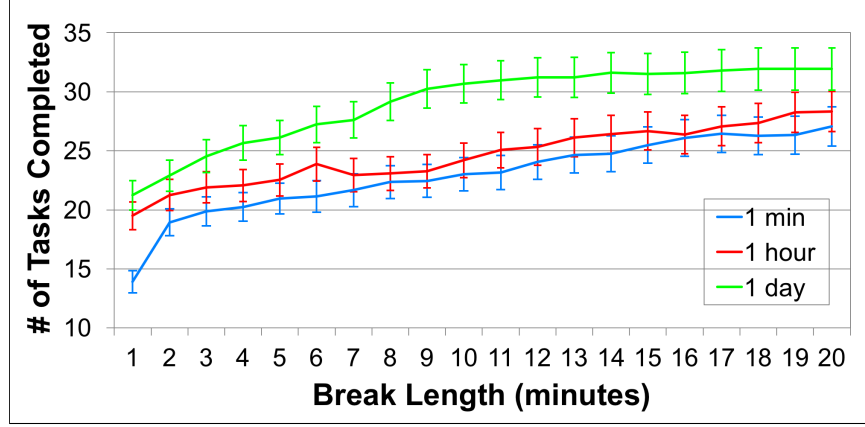


Figure 4.8: The average number of tasks a worker has completed before she takes her first break that is longer than X minutes.

in their working environment, because taking breaks in the tasks is not realistic. Moreover, Figure 4.8 further shows that workers in treatments with longer time allotted also take their first break after completing significantly more tasks. That is, with more in-task flexibility, workers not only take *fewer* breaks between tasks, but also take breaks *later*.

Finally, we study about the relationship between worker’s usage of MTurk and their working behavior in the tasks. More specifically, given a particular worker, we correlate data on her usage of MTurk (i.e., number of years using MTurk, number of hours working on MTurk in the last week, number of income sources out of MTurk) that we collected in the recruiting HIT with her working behavior (e.g., average queue time, average dwell time, etc.) in the sentiment analysis tasks, and we attempt to see whether workers with different characteristics will leverage the in-task flexibility to different degrees. While we don’t find significant differences in the working behavior for workers with different experience levels (i.e., number of years using MTurk) or different levels of dependence on MTurk (i.e., number of income sources out of MTurk), we do observe that workers with various *activity levels* on MTurk (i.e., number of hours working on MTurk in the last week) indeed work on tasks in different ways. For example, workers who worked on MTurk for more than 20 hours in the last week have significantly longer average queue time and average dwell time compared to

workers who worked on MTurk for fewer than 20 hours last week, regardless of the amount of time allotted in the tasks. This implies that workers with higher activity levels on MTurk are more likely to utilize the in-task flexibility through, for example, scheduling their tasks in a more optimal way and taking breaks in the task if needed.

4.4 Towards Measuring the Value of Flexibility

In the previous section, we have experimentally studied the impact of in-task flexibility on the workers. Our experimental results suggest that when workers are granted with more in-task flexibility, they improve their levels of engagement as well as performance in the tasks, and workers also adjust the way that they interact with tasks accordingly. In other words, these results appear to imply that the on-demand, crowd workers *value* the flexibility within the tasks, at least to some degree. A natural follow-up question to ask is *to what degrees* do the workers value the in-task flexibility. Inspired by the experimental approach for eliciting individual’s time preference in economics [Frederick et al., 2002, Hardisty et al., 2013], in this section, we set out to experimentally measure the economic values that workers attach to the in-task flexibility.

In particular, we designed a survey-based experiment with two treatments. The surveys were posted on MTurk, and workers who participated in the experiment were randomly assigned to one of the two treatments upon their arrival at the survey HITs. In both treatments, we first ask workers to imagine the scenario in which they are asked to complete a group of sentiment analysis tasks—it takes them roughly 30 seconds to complete each task and they can complete as many as 100 such imaginary sentiment analysis tasks in total. If a worker is assigned to the first treatment, she will be informed that the requester of the imaginary task allot 1 minute for each task and set the price for each task to be 5 cents. However, the requester doesn’t need the data from the sentiment analysis tasks immediately.

Therefore, he is willing to allot 1 day to each task so that workers can take breaks in the tasks if needed. The worker is then asked to indicate her preference in a sequence of five possible design pairs for this imagined task, where in each pair, the worker is asked to compare whether she would rather complete each sentiment analysis task *within 1 minute for 5 cents*, or *within 1 day for x cent(s)*, where $x \in \{1, 2, 3, 4, 5\}$. On the other hand, if a worker is assigned to the second treatment, she will be informed that the requester has set the time limit for a task as 1 day and the price for a task as 5 cents, but he suddenly needs the data as soon as possible. Hence, the worker is asked to reveal her preferred task designs in a sequence of five possible design pairs, where in each pair, the worker is presented with one option of completing each sentiment analysis task *within 1 day for 5 cents* and another option of completing each task *within 1 minute for y cents*, where $y \in \{5, 6, 7, 8, 9\}$.

We can then calculate the economic value that workers attach to the in-task flexibility for each 30-second task in an indirect way. For example, let's consider a worker in the first treatment—If she prefers the “1 minute, 5 cents” task design over the “1 day, x cent(s)” design for all x values, then the worker is not willing to give up any of her financial gains for the extra amount of time allotted in a task, indicating that she put a value of zero on the in-task flexibility for this task. As another extreme, if the worker always prefers the “1 day, x cent(s)” design regardless of the value of x , it means that in exchange for the in-task flexibility, the worker is willing to give up 4 cents or even more financially, implying that the value of in-task flexibility on this task for her is at least 4 cents. Finally, there is a third scenario in which there is a value $x_0 \in [1, 4]$ such that the worker prefers the “1 minute, 5 cents” design over the “1 day, x cent(s)” for $x \in [1, x_0]$, but she prefers the “1 day, x cent(s)” over the “1 minute, 5 cents” for $x \in [x_0 + 1, 5]$. Suppose in this case, the value of in-task flexibility is δ . The worker's preference then suggests that $x_0 + \delta < 5$ and $x_0 + 1 + \delta > 5$, that is, $\delta \in (4 - x_0, 5 - x_0)$.

In total, 202 workers were assigned to the first treatment, and for 8 of them, we can

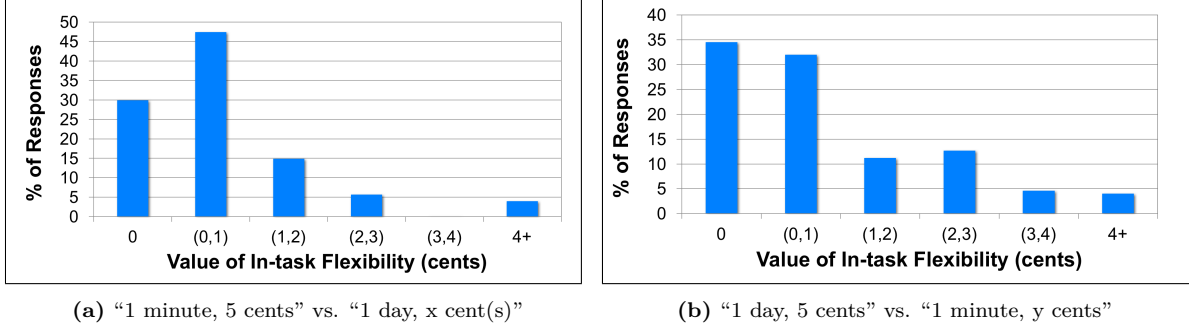


Figure 4.9: Estimate of the economic values of the in-task flexibility

not use the method above to compute a valid economic value for the in-task flexibility due to their inconsistency in the preferences³. Figure 4.9a thus shows the distribution of the computed economic values of in-task flexibility for the rest 194 workers in the first treatment. As the figure suggests, 70.1% of the workers has attached a non-zero value to the possible in-task flexibility that they will be able to get in the task, and 22.7% of the workers are willing to forego at least 1 cent for this 5-cent sentiment analysis task in exchange for extra amount of time in the task. When taking the lower bound for each worker's value of the in-task flexibility⁴, we find that the worker's average value for the in-task flexibility for each imagined task is at least 0.69 cents. Given that the imagined task is a 30-second task, our findings suggest that on average, workers are willing to forego a financial compensation of at least \$0.82/hour to get the in-task flexibility.

Similarly, for a worker of the second treatment, when the worker always prefers the "1 day, 5 cents" design, she effectively indicates that the in-task flexibility provided in this design is equivalent to at least 4 cents for the imagined task; when the worker always prefers the "1 minute, y cents" design, her value for the in-task flexibility is zero; and when there is a

³For example, a worker may indicate that she prefers "1 day, 2 cents" over "1 minute, 5 cents" while she also suggests that she prefers "1 minute, 5 cents" to "1 day, 3 cents."

⁴For example, if we compute a worker's economic value for the in-task flexibility to be $\delta \in (1, 2)$, the lower bound then is 1 cent.

$y_0 \in [5, 8]$ such that the worker prefers the “1 day, 5 cents” design over the “1 minute, y cents” for $y \in [5, y_0]$, but she prefers the “1 minute, y cents” over the “1 day, 5 cents” for $y \in [y_0 + 1, 9]$, we know that the value of in-task flexibility δ must satisfy the inequalities $5 + \delta > y_0$ and $5 + \delta < y_0 + 1$, and thus $\delta \in (y_0 - 5, y_0 - 4)$. Figure 4.9b displays the distribution of the computed economic values for the 197 workers in the second treatment (there are 3 workers whose responses are not self-consistent and their data is excluded from the analysis). Again, we find that 65.5% of the workers has a positive value for the in-task flexibility, with the average value being at least 1.01 cents. In other words, workers ask for an additional \$1.21/hour in order to give up the flexibility that has already been provided in the tasks. Comparing the economic values of in-task flexibility that we compute from workers in the first and the second treatment, we also notice that, interestingly, workers ascribe more value to the in-task flexibility when they have already “owned” it, even just in a virtual sense. This is consistent with a prevalent psychological bias, that is, the *endowment effect* [Kahneman et al., 1991].

4.5 Discussion

In this chapter, we examine how on-demand workers are influenced by the flexibility provided within the on-demand tasks. Our experimental results suggest that granting more in-task flexibility in the on-demand work leads to significant improvement on worker engagement and performance, and further influences the way workers behave when they work on the tasks. These results are consistent with the previous findings on the impact of temporal flexibility for traditional jobs, which imply that flexibility still plays an important role in influencing workers in the context of the on-demand work, even though it is generally believed to be already quite flexible. Furthermore, we also conduct a survey to estimate the economic values that crowd workers attach to the in-task flexibility. Our survey results confirm that a

significant fraction of workers are willing to forego substantial financial compensations for the capability to control their own time within the tasks.

There are a few interesting future directions for extending this work. First of all, in this study, we used a particular type of task (i.e., the sentiment analysis task) to understand the impact of in-task flexibility on crowd workers. The sentiment analysis task is a common type of task on crowdsourcing platforms and can be representative for a large number of simple tasks that mainly require human intuition and judgment (e.g., search query relevance, image annotation, etc.), hence we believe our results in this study is generalizable to many other tasks of different types. However, as we have briefly mentioned in Section 4.3.1, it is also possible that the nature of this particular type of task, to some degree, affects some of our experimental results, such as the interaction effects between time allotted and the provision of task time estimate. It is thus an interesting future work for us to explore whether our findings in this study still hold true for some significantly different types of tasks, such as tasks that are much more complex and time-consuming (e.g., writing, long behavioral experiments, etc.).

Furthermore, in this study, we measure the economic value that workers attach to the in-task flexibility through a survey that asks workers to indicate their preferences in a hypothetical working environment. Such preference elicited is thus “*stated preference*,” and it may not necessarily be consistent with the “*revealed preference*,” which is decided by worker’s actual decisions. Another interesting future work is, therefore, to measure the economic value of in-task flexibility by actually observing worker’s decisions when they are working on tasks with different flexibility levels. In particular, inspired by [Goldstein et al., 2014], one possible approach to consider is to estimate the economic value of in-task flexibility as the *compensating differential*, that is, the extra amount of money a requester would need to pay a worker to complete the same number of tasks as she would complete for tasks with more in-task flexibility.

Our findings also have important implications for both requesters of labor and workers in the on-demand economy. For example, from the requester’s point of view, as higher levels of temporal flexibility within the tasks can lead to improved worker engagement and performance, when designing the tasks, requesters should consider to provide more flexibility in the work whenever possible. In fact, it is already suggested in the Guidelines for Academic Requesters⁵ that requesters should set the “time allotted” limit for the HITs to an amount of time *much longer* than the expected amount of time needed to complete the task. As job flexibility is a part of job autonomy, our results further hint on the importance for requesters to acknowledge each worker as an autonomous identity and thus incorporate autonomy into the work design.

Meanwhile, from the worker’s point of view, we notice that even when we provide excessively long period of time in a task, there is only a very small fraction of workers who actually leverage such flexibility (e.g., fewer than 10% of the workers in the 1-day treatments put a task in the queue for longer than 2 minutes on average). In addition, as we have discussed in Section 4.3.3, it seems that workers who leverage the in-task flexibility more are those ones who are more active on the on-demand platform. We conjecture that these observations are a result of the current common practice, that is, tasks typically have short time limits and thus workers are “trained” to work in a pace that is as fast as possible. As we encourage requesters to consider increase the temporal flexibility in the tasks, it is also beneficial for workers to learn how they can best utilize such flexibility to both increase their efficiency and reduce their fatigue. One important source of learning is the online forums, as many highly active workers who is already quite familiar with leveraging the flexibility in the tasks tend to share their experience on these forums⁶.

⁵http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

⁶For example, there is a discussion thread on the online forum MTurkGrind named “Utilizing Your Amazon mTurk Queue”, which teaches workers to leverage the flexibility of tasks by effectively scheduling tasks

As a final note, we also recognize that a potential drawback of providing too much temporal flexibility in the on-demand work is the possible decrease in the task completion speed, as many workers may choose to work on other tasks that are more urgent first. We have not observed such decrease in our experiments—the task completion speed is similar across treatments with different time allotted—but again, this may be a result of that workers are “trained” to behave in a way as if the flexibility within the task is limited. Understanding the long-term effect of providing more flexibility within individual tasks, as well as examining the trade-off between task flexibility and completion speed, is yet another important direction for future work.

4.6 Acknowledgements

The work in this chapter was produced in collaboration with Mary Gray and Siddharth Suri during an internship at Microsoft Research.

through queues: <http://www.mturkgrind.com/threads/utilizing-your-amazon-mturk-queue.26362/>.

Chapter 5

Understanding the Effects of Financial Incentives

In previous chapters, we aim to better understand how the on-demand economy of today work, and in particular, we are interested in understanding who the crowd of on-demand workers are and how they behave in the work. Through a set of experimental studies, we have showed a number of key characteristics of the workers in on-demand economy: they have significant *temporal variations*, value *social interactions* and desire more *flexibility and autonomy*. Some of these characteristics lead us to pay close attention to certain human needs of on-demand workers (e.g., social needs like sense of belonging, ego needs like autonomy), which is previously overlooked, and presents useful insights for us to build a more desirable on-demand economy in the future.

Following this line of thought, in the second part of this dissertation, we ask how can we make the on-demand economy work better in the future. A variety of factors may come into play in addressing different kinds of human needs of on-demand workers and therefore improve the efficiency and sustainability of the on-demand economy. One such factor is the *incentive*. Indeed, human beings as they are, on-demand workers can be motivated

by various intrinsic and extrinsic incentives in completing their work [Benkler, 2002]. As such, many intrinsic motivations, such as the enjoyment in playing online games [Von Ahn, 2006, Savage, 2012] and the willingness to learn new knowledge [von Ahn, 2013], have been smartly integrated into the on-demand work environment to incentivize workers. For most on-demand crowdsourcing platforms like Amazon Mechanical Turk, however, the primary type of incentive remains to be extrinsic, that is, crowd workers complete tasks in exchange for monetary compensations.

Workers can be incentivized to exert more or less effort and be influenced by their psychological biases when the design of tasks or workflows affects these motives. Thus, a thorough understanding of how incentives affect work quality and worker effort in the on-demand economy is critical for developing methods to improve the effectiveness of these incentives. In this chapter, we focus on empirically examining the effects of extrinsic, financial incentives on workers of on-demand crowdsourcing platforms. Understanding the relationship between financial incentives and worker performance or productivity is, in fact, a very classical and fundamental question in the traditional economy. However, there are a few unique features of the on-demand work environment that make this question relevant again for the on-demand economy.

First of all, unlike that in the traditional economy, by design, the size of tasks in the on-demand economy is very small so that workers can make short-term contributions to these tasks while enjoying high levels of mobility and flexibility. As a result, payments for these tasks are often provided in the form of piece-rate payments (rather than hourly wage), and the amount of payment in each task can be quite small (e.g., earning several cents for tagging an image). In addition, on-demand workers can complete a large number of such small tasks in a very short period of time, enabling them to subsequently interact with financial incentives in the tasks in a much higher frequency compared to that in the traditional economy. These unique features naturally lead one to wonder whether on-demand workers react to small

piece-rate payments in a similar way as employees reacting to financial incentives in the traditional economy. Besides, they also suggest the importance of understanding the effects of financial incentives on work quality and worker effort in the context of a sequence of tasks, rather than just for each individual task.

Secondly, task switching is a very common practice for on-demand workers. While workers may actively choose to switch between different types of tasks to diversify their workload or avoid fatigue or boredom, many task switches are initiated by requesters as a result of the design of working sessions. For example, a requester may ask a worker to identify whether a pre-specified object (e.g., automobile or person) exists in each of a set of pictures and group tasks by the objects of interests; this results in task switches when the “target” object changes. Moreover, in many citizen science projects, tasks of different types are bundled into a single working session. For instance, in Cell Slider¹, a worker is shown an image of blood cells and needs to identify the types of cells, count the number of irregular cells and then estimate the brightness of the stained irregular cell cores; in Citizen Sort², a worker classifies the same group of moth pictures according to shape, color and forewing pattern respectively. One major challenge associated with task switching in the on-demand work is to ensure work quality for *all* tasks in a working session. It is well known that workers perform worse on *switch tasks*, tasks that follow another task of a different type, than on *repetition tasks*, tasks that follow another task of the same type [Rogers and Monsell, 1995, Monsell, 2003]. Therefore, it is straightforward for us to ask what role can financial incentives play to influence worker performance in these task switching settings.

We present two experimental studies in this chapter to address these questions on the effects of financial incentives that stem from the unique nature of the on-demand work—the

¹<http://www.cellslider.net/>

²<http://www.citizensort.org/>

first study (Section 5.1) explores how workers react to financial incentives in a sequence of tasks of the *same* type, while the second study (Section 5.2) examines whether and how financial incentives can be used to affect worker performance in a sequence of tasks of *different* types (i.e., in a task switching setting). In both studies, we focus on understanding the effects of a particular type of financial incentives, that is, the *performance-contingent* financial incentives. By “performance-contingent,” we mean that the amount of reward for a task depends on the quality of work produced in the task, where the quality is evaluated according to some metric of interest to the task requester. Such performance-contingent rewards are often used by requesters to encourage high-quality work from workers, making them a good candidate to study for the on-demand economy. Our experimental results suggest that in a sequence of tasks of the same type, workers are not always shown to be sensitive to the magnitude of performance-contingent financial incentives alone in each individual task, or in other words, workers don’t necessarily respond to the *absolute* magnitude of incentives. Instead, it has been robustly observed that the work quality and worker effort is affected by the changes of performance-contingent financial incentives in the subsequent tasks, or the *relative* magnitude of the incentives. Furthermore, we also find that in a sequence of tasks of different types, performance-contingent rewards are most effective in improving worker performance when they are placed on switch tasks in working sessions with a low task switching frequency. We finally conclude this chapter by discussing the implications of our findings in Section 5.3.

5.1 Placing Financial Incentives in a Task Sequence

In this section, we ask two questions regarding the effects of performance-contingent financial incentives on work quality and worker effort in a sequence of tasks of the *same* type:

- (1) How does the magnitude of financial incentive in each individual task *alone* influence work

quality and worker effort? (2) How does the work quality and worker effort affected by the *change* of incentives in the subsequent tasks?

To answer these two questions, we designed and conducted experiments on Amazon Mechanical Turk (MTurk), where we placed two tasks of the same type together within each Human Intelligence Task (HIT), and various types of tasks as well as different levels of performance-contingent financial incentives were considered. By varying the level of financial incentives in each task of the HIT and thus controlling the changes of financial incentives in the task sequence, we created a set of experimental treatments. The effects of performance-contingent financial incentives were then analyzed by comparing work quality and worker effort across treatments.

5.1.1 Related Work

Understanding the effects of various incentives in the on-demand crowd work settings is a quite active and open research area. For example, researchers have compared the produced work quality on MTurk when 14 different incentive schemes were used, including financial, social and hybrid schemes [Shaw et al., 2011]. They found that while most of the incentive schemes had weak effects on the work quality, two schemes which associated individual worker’s financial incentives with the responses from their peers elicited significantly better performance from the crowd workers.

A few previous studies focused specifically on examining the effects of financial incentives. It was demonstrated that for *performance-independent financial incentives* (i.e., a worker received a fixed amount of payment per task regardless of how well she performed in the task), while a larger amount of incentives motivated workers to complete more tasks, the quality of work in each task was not significantly improved [Mason and Watts, 2010, Rogstadius et al., 2011]. However, when the actual performance-independent financial incentive a worker received was higher than the previously contracted value and a portion of the increased

payment was clearly framed as an unexpected gift from the employers, workers were likely to reciprocate by exerting higher levels of efforts [Gilchrist et al., 2016].

As for the *performance-contingent financial incentives*, Harris [2011] studied the worker performance both with and without such incentives and showed that the existence of performance-contingent financial incentives led to higher work quality. More recently, Ho et al. [2015] carefully examined when, where and why performance-based payments can help improve crowd work quality, and they found that performance-based payments were most likely to encourage high-quality work for tasks that were effort-responsive, that is, tasks that allowed workers to improve the work quality by exerting more effort. Besides, it was also demonstrated that canonical economic games conducted on MTurk were comparable to those conducted in lab settings, suggesting that subjects in both pools responded to performance-contingent financial incentives in a similar way, although the magnitude of such incentives on MTurk might be much smaller [Amir et al., 2012]. Different from these studies, in this section, we attempt to understand not only how the magnitude of performance-contingent financial incentives in an individual task alone affects the worker performance, but also how the *change* of incentive magnitude in a task sequence influences workers.

In a broader context, the relationship between financial incentives and productivity has been extensively studied in economics and psychology prior to the existence of the on-demand economy, yet results from lab or field experiments diverge. While it was demonstrated in a lot of studies that higher level of performance-contingent financial incentives led to increased productivity [Pritchard and Curts, 1973, Lazear, 2000, Camerer and Hogarth, 1999], there were also evidences showing that such incentives had little influences on or even hurt the productivity [Jenkins Jr et al., 1998, Camerer and Hogarth, 1999]. A few explanations for the negative effects of financial incentives on productivity have been discussed. For example, the presence of a small financial incentive may cause the decrease of worker’s intrinsic motivation and further results in poorer performance, which is referred to as the phenomenon of *crowding*

out [Gneezy and Rustichini, 2000, Deci et al., 1999, Frey and Jegen, 2001, Bowles, 2008]. On the other hand, excessively large financial incentives could also exert deleterious influences on work quality, especially for tasks that require mostly intuitions or simple skills, as they may trigger worker’s overreaction [Ariely et al., 2009].

Different theories were proposed to characterize how workers react to financial incentives of varying magnitude. For instance, the model of *gift exchange* in sociology explains that when the employer shows his kindness by offering a wage that is higher than the market-clearing value, a worker tends to exhibit positive reciprocity by providing work in excess of the minimum quality standard [Akerlof, 1982, Dufwenberg and Kirchsteiger, 2000]. The *fair wage-effort hypothesis* [Akerlof and Yellen, 1990] in labor economics, which originates from the theory of equity in social psychology [Adams, 1963], provides another story. It states that if the actual wage a worker receives is less than the amount of “fair” wage in her mind, the worker may supply only a fraction of her normal effort levels. These theories were also supported by different experimental observations—Gneezy and List [2006] found that workers exhibited considerably higher levels of effort in the first few hours on the job when they received a “gift” of increasing wage compared to the advertised rate. Cohn et al. [2014], on the other hand, concluded from their experiments that workers who perceived themselves as underpaid at the base wage reciprocated the increasing hourly wage with improved performance, while those who felt adequately paid or overpaid at the base wage didn’t show significant responses to the wage increases. It is interesting to note that both these two theories seem to hypothesize that workers interpret the provided financial incentives by comparing the magnitude of the incentives to some references—the market wage or the fair wage, and their decisions on how much quality to produce or how much effort to exert in the work are thus influenced by their judgment on the kindness/unkindness or fairness/unfairness of the employers according to the comparisons. This is consistent with the well-known *prospect theory*, which points out that in decision-making, people are likely to set a reference

point and evaluate gains and losses against the reference point rather than deliberating over the absolute outcomes [Kahneman and Tversky, 1979].

While both theories above seem to imply that the work quality and worker effort can be sensitive to the magnitude of the provided financial incentives, an interesting subtlety to consider is that workers may not have a clear conception of the market value or the fair value of their work a priori. Hence, the formation of such conception can be affected by a prominent psychological bias, the *anchoring effect* [Tversky and Kahneman, 1974, Chapman and Johnson, 1994], which describes the common human tendency to rely heavily on the first piece of information offered (i.e., the “anchor”) in decision-making. In workplaces, the anchoring effect relates to a commonly observed worker behavior called *wage entitlement*, which refers to worker’s belief that they are entitled to their existing pay, no matter how high it may be, and the existing wage is further used as worker’s internal reference for accessing the fairness of other wage offers [Bewley, 2007]. In fact, in the post-task survey of Mason and Watts’s study on the effectiveness of financial incentives in crowdsourcing markets, workers systematically reported higher “appropriate” compensation levels than the actual payments they received in the tasks, and the reported compensation level also increased monotonically with the actual payment, suggesting workers might have used the latter as an “anchor” for setting the reference point of the appropriate payment level [Mason and Watts, 2010].

5.1.2 Experimental Design

We designed and conducted an online experiment to understand how on-demand crowd workers react to performance-contingent financial incentives in a sequence of tasks of the same type. In this experiment, we bundled *two* tasks of the same type into one Human Intelligence Task (HIT) and a worker was asked to complete both tasks in the HIT before she got paid. For each task in the HIT, a worker earned a performance-independent payment of 1 cent, which made the base payment for one HIT to be 2 cents. Besides, we also offered workers

with the opportunities to earn performance-contingent bonuses in each task. However, the magnitude of the performance-contingent bonuses for the subsequent two tasks in one HIT may or may not be the same. Specifically, we considered four levels of performance-contingent bonuses: 4 cents, 8 cents, 16 cents and 32 cents.

Treatments. By controlling the level of performance-contingent bonuses in the subsequent two tasks in the same HIT, we created the following three sets of treatments, defined by the bonus levels in the two tasks:

- *4 base treatments:* 4 – 4, 8 – 8, 16 – 16, and 32 – 32;
- *3 treatments with increasing bonus level:* 4 – 8, 4 – 16, and 4 – 32;
- *3 treatments with decreasing bonus level:* 8 – 4, 16 – 4, and 32 – 4.

For the four base treatments, as the bonus level in the subsequent two tasks in the HIT was the same, we were able to investigate whether work quality and worker effort is affected by the magnitude of performance-contingent financial incentives *alone*, or in other words, whether workers react to the “absolute magnitude” of financial incentives. On the other hand, the three treatments with increasing bonus level and three treatments with decreasing bonus level enabled us to examine how workers are influenced by the *changes* of financial incentives in the sequence, or in other words, how workers react to the “relative magnitude” of financial incentives.

Tasks. To see whether the effects of financial incentives are dependent on the nature of tasks, we considered two types of tasks in our experiment:

- *The button clicking (BC) task:* Two buttons of the same size are displayed on the screen, with one of them placed on the top while the other one on the bottom. One of the two buttons is green, which is the “target” button, and the other button is gray. A worker is asked to click on the target button, which will alternate between the top



Figure 5.1: Interfaces of the two types of tasks in our experiment

button and the bottom button, in a three-minute task session. To start the session, the worker needs to press the “Start” button. The worker is instructed to click on the target button as many times as she can, and the amount of time left in the session is also displayed on the screen. If the worker correctly clicks on the target button for more than 400 times in the session, she will earn the pre-specified bonus in that task.

- *The spotting differences (SD) task:* Two pictures are presented on the screen. These two pictures are almost identical with each other except for five non-obvious places. A worker is told the number of differences between the two pictures and is asked to find as many differing places as she can. The worker can spot a found difference by clicking on it in either picture to mark the difference with a red circle. If she finds any marked difference to be wrong, she can also deselect it by clicking in the corresponding red circle again. The worker gets the pre-specified bonus in a task if she spots all five differences correctly. Two different sets of pictures are used in the subsequent two tasks in the spotting differences HIT, and the order of their appearance is randomized. Prior study showed that the difficulty of spotting differences in these two sets of pictures is similar.

The button clicking task primarily requires the motor skills of a worker. A similar task was used by [Horton and Chilton](#) to estimate worker’s reservation wage in paid crowdsourcing

markets [Horton and Chilton, 2010], with the buttons in their task being aligned horizontally on the left and right of the screen. The spotting differences task demands mostly cognitive skills, as comparing two pictures carefully and finding the differences requires significant attention and concentration from workers and may also relate to workers' short-term memory. The interfaces of the button clicking task and the spotting difference task are shown in Figure 5.1a and Figure 5.1b, respectively.

Procedure. To understand whether crowd workers react to performance-contingent financial incentives in a consistent way as time passes by, we conducted the experiment twice: The experiment was initially launched from October to December in 2012 for the first time. Then, after three years, we replicated the experiment with the button clicking task from November to December in 2015, and with the spotting differences task from February to March in 2016.

For both the original and the replicating experiments, to avoid the complication of cultural differences in perceiving financial incentives, we restricted our experiments to U.S. workers. Within one experiment, each worker was allowed to participate in at most one button clicking HIT and one spotting differences HIT so that for either type of tasks, a worker would not be influenced by multiple bonus treatments. Upon arrival, a worker was randomly assigned to one of the ten treatments. The worker would then read the instruction in which we explained our payment rules, and took a qualification test to see whether they understood our payment rules. The worker could only proceed on to the actual tasks after passing the qualification test. To guarantee that workers pay attention to the magnitude of performance-contingent financial incentives in each task, especially the possible changes in the subsequent two tasks, the amount of bonus in a task was revealed *right before* each task instead of all together at the beginning of the HIT. Depending on the type of task in the HIT, the worker would then click buttons or spot differences in each task. The length for a clicking button task was fixed to be three minutes. In contrast, the worker could spend as much time as she wanted to find

differences in the spotting differences task. At the end of the HIT, the worker needed to complete an exit-task survey where she was asked to compare the bonus in the subsequent two tasks and identify whether the bonus in the second task was higher than, equal to or lower than that in the first task.

We decided to keep the magnitude of financial incentives (including both performance-independent and performance-contingent financial incentives) in the replicating experiment to be *exactly the same* as that in the original experiment in order to make sure that results in these two experiments are comparable. However, we were also aware of the fact that crowd workers nowadays have a higher expectation on the effective hourly wage (e.g., \$6–\$22/hour according to the Guidelines for Academic Requesters³) compared to those in the early years. In addition, as we discussed in Chapter 3, workers are also willing to share with other workers, through online forums or other communication channels, about their personal experience with different tasks and requesters [Martin et al., 2014, Gray et al., 2016]. Therefore, to make sure that workers in our experiment will not feel like being treated unfairly and to minimize the possible interference resulted from the external (and potentially negative) discussions about our experimental HIT and request account, we made two small adjustments when we replicated the experiment in 2015 and 2016: (1) We explicitly instructed workers *not* to discuss the HITs in online forums as they were a part of a scientific experiment; (2) In addition to the performance-independent base payment and the performance-contingent bonus, we further provided each worker with a *performance-independent bonus* when she reached the end of the HIT. The magnitude of this performance-independent bonus was computed such that the maximum amount of rewards a worker can earn in a HIT is always 70 cents. For example, if a worker was assigned to the 4 – 4 treatment in the replicating experiment, the performance-independent bonus she could receive at the end of the HIT is

³http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

$70 - (4 + 1) \times 2 = 60$ cents. As the average completion time for our HIT is 6–10 minutes, with the performance-independent bonus, our payment implied an effective hourly wage of \$4.2–\$7/hour for those workers who participated in the replicating experiment.

Data. In the original experiment in 2012, we recruited 1214 workers for the button clicking HITs and 1270 workers for the spotting differences HITs. To confirm that workers were indeed aware of the possible changes of financial incentives in the subsequent two tasks, we eliminated all data from those workers who gave wrong answers to the bonus comparison question in the exit-survey. Such elimination left us with 100 valid data points for each of the 10 treatments in each type of HITs. For the replicating experiment in 2015 and 2016, in total, 1119 workers participated in the button clicking HITs and 1224 workers participated in the spotting differences HITs. Similar to that in the original experiment, we eliminated all data from those workers who incorrectly answered the bonus comparison question in the exit-surveys. Again, after the elimination, 100 valid data points were preserved for each of the 10 treatments within each type of HITs. Only the valid data points are used in the subsequent analysis for both experiments.

To measure the work quality, we recorded the number of times a worker clicked on the “target” button in the three-minute session for a button clicking task, while the number of differences that were correctly identified by a worker is used as the quality metric in a spotting differences task.

As for the worker effort in the task, while there is no straightforward way to differentiate between the quality of work produced by a worker and her exerted effort in a button clicking task, we adopted two natural metrics to represent a worker’s effort in a spotting differences task. Specifically, we considered the log of a worker’s activities in a spotting differences task, which is a sequence of timestamps. For a worker who identified $n \leq 5$ differences correctly in a task, she had a log of $(t_0, t_1, \dots, t_n, t_{n+1})$, with t_0 being the time for her to load the task

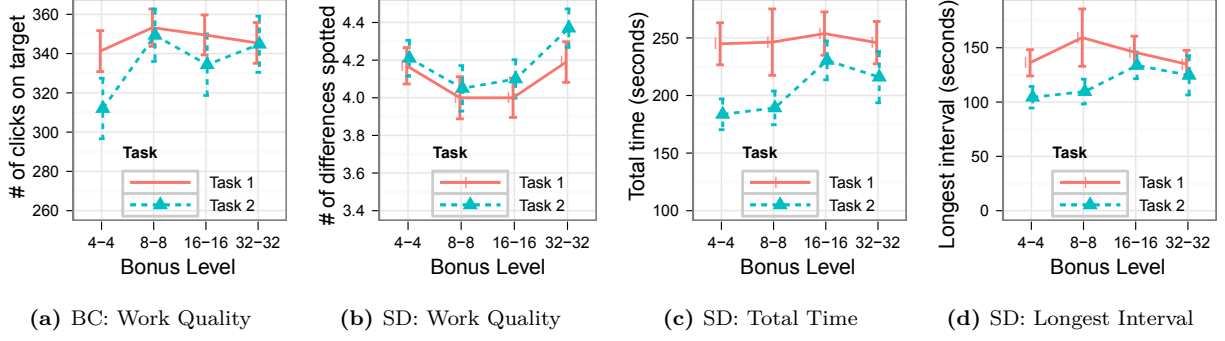


Figure 5.2: Work quality and worker effort in the four base treatments of the original experiment in 2012.

page, $t_i (1 \leq i \leq n)$ being the time at which the i -th difference was correctly identified, and t_{n+1} being the time of task submission. The first metric for assessing the worker’s effort is the *total time* she spent on the task, i.e. $t_{n+1} - t_0$, and the second metric is the longest elapsed time between two subsequent timestamps, that is, $\max \{t_1 - t_0, \dots, t_{n+1} - t_n\}$, which we refer to as *longest interval*. While our first metric captures the duration of the worker’s effort, our second metric characterizes the intensity of the worker’s effort, that is, how hard a worker tries in a task.

5.1.3 Effects of the Magnitude of Rewards Alone

Our first goal is to understand the effects of the magnitude of performance-contingent financial incentives *alone* on work quality and worker effort. We start with the results of our original experiment that was conducted in 2012: The mean values of the work quality metrics for both the first and the second task in each of the four base treatments are presented in Figure 5.2a and Figure 5.2b, with Figure 5.2a plotting the values for the button clicking HITs and Figure 5.2b showing the values for the spotting differences HITs. For worker effort in the spotting differences HITs, the mean values of the two metrics, total time and longest interval, are displayed in Figure 5.2c and Figure 5.2d, respectively. Visually, we find that although the magnitude of financial incentives varies from 4 cents to 32 cents in an individual

Metrics	Original		Replicating	
	Task 1	Task 2	Task 1	Task 2
BC: # of clicks on target	0.82	0.40	0.52	0.31
SD: # of differences spotted	0.33	0.15	0.03	<0.001
SD: total time	0.37	0.29	0.26	0.09
SD: longest interval	0.41	0.24	0.18	0.30

Table 5.1: p-values of the Kruskal-Wallis one-way analysis of variance on base treatments. Left section: results for the original experiment (2012). Right section: results for the replicating experiment (2015-2016). p-values smaller than 0.1 are highlighted in bold.

task, workers exhibit no significant differences in either work quality or worker effort across different treatments, and such observations are valid for both the first task and the second task in the sequence. Interestingly, we notice that within each treatment, in the second task, workers clicked on the target buttons for fewer times in the button clicking HITs or identified more differences correctly in the spotting differences HITs, compared to their performance in the first task. Besides, both the total time and longest interval in the second task of the spotting differences HITs are shorter than that in the first task. These observations suggest that workers may become fatigued/bored or learn from their experiences when they are asked to complete tasks of the same type in a sequence.

Our intuition from visual inspection is confirmed by statistical tests. The test we use is the Kruskal-Wallis one-way analysis of variance (Kruskal-Wallis one-way ANOVA), which is a non-parametric test of the null hypothesis that multiple empirical samples come from the same distribution. The goal then is to examine whether the work quality and worker effort are statistically the same across 4 bonus levels in our base treatments. In particular, consider the tests on the work quality in the first task of the button clicking HITs as an example. For each of the four base treatments, we have a sample of 100 data points on the number of clicks on the target button, and we attempt to understand whether these four samples originate from the same distribution using Kruskal-Wallis one-way ANOVA. The p-values of the tests for both the first task and the second task are reported in Table 5.1 (left section).

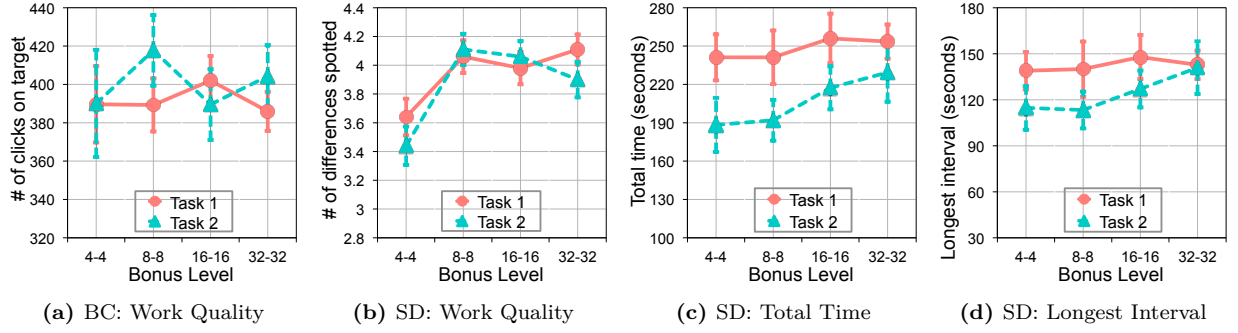


Figure 5.3: Work quality and worker effort in the four base treatments of the replicating experiment in 2015-2016.

As shown in the table, none of the differences in work quality or worker effort are statistically significant across the four base treatments, which implies that the magnitude of performance-contingent financial incentives alone affects neither work quality nor worker effort. This result is consistent with the previous findings on performance-independent financial incentives which claimed that the magnitude of financial incentives had no significant influences on the produced work quality [Mason and Watts, 2010, Rogstadius et al., 2011], suggesting that crowd workers may not adjust their performance according to the absolute magnitude of financial incentives, regardless of whether the incentives are contingent on the performance.

To understand the robustness of this result, we then move on to the replicating experiment to check that after three years, whether it is still true that the magnitude of performance-contingent financial incentives alone has no significant effect on either work quality or worker effort. Similar to our analysis for the original experiment, Figure 5.3 illustrates the mean values of work quality metrics and worker effort metrics across the four base treatments for both the button clicking HITs and the spotting differences HITs in the replicating experiment, and the p-values of the Kruskal-Wallis one-way ANOVA are reported in Table 5.1 (right section).

First, we notice that for the button clicking HITs, the conclusion we draw from the original experiment still holds for the replicating experiment. That is, the magnitude of

performance-contingent rewards alone has no significant effect on the work quality — across the four base treatments with different bonus levels, we observe no obvious pattern in the number of clicks on the target buttons for either the first task or the second task within the HIT, and such observation is also supported by the statistical test result (see the first row of Table 5.1 (right section)).

For the spotting differences HITs, however, our results in the replicating experiment is somewhat mixed. On the one hand, both the upward trend shown in Figure 5.3b and the statistical significant differences reported in the second row of Table 5.1 (right section) suggest that the magnitude of performance-contingent rewards actually has an impact on the work quality in the spotting differences HITs this time. A Tukey post-hoc pairwise comparison further indicates that the significantly low work quality in the 4–4 treatment contributes the most to the observed statistical significant differences in work quality across the four treatments: For the first task, the mean value of the number of correctly spotted differences in the 4–4 treatment is significantly smaller than that in the 8–8 treatment ($p < 0.05$) and the 32–32 treatment ($p < 0.05$); for the second task, the mean work quality in the 4–4 treatment is significantly lower than that in all other three treatments (4–4 vs. 8–8: $p < 0.001$, 4–4 vs. 16–16: $p < 0.01$; 4–4 vs. 32–32: $p < 0.05$); yet for both the first and the second task, there is no significant difference in work quality among the 8–8, 16–16 and 32–32 treatment. On the other hand, we also find that in most cases, how much effort a worker puts into a spotting difference task in the replicating experiment is still not quite affected by the magnitude of performance-contingent reward in that task (except that for the total amount of time a worker spends on the second spotting differences task, there is a marginally significant difference across the four treatments, $p = 0.09$), which is in line with our observations in the original experiment.

Taken together, from both the original experiment and the replicating experiment, we find that the magnitude of performance-contingent financial incentives alone does *not* necessarily

exert any significant influence on either work quality or worker effort. However, compared to our observations in the original experiment, the slightly different findings we have for the replicating experiment also suggest that the worker behavior in reaction to the magnitude of performance-contingent financial incentives alone can change over time. We provide two possible explanations to this observed change: (1) As time passes by, the MTurk worker population may have changed significantly hence workers who participated in our replicating experiment can be fundamentally different from those workers who participated in the original experiment in spite of all our efforts to control the comparability of the two experiments. In fact, it is estimated that half of the workers will be replaced in the MTurk worker pool in about every 7 months [Stewart et al., 2015]. (2) Even if there is no significant change in terms of the composition of worker population, a worker’s interpretation of the absolute magnitude of financial incentives can still be adjusted over time, and such adjustment can be influenced by a variety of external factors, such as the current inflation rate, the increase in federal minimum wage and the formation of the sense of “fair” payment among more MTurk workers. Finally, given that we have observed similar results for the button clicking tasks but different results for the spotting differences tasks in the original and the replicating experiment, we conjecture that how workers react to the absolute magnitude of financial incentives, as well as whether and how the effects of financial incentives on worker performance will change over time, is dependent on the nature of the task.

5.1.4 Improving the Effectiveness of Rewards in a Sequence

Our next goal is to understand how the work quality and worker effort is affected by the change of performance-contingent financial incentives in the subsequent two tasks in a sequence, or in other words, whether and how workers react to the relative magnitude of financial incentives in task sequences. Thus, unless otherwise specified, our analysis in the following is based on the *change* in work quality and worker effort from the first task to the

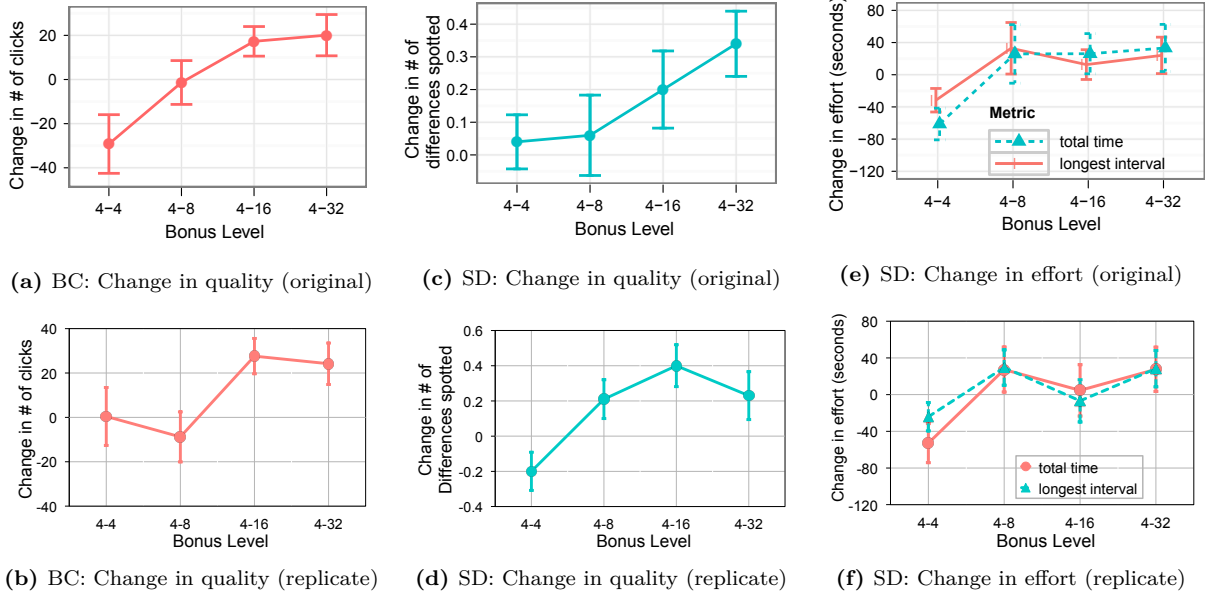


Figure 5.4: Changes in work quality and worker effort for treatments with increasing bonus level. Mean values and standard errors of the changes are plotted. Top row: results for the original experiment. Bottom row: results for the replicating experiment.

second task in a HIT. Specifically, for a metric of work quality or worker effort, the change in it for a HIT equals the value of the metric for the second task minus that for the first task in the HIT. Since samples of these changes do not visually deviate from normal distribution, one-way analysis of variance (one-way ANOVA) and two-sided t-tests, both assuming normal distribution of errors, are used in the subsequent statistical analysis.

We first examine how the work quality and worker effort in HITs changes with the *increase* of bonus levels in the sequence and the results are illustrated in Figure 5.4. In particular, Figures 5.4a, 5.4c, and 5.4e present the changes in work quality and worker effort for the 4 – 4, 4 – 8, 4 – 16, and 4 – 32 treatments in the original experiment. We see a clear upward trend for changes in both work quality and worker effort as the bonus level of the second task increases, except a slight dip for the change in longest interval in the 4 – 16 treatment of the spotting differences HITs. Figures 5.4b, 5.4d, and 5.4f show the similar results for the replicating experiment, suggesting that when the performance-contingent financial incentives

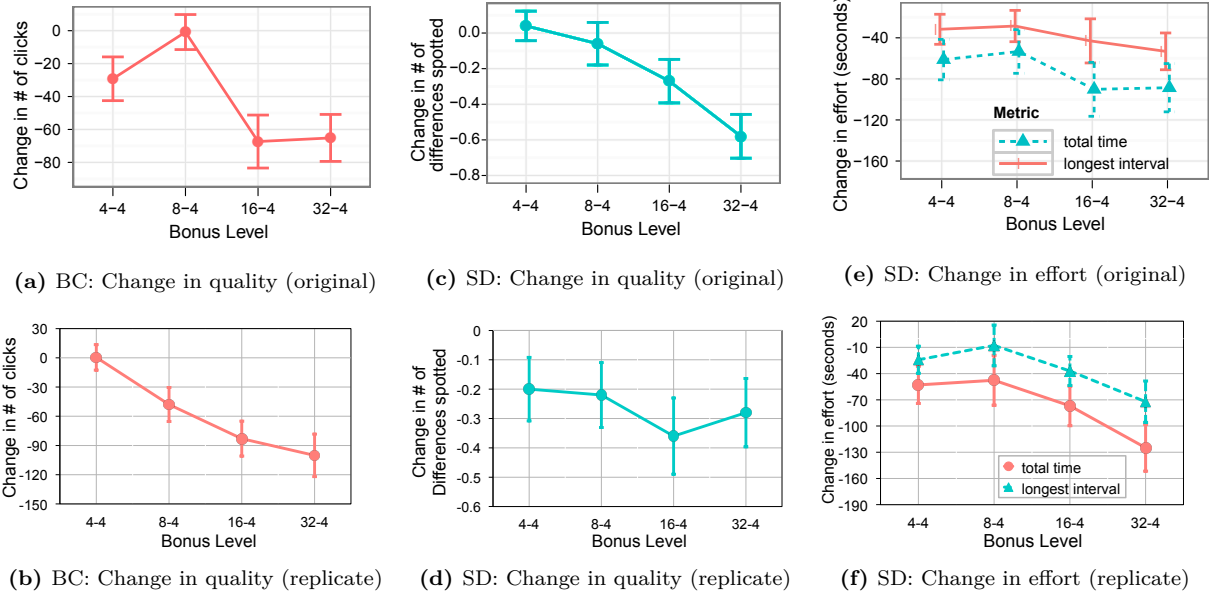


Figure 5.5: Changes in work quality and worker effort for treatments with decreasing bonus level. Mean values and standard errors of the changes are plotted. Top row: results for the original experiment. Bottom row: results for the replicating experiment.

increase in the task sequence, workers are likely to exert more efforts and improve their work quality in the tasks.

Similarly, we next look into how the work quality and worker effort in HITs changes with the *decrease* of bonus levels in the task sequence. As shown in Figure 5.5, for both the original experiment (Figures 5.5a, 5.5c, and 5.5e) and the replicating experiment (Figures 5.5b, 5.5d, and 5.5f), we find a downward trend for all metrics in both HITs with some occasional exceptions (e.g., the change in the number of clicks on target in the 8 – 4 treatment of the button clicking HITs in the original experiment). These results further indicate that when the performance-contingent financial incentives decrease in the task sequence, workers also tend to decrease their effort levels and complete tasks in lower quality.

All of these findings provide supporting evidence to a consistent conclusion, that is, workers in treatments with increasing or decreasing bonus levels *do* react to the relative changes of performance-contingent financial incentives, hence produce work of higher (or

Metrics	Original			Replicating		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	27.86*	46.50***	49.30***	-9.24	27.17 [†]	23.72
SD: change in # of differences spotted	0.02	0.16	0.30*	0.41**	0.60**	0.43*
SD: change in total time	87.09*	87.35**	94.52**	80.18*	57.51	80.45*
SD: change in longest interval	64.58*	44.29*	55.76*	53.68*	17.35	52.45*

(a) Base treatments vs. treatments with increasing bonus (X – X vs. X – Y)

Metrics	Original			Replicating		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	3.06	-52.15**	-64.46***	-76.39***	-70.56**	-118.46***
SD: change in # of differences spotted	-0.11	-0.37**	-0.76***	-0.27 [†]	-0.44**	-0.07
SD: change in total time	3.79	-68.66*	-58.62*	1.56	-38.24	-100.72**
SD: change in longest interval	21.15	-30.86	-42.87 [†]	18.99	-16.19	-70.29*

(b) Base treatments vs. treatments with decreasing bonus (Y – Y vs. Y – X)

Table 5.2: Differences of the mean values for the change in work quality and worker effort for pairs of treatments with the same bonus level in the first task. For a pairwise comparison, treatment A vs. treatment B, the reported value for a metric is the mean change of the metric in treatment B minus the mean change in the metric in treatment A. X is fixed to be 4. The statistical significance of the two-sided t-test is marked as a superscript, with [†], *, **, and *** representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

lower) quality and exert effort of higher (or lower) levels given increased (or decreased) amount of rewards in the second task of the HIT. With this conclusion in mind, it is natural to ask whether it is possible to improve the effectiveness of financial incentives by carefully designing the magnitude of the rewards in a task sequence.

More specifically, we are interested in answering this question from two perspectives: First, for two treatments starting with the same amount of reward in the first task, will workers “recover” their sensitivity to different magnitude of financial incentives in the second task as the bonus level in the second task differs (e.g., produce higher work quality if the reward magnitude in the second task is larger)? Second, for two treatments that change from different bonus levels in the first task to the same bonus level in the second task, will workers perceive the same amount of financial incentives in the second task differently?

The pairwise comparisons shown in Table 5.2a and Table 5.2b present the answer to the first question. Specifically, let X and Y be two bonus levels, $X = 4$, and $X < Y$. Table 5.2a presents the pairwise comparisons of treatment $X - X$ and treatment $X - Y$, and Y varies from 8, to 16, to 32. The values reported in the table are the differences in the mean values of the change in the work quality or worker effort metric between the two treatments in the pair. In particular, consider the comparison on the change in the number of clicks on the target in the button clicking HITs as an example. When $Y = 8$, the value reported in the table is the mean value of the change in the number of clicks on target in the $4 - 8$ treatment minus the mean value of that change in the $4 - 4$ treatment. The statistical significance of two-sided t-tests is noted as a superscript. Similarly, Table 5.2b compares the change in work quality or worker effort metrics between treatment $Y - Y$ and the treatment $Y - X$, with Y varying from 8, to 16, to 32.

Thus, each of the pairs in Table 5.2a compares a base treatment with another treatment which starts from the same bonus level as that in the base treatment but increase to a higher bonus level later. If workers can actually recover their sensitivity to the magnitude of performance-contingent financial incentives due to the change of rewards in subsequent tasks, we expect to see positive numbers in this table, which is indeed the case for all but one comparison in both the original experiment and the replicating experiment. Notably, the improvement in work quality and worker effort also appear to be statistically significant for the majority of the comparisons, and for the only exception case (i.e., the change in work quality for the $4 - 8$ treatment is slightly smaller than that for the $4 - 4$ treatment in button clicking HITs of the replicating experiment), the difference is also not statistically significant. Likewise, each of the pairs in Table 5.2b compares a base treatment to another treatment with the same reward in the first task but a decreased bonus in the second task. If workers become more sensitive to the incentive magnitude in the second task of the sequence, we expect to see negative numbers in this table, which is again mostly true with some exceptions.

Metrics	Original			Replicating		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	2.54	32.48*	20.77 [†]	-37.27*	40.03*	5.89
SD: change in # of differences spotted	0.01	0.10	0.16	0.16	0.32*	0.44*
SD: change in total time	82.97*	49.38 [†]	63.27 [†]	76.46*	43.29	51.64
SD: change in longest interval	82.56*	24.75	34.39*	56.17*	13.99	30.11

(a) Base treatments vs. treatments with increasing bonus (Y – Y vs. X – Y)

Metrics	Original			Replicating		
	Y = 8	Y = 16	Y = 32	Y = 8	Y = 16	Y = 32
BC: change in # of clicks on target	28.38	-38.13*	-35.93*	-48.36*	-83.42**	-100.63***
SD: change in # of differences spotted	-0.10	-0.31*	-0.62***	-0.02	-0.16	-0.08
SD: change in total time	7.91	-30.69	-27.37	5.28	-24.02	-71.92*
SD: change in longest interval	3.17	-11.31	-21.50	16.51	-12.83	-47.95 [†]

(b) Base treatments vs. treatments with decreasing bonus (X – X vs. Y – X)

Table 5.3: Differences of the mean values for the change in work quality and worker effort for pairs of treatments with the same bonus level in the second task. For a pairwise comparison, treatment A vs. treatment B, the reported value for a metric is the mean change of the metric in treatment B minus the mean change in the metric in treatment A. X is fixed to be 4. The statistical significance of the two-sided t-test is marked as a superscript, with [†], *, **, and *** representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

None of the positive differences are statistically significant though. Besides, the decrease in work quality and worker effort also seems to be statistically more significant for larger bonus decreases. Combined together, these results provide a confirmative answer to our first question — given the same initial level of incentives in the first task, the worker become more sensitive to the magnitude of the performance-contingent financial incentive in the second task and can then respond to an increased (or decreased) bonus level with a better (or worse) performance.

Analogically, Table 5.3a and Table 5.3b report comparisons between treatments which start from different bonus levels but end up at the same bonus level in the second task, and thus provide insights for our second question above. Specifically, Table 5.3a compares the base treatment Y – Y to treatment X – Y with increasing bonus, while Table 5.3b compares the base treatment X – X to treatment Y – X with decreasing bonus. Almost all differences in

Table 5.3a are positive, suggesting that even though the absolute magnitude of the bonus in the second task is the same for the two treatments in the pair, when that bonus is increased from a lower bonus level in the first task, workers tend to exhibit higher levels of work quality and worker effort. In contrast, the majority of the numbers in Table 5.3b are negative, implying that when observing the bonus to decrease from a high level to a low level in a task sequence, workers may perform even worse than that in the case where the bonus level is always low. Therefore, results in Tables 5.3a and 5.3b further demonstrate that workers may interpret the performance-contingent financial incentives differently when the reward magnitude changes in a sequence, even if the absolute magnitude of the performance-contingent financial incentives is actually the same.

Finally, to reaffirm that the absolute magnitude of the reward for the second task does not affect work quality and worker effort on the task, but the change of the magnitude of the reward from the first task to the second task does, we fit all of the data in one experiment into the following linear model, and repeat this process for both the original experiment and the replicating experiment:

$$M_{i,2} = C + \alpha \cdot M_{i,1} + \beta \cdot \text{Bonus}_{i,2} + \gamma \cdot \Delta\text{Bonus}_i + \epsilon_i, \quad (5.1)$$

where M is one of the metrics of work quality or worker effort (i.e., number of clicks on the target button for a button clicking task, number of differences correctly spotted for a spotting differences task, total time for a spotting differences task, or longest interval for a spotting differences task), $M_{i,1}$ and $M_{i,2}$ are worker i 's value of this metric on the first and second tasks respectively, $\text{Bonus}_{i,2}$ is the bonus level of the second task in this HIT, and ΔBonus_i is the change of the bonus level from the first task to the second task, that is, $\Delta\text{Bonus}_i = \text{Bonus}_{i,2} - \text{Bonus}_{i,1}$. Note that here we consider the value of a metric, rather than the change in the value of a metric. We include $M_{i,1}$ in the model to account for a

Metric M	C	$M_{i,1}$	$\text{Bonus}_{i,2}$	ΔBonus_i
# of clicks on target in BC task	48.2 ^{**} (14.8)	0.81 ^{***} (0.04)	0.23 (0.48)	1.66 ^{***} (0.36)
# of differences spotted in SD task	1.79 ^{***} (0.14)	0.55 ^{***} (0.03)	0.004 (0.004)	0.01 ^{***} (0.003)
Total time in SD task	168.0 ^{***} (14.8)	0.18 ^{***} (0.03)	0.48 (0.82)	1.87 ^{**} (0.63)
Longest interval in SD task	110.3 ^{***} (11.1)	0.11 ^{**} (0.03)	0.16 (0.66)	1.15 [*] (0.50)

(a) Regression results for data from the original experiment

Metric M	C	$M_{i,1}$	$\text{Bonus}_{i,2}$	ΔBonus_i
# of clicks on target in BC task	-24.2 [†] (14.6)	1.01 ^{***} (0.03)	0.49 (0.57)	2.39 ^{***} (0.44)
# of differences spotted in SD task	2.04 ^{***} (0.13)	0.48 ^{***} (0.03)	0.000 (0.004)	0.01 ^{**} (0.003)
Total time in SD task	152.3 ^{***} (14.4)	0.24 ^{***} (0.03)	0.40 (0.78)	1.52 [*] (0.61)
Longest interval in SD task	107.4 ^{***} (10.5)	0.14 ^{***} (0.03)	0.34 (0.61)	0.78 [†] (0.47)

(b) Regression results for data from the replicating experiment

Table 5.4: Regression results for linear model (5.1). Estimated coefficients and standard errors are reported. The statistical significance is marked as a superscript, with [†], ^{*}, ^{**}, and ^{***} representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

worker’s innate capability on the task. The regression results for all four metrics in the original experiment and the replicating experiment are shown in Table 5.4a and Table 5.4b, respectively. As none of the coefficients for $\text{Bonus}_{i,2}$ is statistically significant while all coefficients for ΔBonus_i are statistically significant for all models, it is clear that the bonus level of the second task does not affect either work quality or worker effort, but the change of the bonus level affects both.

5.2 Providing Monetary Interventions in Task Switching

In the previous section, we have experimentally studied how workers react to performance-contingent financial incentives in a sequence of tasks of the same type. In this section, we study whether and how performance-contingent financial incentives can be used to influence worker performance in a sequence of tasks of different types, that is, in a task switching setting. In particular, we consider the case that performance-contingent financial rewards are placed on a few *selected* tasks in a working session, and these rewards are referred to as *monetary interventions*. The tasks where monetary interventions are placed are called the *intervened tasks*, while other tasks in the session are *non-intervened tasks*. We ask three questions regarding the effects of monetary interventions in task switching: (1) How do monetary interventions affect work quality in intervened tasks? (2) How do monetary interventions affect work quality in non-intervened tasks? (3) Where should monetary interventions be placed to improve work quality in a more effective way — should monetary interventions be placed on switch tasks (i.e., tasks that follow another task of the same type) or repetition tasks (i.e., tasks that follow another task of a different type), and in a working session with high task switching frequency or low task switching frequency?

We conducted a between-subject experiment on MTurk to answer these questions. Each worker was randomly assigned to an experimental condition, in which she was asked to complete a sequence of 96 tasks with two types of tasks interleaving with each other in the sequence. Experimental conditions varied in either the task switching frequency or whether and where monetary interventions were used in the sequence. The effects of monetary interventions were then analyzed by comparing work quality across experimental conditions.

5.2.1 Related Work

Task switching is closely related to a few other concepts, including multitasking, task interruption and resumption. Multitasking refers to either performing two or more types of tasks simultaneously or switching back and forth from one type to another [Salvucci and Taatgen, 2010, Salvucci et al., 2009]. Our setting in this study is similar to the latter form of multitasking. Meanwhile, many studies in the human-computer interaction community explored task switching from the perspective of task interruption and resumption. In these studies, a subject was typically performing a primary task before being interrupted by a secondary task, and the effects of the interruption on the *primary* task were analyzed [Iqbal and Horvitz, 2007, Mark et al., 2008, Bailey and Konstan, 2006]. Unlike such work, in this study, we care about the work quality in *all* types of tasks rather than focusing on a single (primary) type of tasks.

A prominent psychological effect was consistently observed in previous studies regarding the worker performance in task switching—workers usually have worse performance on switch tasks than on repetition tasks [Rogers and Monsell, 1995, Monsell, 2003]. The performance difference between the switch and repetition tasks is called the *switch cost*, which is likely to be a result of the costly cognitive control processes triggered by the task switching (e.g., shift of attention, retrieval of task goals and rules into working memory, etc.) or task-set inertia, that is, the proactive interference between the competing old and new tasks (e.g., persistent activation of the old task and the involuntary inhabitation of the current task) [Mayr and Kliegl, 2000, Allport et al., 1994, Kiesel et al., 2010]. It is also known that more frequent task switching demands more cognitive resources, which may be mentally taxing or cause information overload for workers [Speier et al., 1999]. In contrast, repetition tasks offer opportunities for workers to develop task-specific skills and strategies over time as a result of *learning* and *task specialization* and thus may lead to increased work quality.

The emphasis of this study is on understanding the effects of performance-contingent

financial incentives on the worker performance in task switching. Most prior work on the relationship between financial incentives and the worker performance in the on-demand work setting, including our own work described in Section 5.1, is based on experimental studies, in which workers either complete a task only once or complete a sequence of tasks of the *same* type. To the best of our knowledge, how workers react to financial incentives when they complete a sequence of tasks of *different* types was only studied in the labs. It was observed that if workers could earn additional rewards based on their *overall* performance in a working session of tasks of mixed types, their performance on switch tasks was improved marginally [Nieuwenhuis and Monsell, 2002]. Different from this research, our work considers an alternative way to provide monetary rewards in task switching settings—we place performance-contingent financial incentives on *selected* tasks in a working session; thus, workers can earn additional rewards on the intervened tasks as long as their performance in these tasks meets some pre-specified criteria, yet workers will not be able to earn additional rewards on the non-intervened tasks regardless of how well they perform in those tasks.

Informing a worker that performance-contingent bonuses will be provided on some selected tasks, however, could possibly set an implicit performance goal for the worker, which may further affect her performance on all tasks. There is a large literature on *explicit goal setting* which demonstrates that setting specific and challenging goals often leads to better performance [Locke et al., 1981, Mento et al., 1987, Locke and Latham, 2002]. Furthermore, when the explicit goals are combined with monetary incentives, the worker performance can be further improved [Locke et al., 1981, Pritchard and Curts, 1973]. It is thus interesting to examine whether the implicit goals potentially conveyed by monetary interventions have a similar effect as the explicitly stated goals. If they do, we expect that monetary interventions affect worker performance on not only intervened tasks, but also non-intervened tasks.

5.2.2 Experimental Design

Our experimental design is inspired by two classical task switching experimental paradigms: *predictable task switching*, where switches happen in a predictable way after a constant number of tasks in a sequence, and *task cuing*, where an explicit cue is presented before each task to specify the type of the current task [Kiesel et al., 2010].

Tasks. Two types of tasks are used in our experiments: the *color naming* task and the *word reading* task. In a task of either type, a worker will see a stimuli word on the screen, which is the name of one of the five colors — blue, green, magenta, red and yellow. The word is displayed in a color that may or may not match the word, and the color is also limited to the previous five options. For example, a stimuli word “red” can be written in blue. The two types of tasks are:

- *The color naming task* (Color): A worker is asked to indicate the color in which the word is written, regardless of whether or not that matches the word itself. In the above example, the answer is “blue.”
- *The word reading task* (Word): A worker is asked to indicate what the word denotes, regardless of the color it is written. In the above example, the answer is “red.”

In each task, the worker is instructed to report the answer by typing the initial of it in lower case. For example, the worker can report the answer “red” by typing ‘r’ on the keyboard.

Worker performance in each task is measured in two dimensions:

- *Reaction time* (RT): The elapsed time between the onset of the stimuli and the worker’s response.
- *Accuracy* (or correctness): A binary value indicating whether the reported answer is correct or not.

These two metrics innately *compete* with each other as when workers shorten their reaction

time, they are likely to be less accurate, *ceteris paribus*.


The two types of tasks were initially used in the Stroop test, which revealed the Stroop effect, that is, subjects generally spend more time on naming the colors than reading the words [Stroop, 1935]. They are now widely used by psychologists in studying task switching [Wylie and Allport, 2000, Gilbert and Shallice, 2002, Allport and Wylie, 1999].





Task Sequences. In our experiment, we put 96 tasks, which include 48 tasks of each type, in a human intelligence task (HIT). For different task sequences, the two types of tasks switch at different frequencies.





Specifically, we define a “segment” in a sequence as a consecutive chunk of tasks of the same type and the length of a segment is the number of tasks in it. Thus, for our experiment, if the length of each segment in a task sequence is N , there are $M = 96/N$ segments in that sequence, and the sequence is then referred to as an $N \times M$ sequence. Different types of tasks are assigned to neighboring segments in a sequence. By varying segment lengths, we can control the task switching frequency. We considered five task sequences in our study: 4×24 , 8×12 , 16×6 , 24×4 and 48×2 .

Intervention Treatments. Each worker is asked to complete one of the five task sequences and receives a performance-independent payment of 3 cents for each task completed. Monetary interventions are performance-contingent monetary rewards: a worker can earn an extra bonus of 2 cents on a task with monetary intervention if her reported answer for that task is correct *and* her reaction time is less than 1 second. By varying whether and where the additional bonuses are placed in a sequence, we create three treatments for each of the five task sequences:

- *No Bonus (baseline)*: No bonus is placed on any task in a task sequence.

	No Bonus:	Red	Green	...	Blue	Yellow	Red	...	Blue	Red	Magenta	...	Magenta	Green	Blue	...	Red
Task #	1	2	...	24	25	26	...	48	49	50	...	72	73	74	...	96	
	-	R-NI	...	R-NI	S-NI	R-NI	...	R-NI	S-NI	R-NI	...	R-NI	S-NI	R-NI	...	R-NI	

	Switch Bonus:	Red	Green	...	Blue		Yellow	Red	...	Blue		Red	Magenta	...	Magenta		Green	Blue	...	Red
Task #	1	2	...	24	25	26	...	48	49	50	...	72	73	74	...	96				
	-	R-NI	...	R-NI	S-I	R-NI	...	R-NI	S-I	R-NI	...	R-NI	S-I	R-NI	...	R-NI				

	Repetition Bonus:	Red	Green	...	Blue		Yellow	Red	...	Blue		Red	Magenta	...	Magenta		Green	Blue	...	Red
Task #	1	2	...	24	25	26	...	48	49	50	...	72	73	74	...	96				
	-	R-NI	...	R-NI	S-NI	R-I	...	R-NI	S-NI	R-NI	...	R-I	S-NI	R-NI	...	R-NI				

←

Color

→

←

Word

→

←

Color

→

←

Word

→

Figure 5.6: An illustration of three treatments for the 24×4 sequence. S-NI denotes a switch task without monetary intervention, R-NI represents a repetition task without monetary intervention, S-I refers to a switch task with monetary intervention, and R-I is a repetition task with monetary intervention. The first task of a sequence is neither a switch nor a repetition task.

- *Switch Bonus:* Starting from the second segment in a task sequence, a performance-contingent bonus is offered at the first task in every segment, i.e., bonuses are placed at all switch tasks.
- *Repetition Bonus:* Starting from the second segment in a task sequence, a performance-contingent bonus is offered at a randomly selected non-switch task in every segment, i.e., a bonus is placed at one random repetition task in each segment (except for the first segment).

Figure 5.6 gives a graphical example of the three treatments.

We call a combination of a task sequence and an intervention treatment an *experimental condition*. Thus, there are 15 experimental conditions in our experiment.

Procedure. We posted our HITs on MTurk on weekdays around 12:00-14:00 and 16:00-18:00 (Eastern Standard Time) in a week in March 2014. To avoid network latency as well as cultural differences in perceiving financial incentives, we restricted our HITs to U.S. workers.

We suggested workers who have difficulties in seeing colors or perceiving color differences not to take the HIT. Using a desktop or laptop computer with a keyboard to complete the HIT was recommended. Each worker was limited to take the HIT once.

Upon arrival, a worker is randomly assigned to an experimental condition. The worker then goes through an instruction page, a task and interface tutorial and a qualification test. In the tutorial, the worker is instructed to report the answer to each task as quickly and accurately as possible. If she is assigned to a Switch Bonus or Repetition Bonus treatment, she is also informed of the opportunities to earn extra bonuses at some tasks in the sequence, contingent on her answer in those tasks being correct and reported within 1 second. The worker can only proceed to the actual task sequence after passing the qualification test.

The actual task sequence starts with a task of a random type. For each task in the sequence, the worker will first see a cue word (i.e., either “Color” or “Word”), shown in white on the gray background, which indicates whether the current task is the color naming or the word reading task. For the Switch Bonus and Repetition Bonus treatments, a bonus icon is displayed together with the cue word if monetary intervention is placed on the current task. Each cue is displayed for two seconds and then the worker is automatically redirected to the task page, where a stimuli word is displayed. Both the word and the printing color of the stimuli are randomly chosen from the five alternatives. The type of the current task is displayed again on the top of the task page in case of unawareness. Once the worker reports her answer to the current task, she will be automatically redirected to the cue page for the next task. Finally, after completing all 96 tasks, the worker is asked to complete a post-task survey of demographic information.

Each worker in our experiment got a show-up fee of \$0.20 and a performance-independent payment of \$2.88 ($\0.03×96) after submitting the HIT. Workers in Switch Bonus and Repetition Bonus treatments may get extra bonuses depending on their performance in those tasks where monetary interventions were placed.

Data. We recruited 1305 workers in total from MTurk for our experiment. For each worker, we recorded: (1) the task type and whether there was a monetary intervention for each task that the worker worked on in the sequence; (2) the worker’s reaction time for each task; and (3) the worker’s accuracy for each task.

We noticed that it took some workers an excessively long time to report their answers to some tasks, which might be due to interruptions in their working environment. To eliminate the influences of these “outliers,” we excluded the data from a worker if her reaction time for any of the tasks in her sequence was longer than 20 seconds. Such elimination left us with 1268 valid workers. The data for these workers were then used in the subsequent analysis.

The average age of the valid workers is 30.8, 59.1% of them are male, and all of them use either a desktop or a laptop computer to complete the HITs. No significant demographic or equipment difference is observed for workers in different experimental conditions.

5.2.3 Effects on Intervened Tasks

Our first goal is to understand whether introducing monetary interventions in a task switching setting can incentivize workers to improve their performance on tasks where the interventions are placed. We thus focus on comparing worker performance on *intervened tasks* in treatments with bonuses (i.e., the Switch Bonus and Repetition Bonus treatments) with worker performance on the corresponding tasks in the baseline treatment (i.e., the No Bonus treatment). In the following, the Wilcoxon rank sum test is used to evaluate statistical significance unless otherwise stated.

We first analyze worker performance in terms of reaction time. To get a sense of on average, how fast workers react to the stimuli word in each task when there are no monetary interventions, five *baseline average reaction time sequences* are created using the data from the No Bonus treatment, one for each of the five task sequences. That is, we consider the five experimental conditions that combines each of the five task sequences with the No Bonus

treatment; for each of these experimental conditions, we take all workers who were assigned to that condition and average their reaction times position-wise. For example, the value at position i in the baseline average reaction time sequence for the 4×24 sequence is obtained by averaging the reaction time for the i -th task across all workers who were assigned to the experimental condition that combines the 4×24 sequence with the No Bonus treatment. We denote the baseline average reaction time sequence for task sequence s as RT_s , where $s \in \{4 \times 24, 8 \times 12, 16 \times 6, 24 \times 4, 48 \times 2\}$, and $RT_s(i)$ refers to the value at position i in RT_s , $i \in [1, 96]$.

Worker reaction time in intervened tasks for treatments with monetary interventions is then compared to that in the corresponding position of the baseline average reaction time sequence. For example, consider the comparison between the No Bonus treatment and the Switch Bonus treatment. For any given task sequence s , we use two buckets: the first bucket collects all intervened, switch task reaction times for all workers in the experimental condition that combines the Switch Bonus treatment with the task sequence s ; and for each reaction time value that we add to the first bucket, suppose it comes from position x in the task sequence, we will then put $RT_s(x)$ to the second bucket. We then calculate the average value for both buckets. Figure 5.7a plots the differences of the average reaction time on the intervened tasks between the No Bonus treatment and the Switch Bonus treatment for all five task sequences. The differences in reaction time between the No Bonus treatment and the Repetition Bonus treatment are calculated similarly and plotted in Figure 5.7b. As the figures suggest, the presence of monetary interventions leads to shorter reaction time for the intervened tasks, no matter where the interventions are placed. Further statistical tests report $p < 0.001$ for pairwise comparisons for all task sequences, indicating that the decreases are significant.

We then examine worker performance in terms of accuracy. Similar to the analysis on reaction time, for any given task sequence s , we first create the *baseline average accuracy*

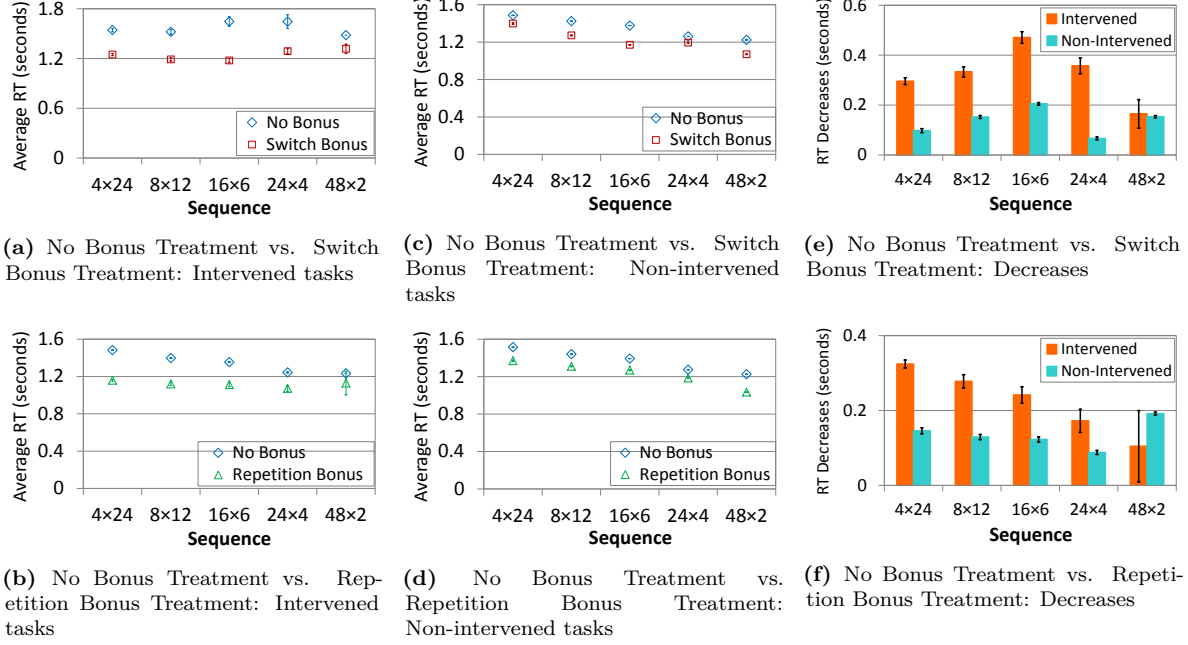


Figure 5.7: Effects of monetary interventions on reaction time for intervened tasks and non-intervened tasks. Error bars represent standard errors of the mean.

sequences, ACC_s , by taking all workers who worked on the task sequence s in the No Bonus treatment and averaging their accuracy position-wise. $ACC_s(i)$ represents the value at position i in ACC_s . Then, for each worker who worked on s in the Switch Bonus (or Repetition Bonus) treatment, we put her accuracy in each task into one of the three categories depending on whether that task appears before, at or after the placement of the monetary intervention in its segment. Furthermore, for each accuracy value for a task at position x that we put into one of the three categories for the Switch Bonus (or Repetition Bonus) treatment, we also add $ACC_s(x)$ to the same category for the No Bonus treatment. Finally, by taking the average of all data in each category, we can see in each treatment how accurate workers are before, at the time or after monetary interventions being placed in the segments and thus investigate whether worker's accuracy improves in the intervened tasks with the extra bonuses.

Figures 5.8a and 5.8b report how worker's accuracy changes within a segment for different

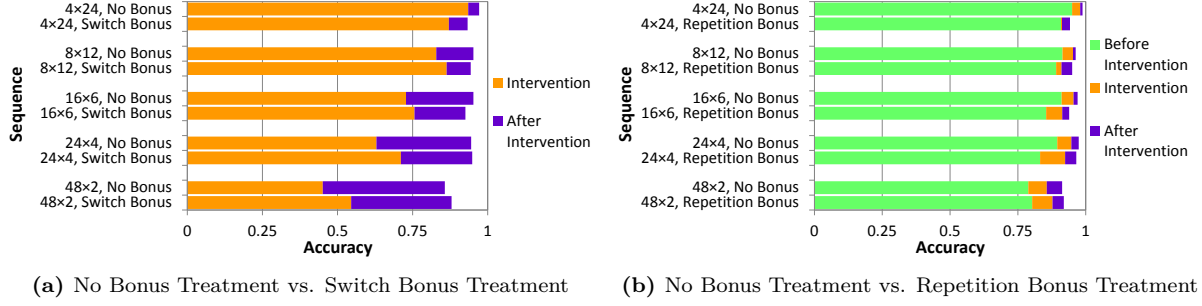


Figure 5.8: Comparison of the average worker accuracy within a segment.

treatments, with Figure 5.8a showing the comparison between the No Bonus treatment and the Switch Bonus treatment, and Figure 5.8b showing the comparison between the No Bonus treatment and the Repetition Bonus treatment. Accuracy is plotted cumulatively: for example, in Figure 5.8b, the average worker accuracy after monetary interventions is the sum of the average accuracy before interventions (green bar), the accuracy increment at intervened tasks (orange bar) and the accuracy increment after interventions (purple bar).

Figure 5.8a shows that in general, the orange bar for the Switch Bonus treatment is longer than that for the corresponding No Bonus treatment, with the 4×24 sequence being the only exception. This indicates that the average accuracy at the switch tasks improves significantly ($p < 0.001$) when monetary interventions are placed on these tasks in all but the 4×24 sequence. The exception of the 4×24 sequence may be resulted from workers being overwhelmed by the mentally-taxing frequent switches and thus find the additional bonuses more disturbing rather than motivating.

When monetary interventions are placed on repetition tasks, the average accuracy at these tasks is often not higher than that in the corresponding No Bonus treatment. To see this, we compare the combined length of the green and orange bars in Figure 5.8b for the Repetition Bonus and No Bonus treatments. The combined length for the Repetition Bonus treatment is shorter than that for the corresponding No Bonus treatment ($p < 0.001$), except for the 48×2 sequence. However, the lower average accuracy at intervened tasks for the

Repetition Bonus treatment can be largely attributed to the low average accuracy in the non-intervened tasks before the intervention, i.e., the green bar is shorter in the Repetition Bonus treatment than that in the corresponding No Bonus treatment for most sequences. This is due to faster reaction on non-intervened tasks when extra bonuses are used and the competition between reaction time and accuracy, which we will detail in the next section. When focusing on the accuracy improvement at intervened tasks and thus comparing the length of orange bars between the two treatments in Figure 5.8b, we find that with monetary interventions, the accuracy improvement at the intervened repetition tasks is significantly larger for sequences with moderate to low task switching frequencies ($p < 0.05$).

To summarize, introducing monetary interventions incentivizes better performance on intervened tasks — workers complete the intervened tasks not only faster but also with either higher accuracy or a larger accuracy improvement. Recall that to earn the bonuses workers need to both react quickly and be accurate. While it may be easy for a worker to submit a response faster, the improved performance in accuracy suggests that workers are indeed motivated by the extrinsic financial incentives to improve her performance along *both* dimensions. To some degree, the performance-contingent financial incentives even lead workers to overcome the innate tradeoff between the two performance metrics on the intervened tasks. Our observations, therefore, imply the effectiveness of performance-contingent rewards in mitigating switch cost (when placed on switch tasks) or promoting faster learning and task specialization (when placed on repetition tasks).

5.2.4 Effects on Non-intervened Tasks

We next attempt to understand the effects of monetary interventions on non-intervened tasks.

The comparisons of the average reaction time for non-intervened tasks in the Switch Bonus treatment and the Repetition Bonus treatment against that in the baseline No Bonus

treatment are displayed in Figure 5.7c and Figure 5.7d respectively. Interestingly, we find that although workers cannot earn extra rewards by completing the non-intervened tasks quickly, they still show a clear tendency in shortening their reaction time significantly ($p < 0.001$) for these tasks. On the other hand, while workers are still very accurate, their accuracy decreases at the non-intervened tasks: for the Switch Bonus treatment, the average worker accuracy for non-intervened tasks is 93.75% across all task sequences, which is slightly lower (by 0.74%) than that for the No Bonus treatment; and for the Repetition Bonus treatment, the average worker accuracy for non-intervened tasks across all task sequences is 91.34%, which is 2.03% lower than that for the No Bonus treatment. The accuracy decreases for non-intervened tasks are statistically significant ($p < 0.001$). These results indicate that with the additional bonuses, workers try to improve their performance in reaction time while maintaining their performance in accuracy even when monetary rewards are not directly applied to the tasks. Yet, the competitive nature of the two performance metrics seems to still dominate in the non-intervened tasks, which means that faster reaction comes with a cost in accuracy for these tasks.

As monetary interventions lead to decreases in reaction time for both intervened and non-intervened tasks, we further compare the magnitude of the decrease between these two categories of tasks. Results are reported in Figures 5.7e and 5.7f. It is clear that no matter where the monetary rewards are placed, the decrease in reaction time for intervened tasks is significantly larger ($p < 0.05$) than that for non-intervened tasks, with the 48×2 sequence in the Repetition Bonus treatment being the only exception (the decrease in reaction time for non-intervened tasks there is marginally larger, with $p = 0.077$).

We provide a unified explanation for our observations on worker performance in both intervened and non-intervened tasks: workers treat the performance-contingency of extra rewards on some selected tasks as an implicit performance goal, which has a similar effect as an explicit goal. Thus, workers attempt to improve their performance for *all* tasks in the

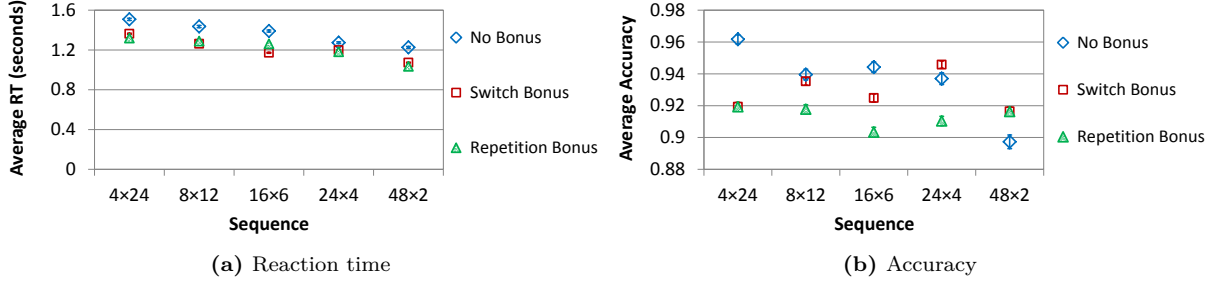


Figure 5.9: Comparison of average worker performance for all 96 tasks in a sequence across different treatments. Error bars represent standard errors of the mean.

sequence (subject to the innate tradeoff between the two performance metrics), regardless of whether monetary interventions are placed on the tasks. For the intervened tasks in the sequence, workers are further incentivized by the extrinsic financial incentives and therefore improve their performance in those tasks to a larger degree, yielding both a faster reaction and a higher accuracy (or a larger accuracy increment).

5.2.5 More Effective Interventions

Finally, we seek to gain some insights into how to more effectively use monetary interventions in a task switching setting to improve worker performance.

Figure 5.9a and Table 5.5 report the comparisons of the three treatments in terms of the average worker reaction time over all 96 tasks for each of the five task sequences. Figure 5.9b and Table 5.6 present similar comparisons for worker accuracy.

First, we look into the baseline No Bonus treatment. As shown in Figures 5.9a and 5.9b, when monetary interventions are not available, as task switching becomes less frequent, worker reaction time gets shorter and worker accuracy also exhibits a downward trend. One-way analysis of variance (ANOVA) further confirms that the differences in reaction time and accuracy across task sequences are statistically significant ($p < 0.001$). In other words, without monetary interventions, by controlling how frequently tasks switch in a sequence, a requester may trade off better performance in average reaction time for better performance in overall

Task Sequences	Reaction Time Mean Values			Reaction Time Differences		
	No Bonus (NB)	Switch Bonus (SB)	Repetition Bonus (RB)	SB – NB	RB – NB	RB – SB
4×24	1.5078	1.3628	1.3193	-0.145 ^{***}	-0.188 ^{***}	-0.044 ^{***}
8×12	1.4355	1.2625	1.2893	-0.173 ^{***}	-0.146 ^{***}	0.027
16×6	1.3904	1.1717	1.2615	-0.218 ^{***}	-0.129 ^{***}	0.090 ^{***}
24×4	1.2731	1.1976	1.1824	-0.075 ^{***}	-0.091 ^{***}	-0.015
48×2	1.2261	1.0725	1.0352	-0.154	-0.191 ^{***}	-0.037 ^{***}

Table 5.5: Average worker reaction time in different experimental conditions and differences of reaction time between conditions. The statistical significance of the Wilcoxon rank sum test is marked as a superscript, with ^{*}, ^{**}, and ^{***} representing significance levels of 0.05, 0.01, and 0.001 respectively. (Unit: seconds)

Task Sequences	Accuracy Mean Values			Accuracy Differences		
	No Bonus (NB)	Switch Bonus (SB)	Repetition Bonus (RB)	SB – NB	RB – NB	RB – SB
4×24	0.9617	0.9193	0.9193	-0.042 ^{***}	-0.042 ^{***}	0.000
8×12	0.9395	0.9353	0.9178	-0.004 ^{***}	-0.022 ^{***}	-0.018 ^{***}
16×6	0.9442	0.9248	0.9034	-0.019 ^{***}	-0.041 ^{***}	-0.021 ^{***}
24×4	0.9370	0.9458	0.9104	0.009 [*]	-0.027 ^{***}	-0.036 ^{***}
48×2	0.8973	0.9164	0.9163	0.019 ^{***}	0.019 ^{***}	-0.000

Table 5.6: Average worker accuracy in different experimental conditions and differences of accuracy between conditions. The statistical significance of the Wilcoxon rank sum test is marked as a superscript, with ^{*}, ^{**}, and ^{***} representing significance levels of 0.05, 0.01, and 0.001 respectively.

accuracy.

When comparing the worker performance in treatments with monetary interventions (i.e., the Switch Bonus or Repetition Bonus treatment) with that in the baseline treatment, we have an interesting observation: while workers can be incentivized to improve their performance in reaction time significantly regardless of the task switching frequency in the sequences (i.e., negative differences in columns “SB – NB” and “RB – NB” of Table 5.5), similar improvement in accuracy can only be achieved when the task switching frequency is low (i.e., positive differences in columns “SB – NB” and “RB – NB” of Table 5.6 only appear in sequences with a low task switching frequency). This observation implies that adding monetary interventions

to sequences with a low task switching frequency could be more effective: instead of trading off speed for accuracy or vice versa, workers perform better according to both metrics; in particular the incentives boost worker’s overall accuracy in the sequence significantly.

In terms of where to place monetary interventions in a task sequence to improve the effectiveness of the extra bonuses, we next compare the worker performance between the Switch Bonus treatment and the Repetition Bonus treatment. We find that while both treatments can effectively improve worker performance in reaction time, placing the performance-contingent rewards on switch tasks generally leads to better performance in accuracy compared to providing extra bonuses at repetition tasks (i.e., differences in the “RB – SB” column of Table 5.6 are mostly negative). This indicates that it is more efficient to use monetary interventions right at the the switching points. With a closer look, this finding can further be attributed to two reasons: first, accuracy improvement at the intervened tasks is significantly larger when bonuses are placed on switch tasks than when they are placed on repetition tasks (+3.49% vs. +0.34%, $p < 0.001$); second, combining extra bonuses with task switches makes workers shift focus to the new type of tasks quicker when tasks switch — compared to the baseline treatment, the average number of tasks it takes for a worker to first submit a correct answer in a segment is decreased by 0.18 (not significant) for the Switch Bonus treatment while increased by 0.15 ($p < 0.05$) for the Repetition Bonus treatment, leading workers in the Switch Bonus treatment to outperform workers in the Repetition Bonus treatment in the early stage of each task segment (e.g., the average accuracy comparison is 88.52% vs. 85.30% for the first half of tasks in each segment, $p < 0.001$).

In sum, monetary interventions can be most effective in motivating better worker performance when they are placed at switch tasks in a sequence with a low task switching frequency. We conjecture that this is due to that monetary interventions are less interruptive in sequences with a low task switching frequency, and the demand for extra attention, which can be stimulated using monetary interventions, is at its height on switch tasks.

5.3 Discussion

We conducted two experimental studies to empirically understand the effects of financial incentives in the on-demand economy. Our results showed that when workers work on a sequence of tasks of the same type, the magnitude of performance-contingent financial incentives alone does not necessarily affect either work quality or worker effort, yet the change of incentive magnitude in subsequent tasks has been consistently observed to significantly affect both of them—increasing the magnitude of incentives helps to improve the work quality and worker effort while decreasing the magnitude hurts. In addition, we found that in a sequence of tasks of different types, occasionally providing performance-contingent financial incentives on some selected tasks not only leads to an improved performance in the intervened tasks, but also cast a spillover effect on the non-intervened tasks. It is also demonstrated that such monetary interventions are most effective in eliciting better worker performance when used on the switch tasks in a sequence with a low task switching frequency.

Both studies provided implications for better designing financial incentives in the on-demand economy by understanding and leveraging the psychological processes of the workers. For example, our findings in the first study generally supported the conjecture of [Mason and Watts](#) on why simply using financial incentives of larger magnitude didn't lead to an improved performance [[Mason and Watts, 2010](#)]*—*they hypothesized the existence of an “*anchoring effect*,” that is, a worker may anchored her conception of “appropriate” payment level based on the actual compensation she received in a task and consistently felt herself being underpaid, which mitigated her motivation to perform better. Indeed, combining their hypothesis with the fair wage-effort hypothesis [[Akerlof and Yellen, 1990](#)], we provide one plausible interpretation for our experimental results in the first study: (1) Because a worker always tends to use the payment in the first task that she encounters in a sequence as a reference point to form a sense of the “fair” amount of payment in her mind, she is likely to feel equally

underpaid no matter how large the magnitude of the performance-contingent reward is and thus exhibits similar performance across the base treatments; (2) With the anchored reference point of fair payment, a worker can interpret the magnitude of the performance-contingent financial incentives in the second task by comparing it to the reference point—when it is higher than the reference point, the worker responds to the “fairer” payment by improving her performance, otherwise the worker decreases her performance. Notice that this interpretation is also in line with the prospect theory [[Kahneman and Tversky, 1979](#)], which claims that people make decisions based on the perceived gains and losses against the reference point rather than the absolute values. In practice, this interpretation suggests that we may leverage the anchoring effect in the task workflow design to improve the effectiveness of financial incentives. As for our findings in the second study, the impact of monetary interventions on worker performance for both intervened tasks and non-intervened tasks lead us to conjecture that workers may in fact set implicit performance goals for themselves when they notice that in some tasks, a part of the monetary rewards is dependent on their performance. It is thus straightforward to further explore how theories in goal setting can inspire us to improve the effectiveness of financial incentives.

While in this chapter, we focus on studying the effects of performance-contingent financial incentives, in reality, we may not be able to use such incentives in some cases. For example, the quality of work for a task may be subjective, not verifiable, or is too costly to be practical to verify. This requires us to have a better understanding of other forms of financial incentives in the on-demand economy. One of the alternative form of financial incentives is the combination of various peer prediction methods [[Miller et al., 2005](#), [Prelec, 2004](#)] with financial rewards, which rewards a worker based on not only her answer but also the answers of her peers. It will be an interesting future direction to study the effects of this type of financial incentives both theoretically and empirically.

5.4 Acknowledgements

This chapter involved collaboration with Yiling Chen and Yu-An Sun. Portions of this chapter previously appeared in the AAAI publication *The Effects of Performance-Contingent Financial Incentives in Online Labor Markets* [Yin et al., 2013] and the HCOMP publication *Monetary Interventions in Crowdsourcing Task Switching* [Yin et al., 2014]. We thank the support of the National Science Foundation under grant CCF-1301976 and the Xerox Foundation on this work.

Chapter 6

Monetary Intervention Design: An Algorithmic Perspective

In the last chapter, we have experimentally studied the effects of financial incentives in on-demand work settings. While these studies provide empirical evidence in support of the effectiveness of using financial incentives to affect worker performance, it remains unclear how can we use financial incentives in a working session (i.e., a sequence of tasks) in an *efficient* way.

For example, consider a requester who plans to place performance-contingent rewards on some selected tasks in a working session to encourage high-quality work. From the requester's perspective, it is not necessarily always beneficial to provide such rewards, as the potentially improved quality also comes with an increase in financial cost. Furthermore, even if providing such rewards is indeed beneficial, the requester still face a series of subsequent decisions such as how many tasks and which tasks should rewards be placed on. Our findings in the last chapter clearly suggest that these decisions are crucial in determining the effectiveness of the financial incentives. Yet, the current common practice among requesters is still to follow some simple fixed or random schemes to offer a performance-contingent reward on *none* of the

tasks, *all* of the tasks or a number of *randomly selected* tasks in a working session, and each worker is also awarded in the same way. Therefore, it is straightforward to ask whether we can design a more efficient way to place monetary interventions by, for example, dynamically adjusting the placement of performance-contingent rewards in a working session.

In this chapter, we provide an initial answer to this question by presenting an algorithmic approach to control financial incentives, which helps the requester to make decisions on whether and when to offer performance-contingent financial rewards (e.g., bonuses) in a working session to maximize the overall utility he derives from the session. In particular, the goal of algorithmically controlling the provision of bonuses in working sessions naturally comes down to two specific problems: First, how can we quantitatively characterize the impact of monetary interventions on work quality? Second, how should we trade off quality against cost?

To address the first problem, we consider to learn statistical models of worker’s reaction to monetary interventions from the empirical data. For example, given a particular type of task at hand, suppose a requester gets access to a set of historic data on worker behavior in working sessions of such task. That is, the requester has the record on whether monetary intervention is provided as well as the work quality on each task of the working sessions for a group of past workers. We can then use this historic dataset to train a statistical model, which characterizes how the quality of on-demand work changes with the provision of monetary interventions. Such model is especially important for a requester to reason about how to provide monetary interventions in an efficient way, because it can be utilized to *predict* work quality for future workers in their working sessions, for both the cases when monetary intervention is placed on a task or not placed. Specifically, for a new worker who starts to work on a task session, after monitoring her performance for a short period of time, the requester can use the learned model to make a prediction on the worker’s performance in a particular task, given whether monetary intervention is placed on this task, as well

as the history of monetary intervention provisions and work quality for all previous tasks that the worker has already completed in the session. As work quality is usually measured with discrete levels (e.g., high-quality or low-quality) and is influenced by the provision of external monetary interventions, this prediction problem is essentially a *categorical time series prediction with exogenous inputs*.

To address the second problem of trading off quality against cost, we propose to augment the learned model with a requester utility function and therefore turn the bonus placement problem into a problem of utility maximization under uncertainty. Depending on the specific model that is used, the uncertainty may originate from the stochastic nature of the model or the varying confidence levels when using the model to predict work quality. We then provide different algorithms to solve for a utility-maximizing, dynamic policy of bonus placement.

In the following, we first describe related work in Section 6.1. In Section 6.2, we present a series of seven models from three categories (supervised learning models, autoregressive models and Markov models) to characterize the impact of monetary interventions on crowd work quality. Using how well they can predict work quality under monetary interventions as the evaluation standard, we also conduct an empirical comparison of these models. As an example, in Section 6.3, we demonstrate our algorithmic approach of dynamic incentive control based on a particular type of Markov model, the first-order input-output hidden Markov model. Through randomized experiments on Amazon Mechanical Turk, we show that our approach significantly improves the requester utility compared to the baseline fixed or random bonus schemes. To the best of our knowledge, this is the first time that the effectiveness of algorithmically-controlled bonus schemes is shown with *real* crowd workers. Section 6.4 discusses possible extensions and future directions.

6.1 Related Work

The problem of modeling worker performance in working sessions in the on-demand, crowd work environment has been explored in a few previous studies. Early work often focuses on estimating a worker’s inherent capability level (sometimes referred to as the error rate) which is *independent* of the working environment, does *not* change over time and determines worker performance in the tasks [Whitehill et al., 2009, Karger et al., 2011, Raykar and Yu, 2012]. Recent work, however, suggests that worker performance can be better modeled when taking its time-variance (e.g., improvement or degradation over time) into consideration [Donmez et al., 2010, Jung et al., 2014, Bragg and Weld, 2016]. In particular, Donmez et al. [2010] proposed a Bayesian time series model, assuming that the latent variable dynamics that governs the change of work quality over time has a uniform offset and correlation, that is $x_t = x_{t-1} + \epsilon_t$. Jung et al. [2014] relaxed this constraint and came up with a generalized model (LAR) with $x_t = c + \phi x_{t-1} + \epsilon_t$. More recently, Jung and Lease [2015a] designed a generalized time-varying assessor model (GAM) that is a logistic regression predictor with features extracted from both generative time-series models (e.g., the estimated ϕ and c from the LAR model) and worker’s behavioral evidence, and they showed that the prediction accuracy on crowd work quality can be significantly improved with this model. Meanwhile, Bragg and Weld [2016] took a different approach and used a parametric hidden Markov model to explicitly model the performance degradation over time.

The time-variance of crowd work quality discussed in all the above studies describes the *organic evolvement* of worker performance, perhaps due to the learning effect or boredom. In reality, however, worker performance can also be influenced by some external factors presented in the working contexts, such as the monetary interventions embedded in the task session. This naturally leads to the interesting questions of how to explicitly take the impact of monetary interventions into consideration for modeling crowd work quality, and furthermore,

how to reward workers in an optimal way such that the requester utility is maximized. To this end, some researchers have explored possible payment strategies based on specific hypotheses or theoretical models on how work quality changes with monetary interventions. For instance, Wang and Ipeirotis [2013] proposed a “payment with reimbursement” scheme based on the conjecture that the fluctuation of payments in task sequences is undesirable as workers may interpret a decrease of financial incentives as a punishment. Ho et al. [2014] used the classical principal-agent model to characterize how a worker makes strategic decisions when provided with a performance-contingent payment and studied how to adaptively adjust such payment over time.

Different from these studies, in this chapter, we present an algorithmic approach to reward workers based on statistical models of worker behavior that are learned from *empirical data*. These models are used to predict crowd work quality under monetary interventions, hence may further provide guidance to the requester for making decisions on bonus placements. More specifically, we adopt some existing models from the literature such as supervised learning models and variants of Markov models. We also propose a few new models. In particular, not many time-series models can be directly applied to the categorical prediction problems with exogenous input sequence like ours — models like discrete autoregressive (DAR) and latent autoregressive (LAR) deal with categorical time series predictions *without* exogenous inputs [Jacobs and Lewis, 1983, Jung et al., 2014], while models like autoregressive with exogenous inputs (ARX) deal with predictions with exogenous inputs for *continuous* variables [Ljung, 1998]. Hence, we propose two variants of autoregressive models, DARX and LARX, for our prediction problem, which are extended from the existing models. We further study the prediction performance of these models in more realistic scenarios, such as when the requester has limited training dataset or limited ground truth. Similar analyses have been conducted previously in different contexts, for the prediction of disengagement [Mao et al., 2013] or predicting temporal work quality without external interventions [Jung and

Lease, 2015b].

Once the models are learned from the empirical data, we combine them with reasoning techniques and therefore make decisions on the placement of bonuses in working sessions. Similar approaches have been used for optimizing the decision making in the on-demand work for different purposes, such as dynamically controlling worker recruitment and testing [Kamar et al., 2012, Bragg and Weld, 2016], task assignment [Dai et al., 2010] and workflow switches [Lin et al., 2012]. Besides, Huang et al. [2010] also worked on optimally designing on-demand work variables, such as how many HITs to post and how many tasks to be bundled in one HIT, following a similar strategy.

Finally, our work is different from several pricing mechanisms proposed to elicit more (or faster) work from rational workers given a fixed budget [Singer and Mittal, 2013, Singla and Krause, 2013, Gao and Parameswaran, 2014]—our goal is to elicit *high-quality* work, and we have no assumption on the rationality of workers.

6.2 Predicting Work Quality under Monetary Interventions

In this section, to address the problem of predicting crowd work quality under monetary interventions, we enumerate 7 models from 3 different categories, and present an empirical comparison on the prediction performance of these models. Specifically, we first treat our prediction as a classification problem and adopt three *supervised learning models* (random forests, support vector machine and artificial neural network). Furthermore, we propose two time-series models (DARX and LARX) that are extended from existing *autoregressive models* to incorporate the exogenous inputs. Finally, by assuming that the change of work quality (or the change of some latent variable related to work quality) is governed by a Markov process, we consider two variants of the *Markov models* (controlled Markov chain and input-output

hidden Markov model) for our prediction. The performance of each model is examined on three datasets that are collected with *real* crowd workers from Amazon Mechanical Turk for different types of tasks, including solving word puzzles, classifying images, and finding typos in the text.

In addition, requesters often face some practical constraints when predicting crowd work quality: (1) *the “cold start” problem*: requesters have very limited training data to start with, hence their knowledge on how workers react to monetary interventions is quite limited at the beginning; (2) *the lack of ground truth*: requesters often get access to the ground truth for only a certain number of tasks, hence they can only evaluate a worker’s performance on *some* tasks in the past when making prediction on her work quality in the current task. Therefore, to better understand the robustness of the models when facing realistic constraints, we conduct further experiments to investigate the performance of different prediction models when the requester has limited training data or limited ground truth.

6.2.1 Prediction Models

Our prediction problem can be formally defined as the following: The requester has collected a training dataset \mathbb{D}_{train} of N workers. Each worker in the training dataset completes a sequence of T tasks. For each worker i ($1 \leq i \leq N$), the requester keeps a record of the sequence of monetary interventions provided to the worker $\mathbf{a}_i = (a_i^1, a_i^2, \dots, a_i^T)$ as well as the sequence of observed work quality $\mathbf{y}_i = (y_i^1, y_i^2, \dots, y_i^T)$. For simplicity, we consider binary levels of monetary interventions and work quality in this study. That is, $a_i^t \in \{0, 1\}$ ($1 \leq t \leq T$) indicates whether a monetary intervention is provided on task t to worker i , with value 1 (or 0) representing a positive (or negative) answer, and $y_i^t \in \{0, 1\}$ refers to the work quality of worker i on task t , with value 1 (or 0) representing high-quality (or low-quality) work. The requester is interested in modeling crowd work quality under monetary interventions through the training dataset and making predictions for a future

worker — given the sequence of monetary interventions $\mathbf{a} = (a^1, a^2, \dots, a^{l-1})$ provided to this worker so far as well as the observed work quality $\mathbf{y} = (y^1, y^2, \dots, y^{l-1})$, what's the worker's performance y^l in the current task (i.e., the l -th task) when monetary intervention level a^l is provided?

Supervised Learning Models

We first treat our prediction as a supervised learning problem. Take worker i 's performance in task t for an example, y_i^t is naturally the *label* for this training instance. We further extract a *feature set* for this instance by focusing on a history window of size L . In particular, the feature set \mathbf{x}_i^t includes:

- *current intervention level*: a_i^t , whether a monetary intervention is provided on the current task;
- *average intervention level*: $\frac{1}{t-1} \sum_{j=1}^{t-1} a_i^j$, the percentage of tasks with monetary interventions among all previous tasks;
- *average performance*: $\frac{1}{t-1} \sum_{j=1}^{t-1} y_i^j$, the percentage of high-quality work in all previous tasks;
- *historical intervention levels*: $a_i^h(t-L \leq h \leq t-1)$, whether monetary intervention is provided on each of the previous L tasks;
- *historical performance*: $y_i^h(t-L \leq h \leq t-1)$, the work quality in each of the previous L tasks;
- *historical intervention changes*: $a_i^{h_2} - a_i^{h_1}(t-L \leq h_1 < h_2 \leq t-1)$, the differences in monetary interventions for any two of the previous L tasks; and
- *historical performance changes*: $y_i^{h_2} - y_i^{h_1}(t-L \leq h_1 < h_2 \leq t-1)$, the differences in work quality for any two of the previous L tasks.

A transformed training dataset is created through extracting the feature-label pair (\mathbf{x}_i^t, y_i^t)

for all workers and all tasks in the original training dataset. A supervised learning model then simply constructs a function $y_i^t = f(\mathbf{x}_i^t)$ that maps the features to the label. We consider three such model in this study:

Model 1: Random Forests (RF). Random forests [Ho, 1998] is a popular ensemble learning technique for classification and regression. Briefly speaking, many decision trees are grown in the random forests. Each tree is constructed by fitting a decision tree to a *random* subset of the training data, and a *random* subset of features are considered for each split within the tree. The prediction for a testing sample is made by classifying it using each decision tree in the forest in turn and then taking the majority vote among all trees.

Model 2: Support Vector Machine (SVM). The general idea of support vector machine [Cortes and Vapnik, 1995] is to map training data points from the original finite-dimensional space to a higher-dimensional space, and search for a *hyperplane* to separate data points from different classes such that the distance between the closest two data points of different classes is maximized. *Kernel functions* are often used to construct non-linear SVM classifiers [Boser et al., 1992]. To make a prediction for a testing sample, we simply map it to the same higher-dimensional space and assign a label to it according to on which side of the hyperplane it falls.

Model 3: Artificial Neural Network (NN). Inspired by the biological neural networks, artificial neural networks are a family of machine learning models that can approximate any function between features and labels [Hornik et al., 1989]. While various network structure can be designed based on the understanding of the specific prediction problem, in this study, we focus on a fully connected multi-layer neural network — in this network, there is a layer of *input* neurons, a layer of the single *output* neuron and one or more layers of *hidden* neurons where each neuron in one hidden layer is connected to *all* neurons in the previous (input or

hidden) layer as well as *all* neurons in the next (hidden or output) layer. Specifically for our problem, each element in the feature set \mathbf{x}_i^t activates an input neuron and the single output neuron produces the label y_i^t . A neuron in a hidden layer takes the weighted sum of output values from the previous layer as the input, and outputs a value after transforming the input with an *activation function*. The weights between any two neurons in the network are estimated through the training data. The prediction of a testing sample can be completed by feeding the input neurons with its features, activating hidden neurons in turn and determining the label until the output neuron is activated.

Autoregressive Models

Next, we introduce two variants of the autoregressive models in time series analysis to address our prediction problem.

Model 4: Discrete Autoregressive Model with Exogenous Inputs (DARX). We extend the Discrete Autoregressive (DAR) model [Jacobs and Lewis, 1983] to incorporate the exogenous inputs. Formally, a DARX model of order p is defined as follows:

$$y_i^t = I_t y_i^{t-D_t} + (1 - I_t) e_t \quad (6.1)$$

where e_t is a binary variable with $Pr(e_t = 1 | a_i^t) = \beta_{a_i^t}$, I_t is a binary variable with $Pr(I_t = 1 | a_i^t) = \lambda_{a_i^t}$, D_t randomly takes a value from the set $\{1, 2, \dots, p\}$ with $Pr(D_t = d | a_i^t) = \alpha_{a_i^t}^d$, and $\sum_{d=1}^p \alpha_{a_i^t}^d = 1$ for $a_i^t \in \{0, 1\}$. Importantly, notice that in the DARX(p) model, the probability distributions for random variables e_t , I_t and D_t are all *conditioned* on the exogenous input a_i^t . This is different from the DAR model where exogenous inputs are not included as a part. As a concrete example, consider when monetary intervention is provided to worker i on task t , that is, $a_i^t = 1$. Then, the DARX(p) model states that, the value of y_i^t (i.e., whether worker i will submit high-quality work in task t) is related to the previously

observed work quality with probability λ_1 (i.e., when $I_t = 1$) and not related with probability $1 - \lambda_1$ (i.e., when $I_t = 0$). When $I_t = 0$, y_i^t is determined by an independent binary variable e_t , which is equal to 1 with probability β_1 . On the other hand, when $I_t = 1$, y_i^t equals to the observation d ($1 \leq d \leq p$) steps ago, that is, y_i^{t-d} , with probability α_1^d .

The DARX(p) model has $2p + 4$ parameters to estimate in total: λ_a , α_a^d and β_a , with $a \in \{0, 1\}$ and $d \in \{1, 2, \dots, p\}$. Given the training dataset, we can search for a set of parameters that best characterizes worker's reaction to monetary interventions as a *population*. To make a prediction for a testing worker with these population-level parameters on her l -th task, we simply draw random variables e_l , I_l and D_l according to the estimated parameters and decide the label of the testing sample with Equation 6.1.

On the other hand, parameters of a DARX(p) model can also be estimated in an online fashion for the *individual* worker that we are currently predicting on. This enables us to make more personalized predictions — we may initialize the model with the population-level parameters, that is, $\lambda_a^1 = \lambda_a$, $\alpha_a^{d,1} = \alpha_a^d$ and $\beta_a^1 = \beta_a$, and we can update these parameters over time as we keep observing the testing worker completes more tasks in the session and obtaining the individual-level model estimates. One way to update the model parameters is to take a weighted average of the old parameters and the newly estimated individual-level parameters at each time step. For instance, suppose the testing worker has completed a sequence of $l - 1$ tasks and the observed sequences of \mathbf{a} and \mathbf{y} lead to an individual-level model with parameters λ'_a , $\alpha_a^{d'}$ and β'_a . We propose to update the model parameters as the following:

$$\lambda_a^l = (1 - \gamma)\lambda_a^{l-1} + \gamma\lambda'_a \quad (6.2)$$

$$\alpha_a^{d,l} = \frac{(1 - \gamma)\lambda_a^{l-1}\alpha_a^{d,l-1} + \gamma\lambda'_a\alpha_a^{d'}}{\lambda_a^l} \quad (6.3)$$

$$\beta_a^l = \frac{(1 - \gamma)(1 - \lambda_a^{l-1})\beta_a^{l-1} + \gamma(1 - \lambda'_a)\beta'_a}{1 - \lambda_a^l} \quad (6.4)$$

The prediction for the l -th task is made based on λ_a^l , $\alpha_a^{d,l}$ and β_a^l , and a new set of individual-level parameters will be estimated after we observe the actual work quality y^l in the l -th task. Notice that γ ($0 \leq \gamma \leq 1$) represents the *learning rate* for parameter updating: when $\gamma = 0$, the prediction is always made with population-level parameters, and when $\gamma = 1$, the prediction is made with individual-level parameters exclusively.

Model 5: Latent Autoregressive Model with Exogenous Inputs (LARX). The second autoregressive model variant is extended from the Latent Autoregressive Model (LAR) [Jung et al., 2014]. Specifically, the LAR model is defined as follows:

$$z_i^t = c + \phi z_i^{t-1} + \epsilon_i^t \quad (6.5)$$

$$Pr(y_i^t = 1) = \frac{1}{1 + e^{-z_i^t}} \quad (6.6)$$

where $\epsilon_i^t \sim N(0, \sigma^2)$ is a random noise, z_i^t is a latent variable that governs the worker's performance, and the observed work quality y_i^t is determined stochastically by z_i^t through the logistic function. To take the impact of monetary interventions on work quality into consideration, we propose a generalized LARX model, with an autoregressive order of p and an exogenous input order of q , by replacing Equation 6.5 with the following formula:

$$z_i^t = c + \sum_{j=1}^p \phi_j z_i^{t-j} + \sum_{j=0}^{q-1} \theta_j a_i^{t-j} + \epsilon_i^t \quad (6.7)$$

Equation 6.7 is essentially an autoregressive model with exogenous inputs (ARX) [Ljung, 1998]. Different from the LAR model, the LARX model assumes that the latent variable z_i^t depends linearly on both its previous values and the exogenous inputs. Given the training dataset, a population-level LARX model can be learned through expectation-maximization algorithms with particle filters [Park et al., 2014]. While the population-level model can be used for prediction, similar to the DARX model, we can also make more personalized

predictions by updating the LARX model parameters over time (e.g., $\phi_j^l = (1 - \gamma)\phi_j^{l-1} + \gamma\phi_j'$) to characterize both the population-level behavior and the individual-level behavior.

Markov Models

Finally, we present two Markov models for predicting crowd work quality under monetary interventions.

Model 6: Controlled Markov Chain (CMC). Controlled Markov chain includes exogenous inputs (often referred to as “actions”) into a Markov chain, and with further addition of reward functions, a CMC will be transformed into a Markov decision process (MDP). A CMC of order p defines that state transition depends only on the recent p states and the current input, that is, $P_a(S_p, \dots, S_1, S_0) = Pr(s_t = S_0 | s_{t-1} = S_1, \dots, s_{t-p} = S_p, a_t = a) = Pr(s_t = S_0 | s_{t-1} = S_1, \dots, s_1 = S_{t-1}, a_t = a)$. For our purpose, we take the observed work quality in each task as the “state.” Thus, the state transition probabilities essentially represent the distribution of the work quality y_i^t in task t , given the monetary intervention level a_i^t in task t and the observed work quality sequence $(y_i^{t-p}, y_i^{t-p+1}, \dots, y_i^{t-1})$ in the past p tasks. A maximum-likelihood estimate of these transition probability parameters can be obtained given the training dataset. For the testing worker, we predict that $Pr(y^l = 1) = P_{a^l}(y^{l-p}, y^{l-p+1}, \dots, y^{l-1}, 1)$.

Model 7: Input-Output Hidden Markov Model (IOHMM). Input-output hidden Markov model [Bengio and Frasconi, 1995] is a variant of the hidden Markov model for learning the mapping between input and output sequences. An IOHMM of order p is defined as follows:

- *inputs:* a_i^t , whether a monetary intervention is provided in task t ;
- *outputs:* y_i^t , the work quality in task t ;

- *hidden states*: $z_i^t \in \{1, 2, \dots, K\}$, the worker's latent state in task t , where K is the total number of hidden states;
- *transition probability*: $P_{tr}(z_i^t | z_i^{t-1}, \dots, z_i^{t-p}, a_i^t)$, the probability of transiting to state z_i^t in task t given the current input a_i^t and state sequence $(z_i^{t-p}, z_i^{t-p+1}, \dots, z_i^{t-1})$ in the previous p tasks; and
- *emission probability*: $P_e(y_i^t | z_i^t, \dots, z_i^{t-p+1}, a_i^t)$, the probability of submitting work of quality y_i^t in task t given the current input a_i^t and the state sequence $(z_i^{t-p+1}, \dots, z_i^{t-1}, z_i^t)$ in the recent p tasks.

An IOHMM can be estimated using the Baum-Welch expectation-maximization algorithm [Bengio and Frasconi, 1996]. To make predictions for the testing worker, we maintain and update a state belief \mathbf{b}_l ($1 \leq l \leq L$) at each step, which is the probability distribution for the worker to stay in different combinations of states in the p tasks before task l . The value of y^l is then computed with \mathbf{b}_l and a^l . For example, when $p = 1$, we have $\mathbf{b}_l = (b_l(1), b_l(2), \dots, b_l(K))$ where $b_l(k)$ ($1 \leq k \leq K$) is the estimated probability for the worker to stay in hidden state k in task $l - 1$. Then, we predict that $Pr(y^l = 1) = \sum_{k=1}^K b_l(k) (\sum_{j=1}^K P_{tr}(j|k, a^l) P_e(1|j, a^l))$, and after we observe y^l , the state belief is updated according to that $b_{l+1}(j) \propto \sum_{k=1}^K b_l(k) P_{tr}(j|k, a^l) P_e(y^l|j, a^l)$.

6.2.2 Evaluation Datasets

To examine the performance of different prediction models, we collected 3 datasets from real crowd workers on Amazon Mechanical Turk:

- **PUZZLE**: consists of 300 workers each completing a sequence of 9 word puzzle tasks in one HIT. In each task, the worker is shown a 12×12 board filled with capital letters and a “target” word on the screen. This target word can be placed on the board horizontally, vertically or diagonally, and for multiple times. The worker is asked to find

the appearances of the target word on the board as many times as possible. The base payment for the HIT is 45 cents. The requester provides extra performance-contingent bonus on 37% of the tasks. When a worker submits a high-quality answer in a bonus task by pointing out more than 80% of all appearances of the target word, she can earn an extra bonus of 5 cents¹.

- **CLASSIFY**: consists of 220 workers each completing a sequence of 10 butterfly classification tasks in one HIT. In each task, the worker sees 5 pictures of butterflies and is asked to classify each picture into three categories of interests: black swallowtail, monarch and machaon. Example pictures of butterflies in each category are provided to workers in the instruction, and workers are further encouraged to search online to better understand the key features for different butterflies. The base payment for the HIT is 50 cents. 29% of the tasks come with extra bonus. When the worker submits a high-quality answer in a bonus task by correctly classifying all 5 pictures in that task, she can earn an extra bonus of 5 cents. The set of butterfly pictures used in the tasks was taken from [Lazebnik et al., 2004].
- **TYP0**: consists of 80 workers each completing a sequence of 10 typo-finding tasks in one HIT. In each task, there is a short paragraph of about 200 words. The worker is asked to proofread it and find out as many typos as possible. The base payment for the HIT is 1 dollar. In 49% of all the tasks, there are extra performance-contingent bonuses. If the worker submits a high-quality answer in a bonus task by finding out more than 75% of all the typos, she will earn a bonus of 10 cents. A similar task was used in a previous study to understand the effects of performance-contingent rewards [Ho et al., 2015].

¹By design, each board in our task contains the target word 11 times (workers are not aware of this fact, however), which means that a worker can earn the extra reward in a bonus task if she identifies the target word for at least 9 times.

6.2.3 An Empirical Comparison of Model Performance

We now report our empirical comparison results on the performance of different models in predicting the crowd work quality under monetary interventions.

Experimental Settings

Given a dataset, we first randomly take 80% of the workers in it and collect their data as the training dataset, while the data for the rest 20% of the workers is used as the testing dataset. For a particular model type (e.g., random forests), we fit a model of that type using the training dataset, and then use the estimated model to make predictions for each worker in the testing dataset. Since predicting the work quality in one task often rely on information about previous tasks, we start making predictions from the fourth task of each sequence. This process is repeated for 20 times, and the average performance of each prediction model across the 20 random splits is then reported.

Baselines. For comparison, in our experiment, we include two baseline models that consider the organic evolvement of worker performance only:

- *running accuracy (RA)*: $Pr(y^l = 1) = \frac{1}{l-1} \sum_{j=1}^{l-1} y^j$, that is, the prediction on the l -th task is made according to the percentage of high-quality work observed in the previous $l - 1$ tasks; and
- *latent autoregressive (LAR)*: the time-series model proposed by [Jung et al. \[2014\]](#)².

Metrics. We use 3 metrics to evaluate the performance of a prediction model:

- *accuracy*: the percentage of tasks in the testing dataset for which the prediction is correct;

²Although GAM is proposed as an improvement of LAR in [\[Jung and Lease, 2015a\]](#), we can not use GAM as a baseline because GAM is tailored to their specific dataset.

- F_1 score: the harmonic mean of precision and recall, i.e., $F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$; and
- \log loss: $-\frac{1}{N_{test}} \sum_{j=1}^{N_{test}} y_j \log(p_j) + (1 - y_j) \log(1 - p_j)$, where N_{test} is the total number of predictions made for the testing dataset, y_j is the true label of the j -th sample, and $p_j = Pr(y_j = 1)$ is the predicted probability of high-quality work for the j -th sample.

Intuitively, the higher the accuracy, the better the model. As some of our datasets are imbalanced³, we provide the F_1 score for further reference. For prediction models that generate probabilistic labels (e.g., RA, LAR, DARX, LARX, CMC and IOHMM), in order to calculate accuracy and F_1 score, we assign a binary label to a testing sample according to a predefined threshold of 0.5, that is, $\hat{y}_j = 1$ when $p_j > 0.5$. Log-loss describes not only whether the prediction is accurate but also whether the prediction is confident, with a smaller value indicating a better model.

Model Selection. Model selection is conducted through cross validation. Specifically, we partition the training dataset into 5 folds, pick each of the five folds to test while using the rest four folds to train models. The model setting with the highest average performance across the five folds (according to log loss) is then selected and a final model is trained with the whole training dataset using this setting.

We fix the size of history window $L = 3$ for all supervised learning models. For RF, we fix the number of trees to be 1,000 and tune on the minimum number of samples on a leaf; for SVM, we tune on the choice of kernel function (e.g., linear, polynomial, radial basis function, sigmoid); and for NN, we tune on the choice of activation function (e.g., logistic sigmoid, hyperbolic tan, rectified linear), the number of hidden layers (1 or 2) and the number of neurons in each hidden layer. For autoregressive models, we experiment with different learning rates $\gamma \in \{0, 0.01, 0.05, 0.1, 1\}$ for both DARX and LARX. While we also tune on

³The percentages of tasks with high-quality work in the PUZZLE, CLASSIFY and TYPO datasets are 76.8%, 55.5% and 63.4%, respectively.

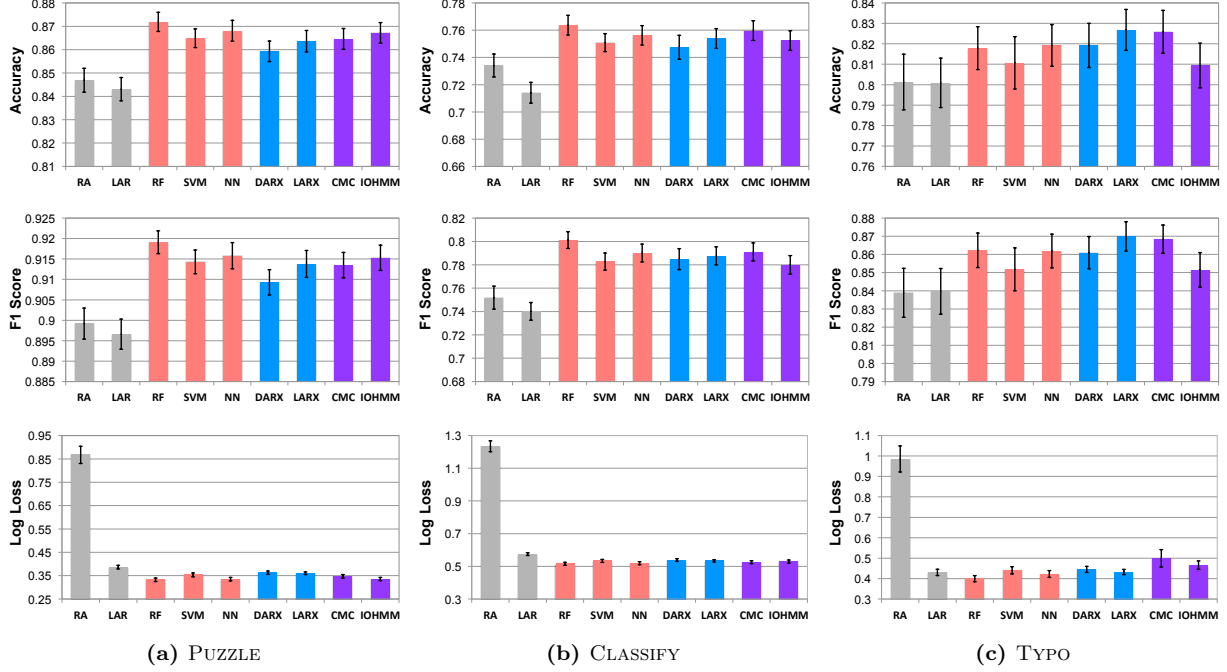


Figure 6.1: Performance comparisons for all prediction models on the three datasets (Top row: accuracy; middle row: F₁ score; bottom row: log loss). Means and standard errors of the mean are reported given 20 random splits of training and testing data.

the autoregressive order ($p \in \{1, 2, 3\}$) for DARX, to have a direct comparison between LAR and LARX, we set $p, q = 1$ for LARX. Finally, for the Markov models, we experiment with 3 types of CMC with $p \in \{1, 2, 3\}$ and 4 types of IOHMM: first-order IOHMMs with different number of hidden states $K \in \{2, 3, 4\}$, and a second-order IOHMM with $K = 2$.

A Comparison on Prediction Performance

Figure 6.1 compares the prediction performance of all 9 models (2 baseline models and 7 proposed models) on the three datasets. We first observe that the 7 proposed models almost always outperform the 2 baseline models on all evaluation metrics. For each dataset, the best-performing proposed model obtains a 2.2%–8.2% improvement on accuracy and F₁ score over the baseline models, and the log loss is also significantly decreased, especially compared to the running accuracy model. This suggests that when monetary interventions are provided in working sessions, it is necessary to explicitly model the impact of monetary interventions

Metric	Dataset	RA	LAR	SVM	NN	DARX	LARX	CMC	IOHMM
Accuracy	PUZZLE	0.025 ^{***}	0.029 ^{***}	0.007 ^{***}	0.004 [*]	0.013 ^{***}	0.008 ^{***}	0.007 ^{**}	0.005 [†]
	CLASSIFY	0.030 ^{***}	0.050 ^{***}	0.013 ^{**}	0.007 ^{**}	0.016 ^{***}	0.010 ^{**}	0.004	0.011 ^{**}
	TYPO	0.017 [*]	0.017 [*]	0.007	-0.001	-0.001	-0.009	-0.008	0.009 [†]
F ₁ score	PUZZLE	0.020 ^{***}	0.023 ^{***}	0.005 ^{***}	0.003 [*]	0.010 ^{***}	0.005 ^{**}	0.006 ^{***}	0.004 [*]
	CLASSIFY	0.049 ^{***}	0.061 ^{***}	0.018 ^{***}	0.011 ^{***}	0.016 ^{***}	0.013 ^{***}	0.010 ^{***}	0.021 ^{***}
	TYPO	0.023 ^{***}	0.023 ^{**}	0.011 [*]	0.000	0.001	-0.008	-0.006	0.011 [*]
Log loss	PUZZLE	-0.535 ^{***}	-0.055 ^{***}	-0.021 ^{***}	-0.003	-0.030 ^{***}	-0.028 ^{***}	-0.015 ^{***}	-0.003
	CLASSIFY	-0.719 ^{***}	-0.059 ^{***}	-0.018 ^{***}	-0.004	-0.023 ^{***}	-0.018 ^{***}	-0.010 ^{***}	-0.014 ^{***}
	TYPO	-0.586 ^{***}	-0.031 ^{***}	-0.041 ^{***}	-0.023 ^{***}	-0.046 ^{***}	-0.033 ^{***}	-0.100 ^{**}	-0.066 ^{***}

Table 6.1: Performance comparison between random forests (RF) and other prediction models. The differences in mean values for each metric are reported. The statistical significance of paired t-test is marked as a superscript, with [†], ^{*}, ^{**}, and ^{***} representing significance levels of 0.1, 0.05, 0.01, and 0.001 respectively.

in order to characterize the temporal crowd work quality accurately and confidently.

Among all prediction models, the *random forests* model seems to outperform other models as its high performance has been consistently observed across all datasets. In fact, random forests is the best-performing prediction model according to all three metrics on the PUZZLE and CLASSIFY dataset, and it is also the best-performing model on the TYPO dataset according to the log loss value. Table 6.1 presents a detailed comparison between random forests and other models. In particular, given a specific metric, we have evaluated that metric 20 times for each prediction model as there are 20 random splits of training and testing data. Thus, for each model, we obtain a performance vector with 20 elements. To compare the performance of random forests with another model, we take the average for the corresponding performance vectors of both models and compute the difference in the average values (e.g., average accuracy of random forests – average accuracy of DARX), which are reported in Table 6.1. We further use *paired* t-test to examine whether these differences are statistically significant, and the results are noted as superscripts in Table 6.1. As we can see in the table, compared to other models, random forests almost always has a significantly higher accuracy (i.e., positive differences for accuracy), higher F₁ score (i.e., positive differences for F₁ score) and lower log loss (i.e., negative differences for log loss), and none of the differences in unexpected directions (e.g., negative differences for accuracy) are statistically significant. These results suggest that

in practice, the random forests model gives high prediction performance for various types of tasks and thus is a good candidate model to use for requesters who are interested in making predictions on crowd work quality. We leave the problem of understanding why the random forests model is consistently accurate for future study.

A closer look at the estimated random forest model further provides us with a few practical insights for understanding the role of monetary interventions on worker performance. On the one hand, we find that the *average performance* is the most important feature for predicting work quality in the current task; on the other hand, it is observed that among all intervention-related features (i.e., current intervention level, average intervention level, historical intervention levels, historical intervention changes), the *average intervention level* is the most informative one for the prediction.

Prediction with Limited Training Data

Next, we examine the performance of different models when the requester has limited training data to start with. To mimic the realistic scenario for the requester to obtain more training data over time, given a particular training dataset, we first randomly take 5% of the workers in it and train the models using *only* the data from these workers. After examining the performance of these models on the testing dataset, we pick another random 5% of the workers in the original training dataset who are *not* previously selected, and *combine* their data with the data from the first 5% workers to create a training dataset that consists of 10% of the workers in the original training dataset. Following the similar process, we construct two more training datasets, with 20% and 50% of the workers in the original training dataset, respectively⁴.

⁴For the TYPO dataset, we only construct two datasets with 20% and 50% of the workers in the original training dataset (the number of workers in the these two training datasets are 13 and 32, respectively) because the total number of workers in this dataset is relatively small.

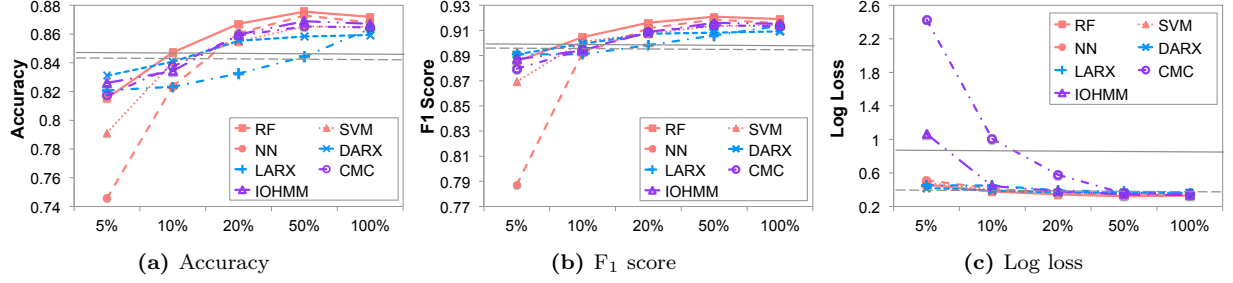


Figure 6.2: Performance comparisons for all prediction models on the PUZZLE dataset when training data is limited. The solid and dashed gridlines are the performance references for the RA and LAR models, respectively (training datasets are not required for these two baseline models).

Figure 6.2 illustrates the performance of different models on the PUZZLE dataset when the models are estimated from the 5%, 10%, 20%, 50% and the full training datasets. The performance of the prediction models improves as the amount of training data increases — with the training data from only 5% of the workers (i.e., 12 workers), all the 7 proposed models are actually inferior to the baseline LAR model according to all three metrics. Some models are especially sensitive to the size of the training dataset. For example, when the training data is very limited, SVM and NN suffers from a significantly lower accuracy and F₁ score, while CMC and IOHMM models have very high log loss values. On the other hand, once the size of the training dataset has been increased to include 20% of all workers (i.e., 48 workers) in the original training dataset, almost all proposed models outperform both RA and LAR on all metrics. When the size of the training dataset further increases, while the prediction performance of different models keeps improving, the marginal benefit of extra training data also decreases. Importantly, we notice that even though the model is trained on only a fraction of the workers in the original training dataset, the random forests model still presents better prediction performance than other models in most cases, which suggests the robustness of this model against the limited training data. Similar results are also observed in the CLASSIFY and TYPO datasets.

Therefore, as a practical implication, a requester may consider to use the LAR model to

predict crowd work quality in task sessions at the initial stage when they just start to recruit workers to work on their tasks. After collecting a small training dataset (e.g., a dataset of about 50 workers), the requester can switch to models that explicitly consider the impact of monetary interventions, especially the random forests model, to obtain more accurate predictions with higher confidence.

Predictions with Limited Ground Truth

Finally, we consider the scenario when the requester only has access to limited amount of ground truth. Ground truth information is quite valuable in crowd work because in many cases, the requester will not be able to assess the work quality in a task without the ground truth. So far, we have assumed that the requester knows the ground truth to all his tasks hence he can evaluate the work quality for *every* task in a task session, and all the seven proposed models rely on the observation of past work quality (i.e., the sequence \mathbf{y}) when making predictions on work quality in the current task. To understand how the prediction performance of different models are influenced when this assumption is violated, that is, when the requester can only check the work quality for a limited number of tasks in the session, we conduct a new set of experiments.

In particular, given a specific split of training and testing data, prediction models are learned using the full training dataset as previously described⁵. When making predictions for workers in the testing dataset, we fix the first three tasks in each worker’s session to be tasks with ground truth in order to obtain an initial record of the worker performance. Then, for the rest of the tasks in the session, we randomly select a certain portion (r) of them to be

⁵We assume that the requester can still evaluate the work quality on every task in the training dataset. This assumption is realistic, for example, if the requester bundles multiple tasks with ground truth into a single session and provide such task sessions to workers in the initial phase when the training dataset is collected. Models estimated from such training dataset can be used to predict work quality on tasks without ground truth when tasks with or without ground truth are similar (e.g., have similar difficulty levels).

tasks with ground truth, hence work quality is only observable for these tasks. We vary this percentage, that is, $r \in \{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$, and examine the performance of our prediction models in each of these cases when the ground truth is limited to different degrees.

For the simplicity of illustration, in this experiment, we focus on the two baseline models and three of the proposed models — RF, LARX and IOHMM, one from each category. For IOHMM, the lack of ground truth can be taken care of by simply updating the state belief in a different way when work quality is not observable⁶. For other models, we take a Monte Carlo approach to address the prediction problem: We maintain a set of $M = 100$ work quality sequences $Q = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M\}$, where $\mathbf{q}_m = (q_m^1, q_m^2, \dots, q_m^{l-1}) (1 \leq m \leq M)$ is a sequence of “simulated” work quality for all the $l - 1$ tasks provided to the worker so far. To forecast the worker’s performance on the l -th task, we first make a prediction with each of the M work quality sequences and then take an average of all M predictions. That is, $Pr(y^l = 1) = \frac{1}{M} \sum_{m=1}^M p_m$, where p_m is the predicted probability of high-quality work on task l assuming that \mathbf{q}_m is the observed work quality sequence for the past $l - 1$ tasks. After the prediction, if the ground truth for task l is available hence the requester can actually decide the work quality y^l , we update \mathbf{q}_m by setting $q_m^l = y^l$; otherwise, we sample a work quality \hat{y}^l according to $Pr(\hat{y}^l = 1) = p_m$, and then update \mathbf{q}_m as $(q_m^1, q_m^2, \dots, q_m^{l-1}, \hat{y}^l)$.

Figure 6.3 plots for each of the 5 models, the change of average prediction performance as the amount of tasks with ground truth increases in the PUZZLE dataset. We find that RF, LARX and IOHMM models almost always make more accurate predictions with higher confidence compared to the baseline RA and LAR models. Among RF, LARX and IOHMM, the RF and IOHMM models are more robust when the requester has limited access to the

⁶For example, when the order of the IOHMM is $p = 1$, the state belief is updated according to the formula $b_{l+1}(j) \propto \sum_{k=1}^K b_l(k) P_{tr}(j|k, a^l)$ if the requester doesn’t have ground truth for the l -th task, rather than according to $b_{l+1}(j) \propto \sum_{k=1}^K b_l(k) P_{tr}(j|k, a^l) P_e(y^l|j, a^l)$.

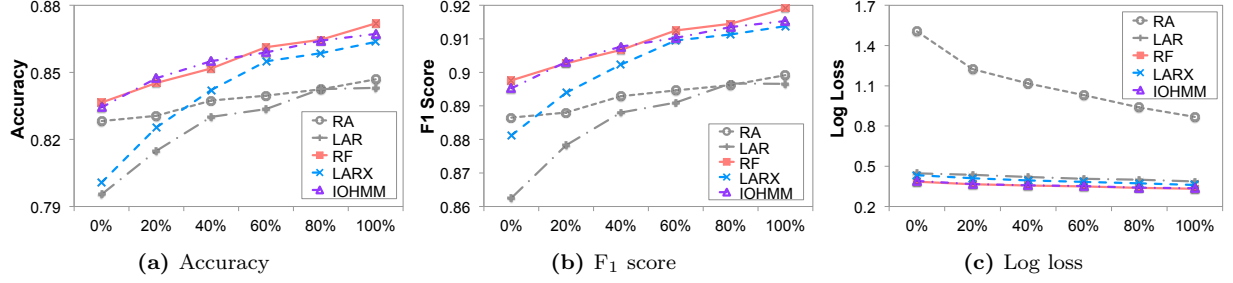


Figure 6.3: Performance comparisons for different prediction models on the PUZZLE dataset when ground truth is limited.

ground truth information. In particular, the RF and IOHMM model outperforms the two baseline models as well as the LARX model regardless of how small the fraction of tasks with ground truth is, and the prediction performance of RF and IOHMM when only 20% of the tasks has ground truth even exceeds the performance of the baseline models when the work quality is always observable for all tasks. Similar results are observed on other datasets, and they provide further supporting evidence for using the random forests model to predict crowd work quality under monetary interventions — it can not only make consistently accurate predictions for various types of tasks or given small set of training data, but also presents robust performance under limited supervision.

6.3 Controlling the Provision of Monetary Interventions Dynamically

In this section, we demonstrate our algorithmic approach to dynamically decide the provision of monetary interventions based on the statistical models of worker behavior that we learn from the empirical data. As an example, we show how can we algorithmically control the placement of performance-contingent bonuses in working sessions when the *first-order input-output hidden Markov model* is used to characterize the impact of monetary interventions. While in Section 6.2, we find that the random forests model performs the best for predicting

crowd work quality under monetary interventions in general, we choose to demonstrate our algorithmic approach with the input-output hidden Markov model (IOHMM) due to its nice connection with partially observable Markov decision process (POMDP), a mathematical framework for modeling decision making under uncertainty, which we will detail later in this section. Embedding other statistical models, such as the random forests model, into our algorithmic monetary intervention control framework, is therefore an interesting future direction that we will briefly discuss in Section 6.4.

6.3.1 Making Bonus Decisions with IOHMM

Similar as that in Section 6.2, we assume that the requester has collected a historic dataset on worker behavior in working sessions. For each task in a particular worker’s working session, the requester records whether a bonus is provided and the worker performance in it. With the historic dataset for a group of workers, a first order IOHMM for this worker population can be learned through an expectation-maximization algorithm [Bengio and Frasconi, 1996].

Given the learned IOHMM \mathcal{M} to describe the impact of bonuses on work quality in a working session, we next solve the problem of deciding whether or not to place a bonus on each task for a new incoming worker in her working session.

To quantify the requester’s tradeoff between work quality and financial cost, we assume that the requester obtains a utility of w_h (or w_l) when he gets a high-quality (or low-quality) answer, while the economic cost for paying a performance-contingent bonus is c . We further assume that the requester has a quasi-linear utility function $U = w_h N_{HQ} + w_l N_{LQ} - c N_B$, where N_{HQ} (or N_{LQ}) and N_B represents the number of high-quality (or low-quality) answers elicited and the number of times a performance-contingent bonus is incurred, respectively. We consider the scenario when the requester can continuously make observations on worker performance over time (i.e., the requester has access to the ground truth information for all tasks), and we are interested in dynamically controlling the placement of bonus in a working

session of T tasks in an *online* fashion given the observed worker performance. That is, we keep making decisions on whether the requester should provide a bonus to a worker in her *next* task given the history of inputs and outputs in all tasks that the worker has completed so far in the session.

In particular, for a worker who has completed t_c tasks, we estimate the distribution of her current state as $\mathbf{b}(t_c)$ (i.e., the “state belief”) based on \mathcal{M} . We define $EU_{max}(\mathbf{b}, a, l)$ as the *maximum* expected utility a requester can obtain in the next l tasks given that the current state belief is $\mathbf{b} = (b(1), \dots, b(K))$ (K is the total number of hidden states estimated in the IOHMM), the input level for the next task is a , and input levels for later tasks follow the optimal policy. Thus, the optimal input level for the next task is $a^{t_c+1} = \operatorname{argmax}_{a \in \{0,1\}} EU_{max}(\mathbf{b}(t_c), a, T - t_c)$ — when $a^{t_c+1} = 1$, we offer a bonus on the next task; otherwise, we don’t.

$EU_{max}(\mathbf{b}, a, l)$ can be calculated recursively. Specifically, suppose the worker is currently in state k ($1 < k < K$). The requester’s expected utility in the next task if he places an input a on it can be computed as:

$$R(k, a) = \sum_{i=1}^K P_{tr}(i|k, a) \cdot (P_e(0|i, a)w_l + P_e(1|i, a)(w_h - \mathbb{I}(a = 1)c)) \quad (6.8)$$

Naturally, the requester’s expected utility in the next task when his current state belief about the worker is \mathbf{b} and the next input is a can be denoted as:

$$R(\mathbf{b}, a) = \sum_{k=1}^K b(k)R(k, a) \quad (6.9)$$

When there is only one task left in the session, that is, $l = 1$, $EU_{max}(\mathbf{b}, a, l) = R(\mathbf{b}, a)$; otherwise, we will need to consider both the immediate utility the requester will be able to obtain in the next task, as well as the maximum expected utility that the requester will be able to obtain in the later tasks, averaging on the possible outcomes that the requester may

observe in the next task:

$$EU_{max}(\mathbf{b}, a, l) = R(\mathbf{b}, a) + \sum_{y \in \{0,1\}} \left(\sum_{k=1}^K b(k) \sum_{i=1}^K P_{tr}(i|k, a) P_e(y|i, a) \right) V(\mathbf{b}'_{a,y}, l-1) \quad (6.10)$$

where $V(\mathbf{b}, l) = \max_{a \in \{0,1\}} EU_{max}(\mathbf{b}, a, l)$ and $\mathbf{b}'_{a,y}$ is the updated state belief if the input for the next task is a and the observed output is y , which can be computed as the following:

$$b'_{a,y}(k) \propto \sum_{i=1}^K b(i) P_{tr}(k|i, a) P_e(y|k, a) \quad (6.11)$$

And finally, in preparation for the decision making in future tasks, we need to update the belief state after implementing the input level a^{t_c+1} and observing the actual output level y^{t_c+1} , that is, $\mathbf{b}(t_c + 1) = \mathbf{b}'_{a^{t_c+1}, y^{t_c+1}}$.

Heuristic Solutions

It turns out that the decision making problem above is equivalent to solve a finite-horizon partially observable Markov decision process (POMDP), with “*action*” corresponding to “input” in \mathcal{M} and the *reward* of taking action a in state k being $R(k, a)$ as defined in Equation 6.8. In practice, finding exact solutions for POMDPs are often computationally intractable [Papadimitriou and Tsitsiklis, 1987]. Therefore, we list a few heuristic algorithms to solve the problem approximately.

Algorithm 1: n -step look-ahead. When making decisions for whether to place an extra bonus on the next task, we look ahead for at most n tasks. That is, $a^{t_c+1} = \operatorname{argmax}_{a \in \{0,1\}} EU_{max}(\mathbf{b}(t_c), a, n')$, where $n' = \min(n, T - t_c)$. A similar strategy is used in [Dai et al., 2010] for optimal control of crowdsourcing workflows.

Note that if a worker’s hidden state z^{t_c} after completing t_c tasks as well as her states in

the future tasks can be accurately identified, the finite horizon POMDP degenerates into a finite horizon MDP, for which we can calculate the optimal policy efficiently. In particular:

$$Q_t(k, a) = R(k, a) + \sum_{i=1}^K P_{tr}(i|k, a) V_{t-1}(i) \quad (6.12)$$

$$V_t(k) = \max_{a \in \{0,1\}} Q_t(k, a); V_0(k) = 0 \quad (6.13)$$

$$\pi_t(k) = \operatorname{argmax}_{a \in \{0,1\}} Q_t(k, a) \quad (6.14)$$

where $Q_t(k, a)$ (referred to as the “Q-function”) is the maximum expected utility to obtain when taking action a in state k with t steps to go, $V_t(k)$ is the maximum expected utility when the current state is k and there are t steps to go, and $\pi_t(k)$ is the optimal MDP policy on state k when the length of the horizon is t . The optimal MDP policy can be computed with the value iteration or policy iteration algorithm. We thus consider two algorithms that leverage the solution for the underlying MDP of the POMDP problem:

Algorithm 2: MLS-MDP. We infer the most likely sequence (MLS) of hidden states up to the current task (i.e., $\mathbf{Z}^{\hat{1}:t_c} = (z^{\hat{1}}, \dots, z^{\hat{t}_c})$) using the Viterbi algorithm [Viterbi, 1967] and estimate z^{t_c} as $z^{\hat{t}_c}$. The input level for the next task a^{t_c+1} is then set to be $\pi_{T-t_c}(z^{t_c})$, that is, the optimal MDP policy on state z^{t_c} when the length of horizon is $T - t_c$.

Algorithm 3: Q-MDP. We first calculate $Q_{T-t_c}(k, a)$ for the underlying MDP, which is the Q-function value for taking action a on state k with $T - t_c$ steps to go. Then, $a^{t_c+1} = \operatorname{argmax}_{a \in \{0,1\}} \sum_{k=1}^K p_k Q_{T-t_c}(k, a)$, with p_k being the k -th element of $\mathbf{b}(t_c)$ [Littman et al., 1995].

6.3.2 Experimental Evaluation with Real Crowd Workers

To examine whether our algorithmic approach of placing bonuses can effectively improve requester utility in working sessions in the real on-demand work environment, we designed and conducted an online experiment on MTurk.

Experimental Settings

In this experiment, we used the word puzzle tasks as described in Section 6.2.2. In particular, each worker completed 9 tasks in a working session (i.e., one HIT). A worker earned a performance-independent reward of 5 cents in each task, that is, the base payment for the HIT was 45 cents. In addition, we also informed the worker that some tasks in the session are “bonus tasks” (specified with a bonus icon), in which she may earn an extra bonus of 5 cents if she submits a *high-quality* answer to it by pointing out more than 80% of all appearances of the target word.

Corresponding to the two steps in our algorithmic approach, we divide our experiment into two phases.

In the first phase, we collected a training data set by recruiting 50 MTurk workers to participate in our experiment. For each of the 9 tasks that a worker completed in the HIT, we randomly set it as a bonus task with a 20% chance; whether that task was a bonus task and whether the worker submitted a high-quality answer to it (i.e., found out the 80% of all appearances of the target word) was recorded. We then learned an IOHMM to understand the impact of bonuses on worker performance in the word puzzle task sequences using the collected data set. Specifically, we ran the expectation-maximization algorithm with 100,000 random restarts, and each run was terminated after convergence or 500 iterations, whichever was reached earlier. In searching for a parsimonious model, we experimented on a range of values for the number of hidden states ($K = 1 \sim 7$) to train different IOHMMs, and the IOHMM with the maximized Bayesian information criterion (BIC) score [Schwarz et al.,

1978] was selected to be used in the second phase. In our experiment, $K = 2$ for the selected IOHMM.

The second phase of our experiment is the testing phase, in which we had 6 experimental treatments and each treatment corresponded to one bonus scheme. In particular, we included 3 dynamic bonus schemes that were designed according to our algorithmic approach using different heuristics (i.e., 2-step look-ahead⁷, MLS-MDP and Q-MDP). When these schemes were used in treatments, we kept track of a worker’s performance in the session and used the learned IOHMM to strategically make a decision on whether to offer an extra bonus to the worker on the next task. As a comparison, we also considered 3 fixed or random baseline bonus schemes: not placing bonus in any task (No Bonus), always placing bonuses in all tasks (All Bonus) and randomly choosing 50% of the tasks to place bonuses (50% Bonus). The utility parameters we used in the experiment are $w_h = 0.15$, $w_l = 0$ and $c = 0.05$ ⁸.

To make sure the IOHMM learned from the training phase would be useful for the testing phase, we recruited workers from the same pool by running our second phase experiment exactly 2 weeks after the first phase experiment around the same time. Each worker was randomly assigned to one treatment and 50 workers were recruited for each treatment. All workers in a treatment were paid according to the bonus scheme of that treatment. We again collected data on the presence of bonus and work quality for each task and each worker.

Our experiment was limited to U.S. workers and each worker was allowed to take the HIT only once.

⁷We set $n = 2$ to balance between performance and efficiency based on simulations for workers who indeed behave according to the learned IOHMM.

⁸ c was set to be 0.05 as the bonus magnitude used in the experiment is \$0.05. We set $w_h = 0.15$ and $w_l = 0$ to ensure that a dynamic bonus scheme doesn’t become a No Bonus or All Bonus scheme.

Interpreting the Learned IOHMM

The IOHMM we learned from the training dataset is as follows: the initial state belief is estimated as $\mathbf{b}(0) = (0.67, 0.33)$, suggesting that at the beginning of the working session, 67% of the workers start from state 1 while the rest 33% start from state 2. To further understand what states 1 and 2 are, we look into the estimated emission probability matrices:

$$\mathbf{E}^0 = \begin{pmatrix} 0.10 & 0.90 \\ 0.88 & 0.12 \end{pmatrix}, \quad \mathbf{E}^1 = \begin{pmatrix} 0.13 & 0.87 \\ 0.61 & 0.39 \end{pmatrix}$$

The first and second row of the matrices corresponds to state 1 and 2, respectively; and the first and second column of the matrices represents the probability for submitting low-quality work and high-quality work, respectively. Therefore, we find that when workers are in state 1, when a bonus is not provided in a task, she will submit high-quality work 90% of the time (hence submit low-quality work 10% of time). Meanwhile, when a bonus is provided in a task, a worker in state 1 will submit high-quality work 87% of the time. In other words, workers in state 1 almost constantly submit high-quality work regardless of whether the monetary intervention is presented in a task or not. In contrast, if a worker is in state 2, only 12% of her submissions is of high-quality if no bonus is provided, but using an additional performance-contingent bonus, she can be incentivized to submit high-quality work 39% of the time.

Furthermore, the transition probability matrices are estimated as follows (round to 2 decimal places):

$$\mathbf{T}^0 = \begin{pmatrix} 0.92 & 0.08 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{T}^1 = \begin{pmatrix} 1 & 0 \\ 0.09 & 0.91 \end{pmatrix}$$

The first and second row corresponds to state 1 and 2 at time t , and the first and second column corresponds to state 1 and 2 at time $t + 1$, respectively. Interestingly, we find that

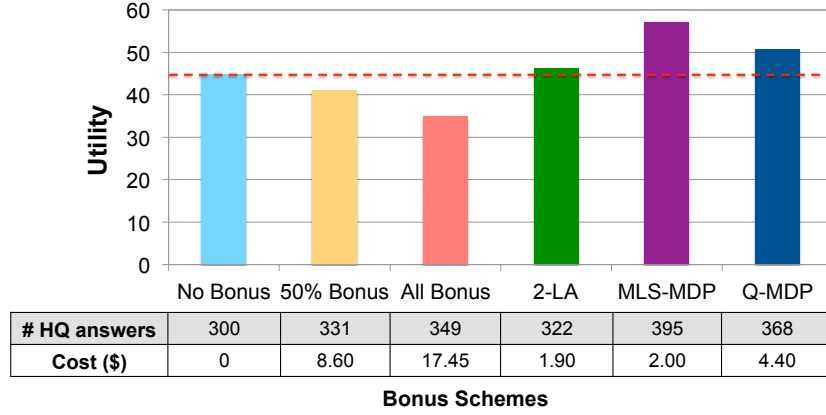


Figure 6.4: The requester’s utility across 6 treatments in the second phase MTurk experiment.

when we don’t provide bonus in a task, state 1 workers (i.e., workers who constantly submit high-quality work) may “slack off” and switch to state 2 with a small probability of 0.08, yet they will stay in state 1 if a bonus is provided in the task. In addition, if we don’t provide bonus in a task, state 2 workers will stay in state 2, but with a small chance of 9%, they may be incentivized to transit to state 1 and start to constantly submit high-quality work when a bonus is placed on a task.

Comparing the Requester Utility under Different Bonus Schemes

Figure 6.4 compares the overall utility a requester derives from *all* 50 workers in the working session across the 6 treatments of our second phase experiment. As we can see in the figure, among the 3 baseline bonus schemes, the best scheme is not to pay bonus on any task at all. Yet, following our algorithmic approach, the 3 dynamic bonus schemes (i.e., 2-step look-ahead, MLS-MDP and Q-MDP) lead to an increase of 3.11%, 27.22% and 12.89% in the requester utility, respectively, compared to the No Bonus scheme. To see whether the utility improvement is material, we decompose the overall requester utility in each treatment into the number of high-quality answers the requester elicits (hence the economic benefits) and the cost the requester pays to encourage better performance (see the table in Figure 6.4). Results

	No Bonus	50% Bonus	All Bonus
2-LA	1	0.007	<0.001
MLS-MDP	0.03	<0.001	<0.001
Q-MDP	0.38	<0.001	<0.001

Table 6.2: p-values of the Wilcoxon rank-sum tests for pairwise comparisons.

suggest that the requester can elicit *more* high-quality work with *lower* cost by applying our dynamic bonus schemes. For example, compared to the 50% Bonus scheme, a requester using the 2-step look-ahead scheme obtains a similar number of high-quality answers with a 77.9% saving in money; while a requester following the MLS-MDP (or Q-MDP) scheme elicits 19.3% (or 11.2%) more high-quality answers with roughly a quarter (or a half) of the cost.

The statistical significance of the improvement in utility brought by our algorithmic approach is further examined through Wilcoxon rank-sum tests. Specifically, we compute the utility a requester obtains from *each* worker in every treatment. Thus, in total, we have 6 samples of requester utility with 50 data points in each sample. We conduct pairwise comparisons to test whether a pair of samples have the same mean value, and the p-values of the tests are reported in Table 6.2. Almost all pairwise comparisons are statistically significant at the $p = 0.05$ level, which suggests that our approach helps the requester to improve his utility.

Finally, to get a qualitative intuition of how our dynamic bonus schemes work, we pick a few exemplary workers from the treatment in which the MLS-MDP bonus scheme is used and take a close look at how they are awarded in the working session. Figure 6.5 displays for each of the 4 selected workers, whether a bonus is provided and whether her submitted answer is of high quality for each task in the working session. The comparison between worker A and worker B first suggests that our algorithmic approach can effectively differentiate “diligent” workers from “lazy” workers and reward them differently: For a diligent worker (worker A, a worker who is likely to stay in state 1) who always submits high-quality answers, there is

Worker	Inputs & Outputs in the Working Session									
A	Bonus?	×	×	×	×	×	×	×	×	×
	High-quality?	1	1	1	1	1	1	1	1	1
B	Bonus?	×	✓	✓	✓	✓	✓	✓	✓	✓
	High-quality?	0	0	0	0	0	0	0	0	0
C	Bonus?	×	✓	✓	✓	✓	✓	×	×	×
	High-quality?	0	0	0	1	1	1	1	1	1
D	Bonus?	×	×	×	×	✓	✓	✓	×	×
	High-quality?	1	1	0	0	1	1	1	1	0

Figure 6.5: Examples for offering bonus to a worker in the working session based on the MLS-MDP bonus scheme.

no need for the requester to place extra bonuses in the working session; however, for a lazy worker who can be responsive to financial incentives (worker B, a worker who is likely to stay in state 2), the requester keeps offering bonuses with the hope that he may increase the work quality through providing additional motivation to the worker⁹. In fact, the MLS-MDP bonus scheme strategically focuses on incentivizing lazy workers — on average, the requester offers a bonus on 6 tasks to a worker who performs well in at most half of the tasks in the working session, while for a worker who performs well in more than half of the tasks, the requester offers a bonus on only 0.49 tasks. Furthermore, our algorithmic approach also seems to offer bonuses at the right timing. On the one hand, for a worker who starts a session with unsatisfying performance (worker C), the requester keeps placing bonus on each task to incentivize better performance until the worker stabilizes in submitting high-quality answers; on the other hand, for a worker who slacks off from her initial good performance (worker D), the requester provides extra incentives in time to bring back hard working from the worker.

⁹Since the bonus is performance-contingent, offering bonus per se is *not* costly; the cost will only be incurred when the work quality in a bonus task meets the predefined standard.

6.3.3 Examining the Robustness through Simulation

The performance of the dynamic bonus schemes in our MTurk experiment suggests the promise of algorithmically controlling the provision of financial incentives in on-demand work. However, one may wonder that following our algorithmic approach, whether the high requester utility can always be obtained in various worker populations where workers potentially behave in different ways. To understand the robustness of our approach, we further ran simulations on two synthesized datasets and each dataset was generated according to a predefined worker behavior model.

Specifically, given a particular worker behavior model, in the training phase, we generated a dataset of 3,000 workers — each worker completed a session of 50 tasks (20% of them were randomly selected as bonus tasks) and decided her performance in each task probabilistically according to the given model. In the testing phase, we considered the same 6 bonus schemes as those in our MTurk experiment. Six groups of testing data were thus generated: each group was assigned to a unique bonus scheme and was composed of 100 workers; each worker completed a session of 10 tasks and was paid according to the bonus scheme of her group, while her performance in each task was controlled by the same behavior model as that used for generating the training dataset. The requester’s utility under a specific bonus scheme was calculated as the sum of utilities that the requester derived from all 100 workers of the corresponding group, with w_l and w_h set to be 0 and 1, respectively. We repeated the simulation 30 times and reported the mean value of the requester’s utility for each scheme.

Model 1: Workers with Two Capability Levels

In our first worker behavior model, we assume that when a performance-contingent bonus is placed (or not placed) in a task, the probability for worker i to submit a high-quality answer in a task is acc_i^h (or acc_i^l). This model is consistent with previous empirical observations that the existence of performance-based bonuses can affect the worker performance [Harris, 2011,

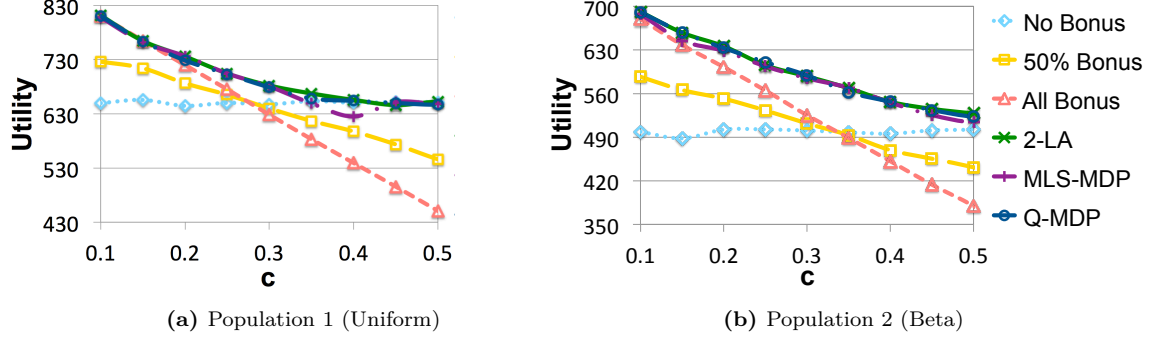


Figure 6.6: The requester’s utility when workers have two capability levels in reaction to the placement of bonus. Error bars are omitted as they are too small.

Ho et al., 2015].

We constructed two different worker populations based on this worker behavior model. To test the performance of different bonus schemes when a requester’s willingness to reward workers differs, we also varied the economic costs of the performance-contingent bonus c (i.e., the magnitude of bonus) from 0.1 to 0.5 in the simulation. Figure 6.6a demonstrates the simulation results for a population that is composed of two types of workers: for the first type, $acc_i^l = 0.5$ and $acc_i^h = 0.9$; while for the second type, $acc_i^l = 0.8$ and $acc_i^h = 0.9$. Each worker in the population is drawn uniformly randomly from the two types. Similarly, Figure 6.6b corresponds to another population where each worker draws her accuracy from Beta distributions: $acc_i^l \sim Beta(2, 2)$ and $acc_i^h \sim Beta(6, 2)$. As the figures suggest, for both populations, when the magnitude of bonus is small (or large) enough such that adding a bonus is always beneficial (or too costly), the dynamic bonus schemes lead to similar requester utility as the All Bonus (or No Bonus) scheme does. However, when the bonus magnitude is moderate, the dynamic bonus schemes robustly result in higher requester utility compared to any single baseline bonus scheme.

Model 2: Workers Influenced by Reference Payment Levels

Both previous literature and our own work as described in Section 5.1 suggest another possible worker behavior in reaction to financial incentives in a working session. That is, a worker maintains and updates a reference point of “appropriate” payment level when she completes tasks in a working session and decides her performance in each task by comparing the provided payment with the reference of that time [Popescu and Wu, 2007, Akerlof and Yellen, 1990]. To see how our algorithmic approach performs when workers indeed behave in this way, we define the second worker behavior model. In particular, we assume that each worker behaves according to an IOHMM, with the hidden state z^t corresponding to the reference payment level r_{z^t} in the worker’s mind in the t -th task¹⁰. Each worker i is further characterized by her skill level α_i and her responsiveness to financial incentives β_i , and the emission probability in task t is parameterized as $P_e(1|z^t, a^t) = \frac{1}{1+e^{-\alpha_i-\beta_i(a^t-r_{z^t})}}$. Intuitively, the larger α_i or β_i is, the worker is more skilled or more responsive to rewards hence more likely to produce high-quality answers.

For manageability, we assume in the simulation that each worker has $K = 3$ hidden states (i.e., 3 discrete reference payment levels), that is, $\mathbf{r} = (r_1, r_2, r_3) = (0.2, 0.6, 1.2)$. Each worker updates her reference payment level in the working session according the transition probability matrices:

$$\mathbf{T}^0 = \begin{pmatrix} 0.8 & 0.15 & 0.05 \\ 0.3 & 0.6 & 0.1 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}, \quad \mathbf{T}^1 = \begin{pmatrix} 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.05 & 0.1 & 0.85 \end{pmatrix}$$

where the (i, j) -th element of matrix \mathbf{T}^a represents the probability for a worker to update her reference payment level from r_i to r_j given the current input level a , i.e., $P_{tr}(j|i, a)$.

¹⁰The reference payment level is defined relative to the magnitude of the bonus, e.g., $r_k = 0.5$ means that in state k , the worker considers a half of the current bonus as an appropriate payment.

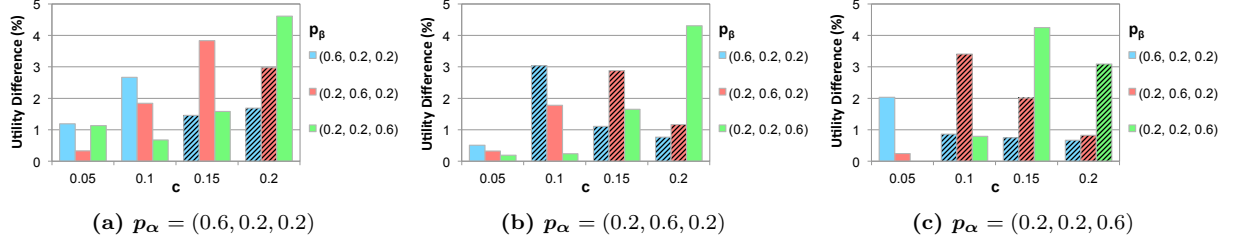


Figure 6.7: The requester’s average utility increase (in percentage) when following a dynamic bonus scheme rather than the best-performing baseline scheme; workers are influenced by their reference payment levels in mind. Shaded (Unshaded) bars indicate that the best-performing baseline scheme is the “No Bonus” (“All Bonus”) scheme in that condition.

Furthermore, we assume that there are 3 possible skill levels for a worker, i.e., $\alpha_i \in \{0, 1, 3\}$, and each worker i draws her skill level according to the categorical distribution $\mathbf{p}_\alpha = (p_\alpha^1, p_\alpha^2, p_\alpha^3)$, where $p_\alpha^1 + p_\alpha^2 + p_\alpha^3 = 1$ and p_α^1 (or p_α^2, p_α^3) represents the probability that $\alpha_i = 0$ (or 1, 3). Similarly, we also consider 3 levels of responsiveness to financial incentives, that is, $\beta_i \in \{0, 1, 3\}$, and each worker draws her β_i according to another categorical distribution $\mathbf{p}_\beta = (p_\beta^1, p_\beta^2, p_\beta^3)$. Varying \mathbf{p}_α and \mathbf{p}_β thus provides us the flexibility to construct a number of populations in which various types of workers are mixed in different proportions. For each population and 4 selected bonus magnitude (i.e., $c = 0.05, 0.1, 0.15, 0.2$), Figure 6.7 displays the utility difference (in percentage) a requester obtains by following a dynamic bonus scheme (averaged over the 3 dynamic schemes) over following the *best-performing* baseline scheme to control bonuses in the session: on the one hand, the best-performing baseline scheme *differs* across various populations and magnitude of bonus, suggesting that following a single baseline bonus scheme can’t guarantee a high requester utility in all conditions; on the other hand, following our dynamic bonus schemes, the requester can consistently obtain similar or higher utility than what he could have obtained by following the *best-performing* baseline scheme, which again implies that the improved performance of our approach is robust.

6.4 Discussion

In this chapter, we first model the impact of financial incentives on worker performance in the on-demand work environment using quantitative models. We present a wide range of models from 3 categories, including supervised learning models, variants of autoregressive models and Markov models. We further conduct an empirical comparison on the performance of these models in predicting crowd work quality under monetary interventions for different types of tasks, as well as in different realistic scenarios, such as when the training data is limited or the amount of available ground truth information is limited. In addition, based on one particular type of statistical model of worker behavior (i.e., the first-order input-output hidden Markov model), we propose an algorithmic approach to dynamically control the provision of monetary intervention in on-demand working session. Our MTurk experiment and simulation results suggest that our approach can robustly lead to significant improvement in requester utility compared to several fixed and random bonus schemes, which are the common practice of today.

There are many interesting future directions for this work. First of all, previous studies have shown that worker’s behavioral traces in crowdsourcing tasks, such as how long they stay in a task and how they interact with the task interface, can be effective in predicting worker performance [Rzeszutarski and Kittur, 2011, Sameki et al., 2015]. It is therefore an interesting future work to examine whether these behavioral traces can be integrated into the current models to further improve the prediction performance on crowd work quality under monetary interventions.

Secondly, while in Section 6.3, we demonstrate our algorithmic approach to control the provision of monetary intervention in working sessions based on the input-output hidden Markov model, in fact, all the statistical models in Section 6.2 can be used as the empirical worker behavior model and hence be incorporated into the algorithmic incentive control

framework. In particular, given any model that we describe in Section 6.2, we can take the n -step look-ahead algorithm to decide the placement of bonuses by first predicting work quality in the next n tasks using the given model and then selecting the bonus level for the next task that maximizes the expected utility for the next n tasks. In Section 6.2, we have identified the random forests model to be an excellent model for requesters to use to predict work quality in practice. It will be interesting to empirically examine that, compared to other statistical models, whether the improvement in prediction accuracy brought up by the random forests model can further guide requesters to place bonus in a more intelligent way and hence increase their utility.

Our algorithmic approach of incentive control has a few limitations, though. For example, currently, our approach leads to a policy that provides more bonus opportunities for workers with lower accuracy. This may incentivize workers to strategically game with our reward policy, or drive high accuracy workers away in the long term. Furthermore, this may raise ethical concerns—our approach essentially provides a “personalized” bonus scheme to each worker based on her past performance, and such “differential pricing” strategy naturally leads one to ask how we can ensure the fairness of payment in the on-demand work environment, especially given that workers may talk to each other as shown in Chapter 3. In short term, the concerns for worker’s strategic behavior and complains about payment fairness may not be obvious, because workers have limited understanding on how exactly the underlying algorithm for incentive provision works. In addition, workers also need to have an accurate estimation on how well they perform in each task in a session to fully understand the mechanics of our dynamic bonus policy. However, as workers interact with the algorithm for a long time, their knowledge about the algorithm will inevitably improve. Therefore, it is necessary for us to understand the long-term impact of our algorithmic approach and explore more incentive-compatible, sustainable monetary intervention control approach in the future (e.g., by providing high accuracy workers with more work opportunities or higher bonus levels).

Another limitation of our current algorithmic approach is that we currently assume the requester has the access to the ground truth information for all tasks and thus can monitor a worker’s performance over time for all tasks that she has completed. In reality, the requester may only have limited ground truth information, which leads to another interesting future direction of designing algorithmic approach for incentive placement when the requester can only conduct *spot check*, that is, checking the work quality in a few selected tasks. In that case, the requester needs to not only decide when to place monetary interventions in a working session, but also when to check the work quality in the working session.

Finally, although in this chapter, we focus on discussing our algorithmic framework to optimize the placement of monetary interventions in the on-demand work sessions, the same models and algorithms can actually be extended to different contexts of deciding the provision of any *external interventions* in the on-demand work, such as the provision of performance feedback [Dow et al., 2012], the switch of workflows [Lin et al., 2012] and the deliver of communication messages [Segal et al., 2016]. The optimization goal can also generalize from increasing the requester utility to other dimensions of interests, such as improving the worker engagement.

6.5 Acknowledgements

The research presented in this chapter was conducted with Yiling Chen. Portions of this chapter previously appeared in the HCOMP publication *Predicting Crowd Work Quality under Monetary Interventions* [Yin and Chen, 2016] and the IJCAI publication *Bonus or Not? Learn to Reward in Crowdsourcing* [Yin and Chen, 2015]. We thank the support of the National Science Foundation under grant CCF-1301976 and the Xerox Foundation on this work.

Chapter 7

Designing for Intrinsic Motivation: A Case Study on Curiosity

So far, we have been focusing on studying how *financial incentives* can be used to increase the *extrinsic motivation* of workers in the on-demand economy. It's very natural and straightforward to consider using extrinsic incentives like financial rewards to motivate workers in the on-demand work environment, because it is a common practice to compensate labor with money. Recently in both the research community and the industry, however, there is an increasing attention on promoting *intrinsic motivation* in the incentive design of the on-demand work. The hope is that by integrating intrinsic motivation and providing a truly enjoyable and inherently satisfying work experience to on-demand workers, we may be able to obtain higher levels of engagement and performance from workers. Therefore, on the basis of various types of intrinsic motivations, a wide range of incentive mechanisms have been proposed and studied in different contexts of on-demand work, including both paid crowdsourcing markets and volunteer-based platforms like citizen science projects [Raddick et al., 2013, Rotman et al., 2012, Shirk et al., 2012]. For example, it has been showed that framing tasks as something meaningful to do [Ariely et al., 2008, Chandler and Kapelner,

2013, Rogstadius et al., 2011, Shaw et al., 2011], applying gamification elements like points, badges and leaderboards [Feyisetan et al., 2015, von Ahn and Dabbish, 2008, Bowser et al., 2013, Cooper et al., 2010], and inserting “micro-diversions” (i.e., breaks during which workers can engage with an enjoyable, task-irrelevant activity) in a long (1+ hours) sequence of tasks [Dai et al., 2015] can all exert a positive impact on worker engagement and productivity. There are also numerous attempts to engineer virtual reward systems [Easley and Ghosh, 2013], provide social comparisons and visualizations [Grevet et al., 2010, Huang and Fu, 2013, Kinnaird et al., 2013, Marlow and Dabbish, 2015, Rashid et al., 2006], and introduce direct communications [Segal et al., 2015] in order to encourage crowd participation.

Despite all these effort on designing intrinsic motivation for the on-demand work, there is one kind of intrinsic motivation that is commonly observed in our daily life from early childhood education to scientific discovery, but has not been explored in the on-demand work settings yet. This intrinsic motivation is *curiosity*, which is defined as “the desire to know, to see, or to experience that motivates exploratory behavior directed towards the acquisition of information” [Litman, 2005]. Therefore, in this chapter, we present a study which examines the potential for curiosity as a “new” type of intrinsic motivational driver to incentivize crowd workers in the on-demand economy. In particular, our study is inspired by *information gap theory* [Loewenstein, 2005], a contemporary model of curiosity which posits that curiosity arises due to a gap between what one knows and what one wants to know. According to this theory, when people are made aware of this gap in their knowledge, they become curious and engage in information-seeking behavior to complete their knowledge and resolve the uncertainty. This innate desire to satisfy one’s curiosity suggests a way to design and structure on-demand work: if tasks can be designed to stoke one’s curiosity, and completing the task provides the requisite information to satisfy that curiosity, then the requesters may be able to create a more enriching experience for workers.

We report results from a set of experiments in which we explicitly incorporate mechanisms

to induce curiosity in workers performing an audio transcription task. Importantly, the curiosity stimuli are related to the task itself, creating a synergy between completing the task and satisfying one’s curiosity. Our results indicate that curiosity can be an effective means of motivating on-demand workers. In particular, we operationalize the concept of curiosity in the task interface design using ideas from information gap theory and show that workers are more likely to complete more tasks, while maintaining a high level of accuracy, when presented with curiosity-inducing stimuli. A closer look at the experimental data further suggests that the magnitude of the effects of curiosity interventions are influenced by both the personal characteristics of the worker and the nature of the task—it is observed that individual workers respond differently to curiosity interventions, and there is also an interaction between curiosity interventions and the task characteristics (e.g., the inherent interestingness of the task), implying that the effects of curiosity interventions are larger for tasks that are less interesting.

7.1 Related Work

Curiosity is an old, yet critical, concept in the psychology of motivation. Various exploratory or information-seeking behaviors have been defined as “curiosity.” For example, animals’ orienting response (i.e., their immediate response to changes in their environment, such as change in illumination or unusual sound) is considered “perceptual curiosity,” whereas humans’ desire for information and knowledge is categorized as “epistemic curiosity” [Berlyne, 1954a]. *Trait* curiosity is a persistent personality attribute, while curiosity aroused by external situations is called *state* curiosity, which is the focus of this study. Curiosity can be triggered by stimuli that are novel (e.g., unexpected changes or violated expectations [Kang et al., 2009]), conflicting (i.e., arousing two or more incompatible responses), uncertain (i.e., leading to outcomes that one is not sure about), and complex (e.g., presenting variety

and diversity) [Berlyne, 1954a]. Although curiosity has been consistently recognized as an important influence on behavior [Loewenstein, 2005], there is no single, agreed-upon model to characterize curiosity’s motivational nature [Silvia, 2014]. Instead, psychologists have proposed a number of theories [Berlyne, 1960, Speilberger and Starr, 1994, Festinger, 1954, Litman, 2005, Berlyne, 1960, 1954a,b, Naylor, 1981, Silvia, 2014] to explain curiosity and people’s information seeking behavior.

The Information Gap Theory

In this study, we focus on one well-established theory of curiosity, Lowenstein’s *information gap theory* [Loewenstein, 2005]. This theory posits that curiosity arises when there is an information gap between what one knows (knowledge baseline) and what one wants to know (information goal). The information goal is subjective, meaning that the same stimuli will arouse different levels of curiosity for each person, depending on the individual’s perception about what she does or does not know.

More formally, information gap theory represents information as a unidimensional concept quantified by an entropy coefficient, $I = -\sum_{i=1}^n p_i \log_2 p_i$. Given this, an individual’s knowledge gap can be measured by the difference between the entropy of the information goal and knowledge baseline. This quantification of information is approximate and serves as a crude proxy that nonetheless provides a way to make predictions about how curiosity might increase or decrease depending on the availability of information, and an individual’s perception of the knowledge gap. Our research adopts a common methodology [Loewenstein, 2005], which only considers ordinal predictions (e.g., “curiosity will increase with information”), and does not attempt to measure the precise magnitude of the information gap.

Using this formulation, curiosity is expected to increase with the accumulation of information, as it creates a sudden shift of attention from focusing on the known (i.e., the existing information) to the unknown (i.e., the missing information). Inspired by approach-gradient

theory [Miller, 1959, Koffka, 1935] and Gestalt psychology [Metzger, 2006], the theory further predicts that the motivation to seek information becomes most intense as one approaches the answer, creating the urge to “complete” the picture. This is used to explain why curiosity is greater for insight problems (where a single piece of information may resolve the entire problem) than for incremental problems (where a single piece of information only provides small progress towards a solution). For example, in a series of experiments, subjects were shown a list of vocabulary words, and asked to evaluate for each word whether they knew the definition, knew the definition only by the tip-of-their-tongue, or did not know the definition [Litman et al., 2005]. Their results showed that individuals were most curious about those tip-of-their-tongue vocabulary words, for which they had some, but not complete, knowledge. In other words, people are unlikely to be curious about information on a topic that they have zero or complete knowledge of, while curiosity is at its height when the information gap becomes quite small, but is not completely closed.

A second implication is that people are more likely to become curious if they have prior knowledge about a particular domain, since a higher knowledge baseline creates a smaller information gap. Indeed, Jones [1979] found that knowledge about a particular domain is correlated with curiosity in that domain, and Berlyne [1954a] found that questions about more familiar entities evoke greater curiosity.

Finally, curiosity requires *attention* to the information gap: to induce one’s curiosity, the information gap must be *salient*, so that the individual can recognize that some information is missing. Also, people will only expose themselves to curiosity-inducing stimuli if there is a non-trivial chance that their curiosity *will be satisfied*, and without long delays. Based on these two implications, Lowenstein suggests that one way to induce curiosity is to ask people to make guesses, which makes the information gap more salient and accurately perceived [Loewenstein, 2005]. The curiosity-inducing designs that we propose and evaluate in this study are based on these key insights.

Practical Applications and Related Constructs of Curiosity

The idea of withholding information to induce the sense of curiosity, which we will describe later in the design of curiosity interventions for the crowd work, have been broadly studied and applied in various settings such as games [Malone, 1981], software engineering [Wilson et al., 2003], interactive designs [Gaver et al., 2003, Tieben et al., 2011], business [Anderson, 2009, Menon and Soman, 2002] and education [Markey and Lowenstein, 2014, Pluck and Johnson, 2011, Schmitt and Lahroodi, 2008, Zion and Sadeh, 2007]. For example, the toy company HotWheels designs “mystery cars” for which the identities are unknown until purchase to boost sales, and the social media company LinkedIn binds the upgrade to the premium account with the reveal of hidden profiles of one’s followers [Anderson, 2009]. Some of the methodologies for stimulating and sustaining curiosity in the classroom include the use of questions, and problem-solving sessions where students each hold partial information, thus requiring them to exchange information in order to accomplish a joint task [Pluck and Johnson, 2011, Schmitt and Lahroodi, 2008, Zion and Sadeh, 2007].

In academic research, curiosity is a distinct construct, though closely related to reinforcement schedules [Gollub, 2001, Zeiler, 1977], as well as goal-setting theories of motivation [Locke and Latham, 2002], which have been studied in several social computing contexts [Beenen et al., 2004, Grevet et al., 2010, Zhu et al., 2012]. The idea of *suspense* [Caplin and Leahy, 2001, Langer, 2014] is a close counterpart to curiosity, with the key difference being the level of emotional engagement with the uncertain outcome.

7.2 Experimental Design

To understand the how curiosity can possibly be used as an intrinsic motivator in the design of on-demand work, we conducted a experimental study on Amazon Mechanical Turk (MTurk). In the following, we describe the operationalization of the concept of curiosity in

task interface design using ideas from the information gape theory, the task that workers were asked to perform in the study, the curiosity interventions that we designed, specific hypotheses driving our study, and various analysis methods we employed.

7.2.1 Operationalizing Curiosity

The central idea in this research is to embed curiosity-inducing designs in task interfaces of the on-demand, crowd work to improve worker engagement and performance. Guiding our designs are the two primary tenets of information gap theory: curiosity can be induced if (1) people are aware of a salient information gap in their knowledge, and (2) people are provided with a means to help them close the gap. Using these constructs as a foundation, we created *curiosity interventions* which consist of three design concepts:

- **Information goal:** To induce curiosity, we create an information gap by posing a *question* that is *relevant* to the current task at hand.
- **Gap salience:** To make workers aware of what they do not know, we prompt them to *guess the answer to the question*.
- **Incremental reveal:** To help workers close the gap, we reveal information as workers progress with their tasks.

We refer to interfaces that employ these three design concepts as *curiosity-inducing stimuli*, and use them to examine the effects that curiosity can have on crowd workers. In this study, we focus exclusively on curiosity-inducing stimuli that support audio transcription tasks, which we will describe in more detail later. Audio transcription task is a good candidate for this study because it is a common type of task on on-demand work platforms that is non-trivial, and at times difficult, for the inexperienced. Hence, effective motivation is especially needed to engage workers. In addition, transcription is a highly *decomposable* task, as an audio clip can be broken down into arbitrarily small units, which enables us to stoke

workers’ curiosity by using each unit as a piece of the “puzzle.” For example, we may reveal one more sentence after each task to complete the story, or one more clue about the identity of the person being discussed in the article. Such techniques are used in the design of our curiosity interventions.

7.2.2 Task and Procedure

The task used in all of our experiments is audio transcription. The audio files were created by a colleague reading an excerpt from articles drawn from a variety of sources (e.g., novels, editorials, news, textbook). Each audio file is cut into 30 individual audio transcription tasks, which vary in length and difficulty. We then combine the 30 transcription tasks for the same audio file in a fixed, but not sorted, order, and bundle them into one HIT (Human Intelligence Task).

In the HIT, workers are asked to perform *at least* three transcription tasks for a base pay of 45 cents, but can quit at any point thereafter by clicking on the “stop now” button, which brings them to a questionnaire. If workers choose to continue after the 3rd task, they earn a 1-cent bonus for completing each additional transcription task. We indicate that all 30 transcription tasks in the HIT come from the same article. A progress bar present on the interface shows how many tasks are remaining. This design is common for studying intrinsic motivation in paid crowdsourcing environment (e.g., as used in [Chandler and Kapelner, 2013]) — that is, by implementing a low payment scheme, we can infer that the reasons for workers to continue are intrinsically, rather than extrinsically, motivated. In addition, we avoid variable payments to minimize confounding motivations such as anticipation for surprise bonuses.

Workers are randomly assigned to one of the experimental conditions (described in the next subsection) as they sign up for the HIT. All our experiments follow a between-subject design, that is, our system ensures that each worker takes our study only once. The HIT is restricted

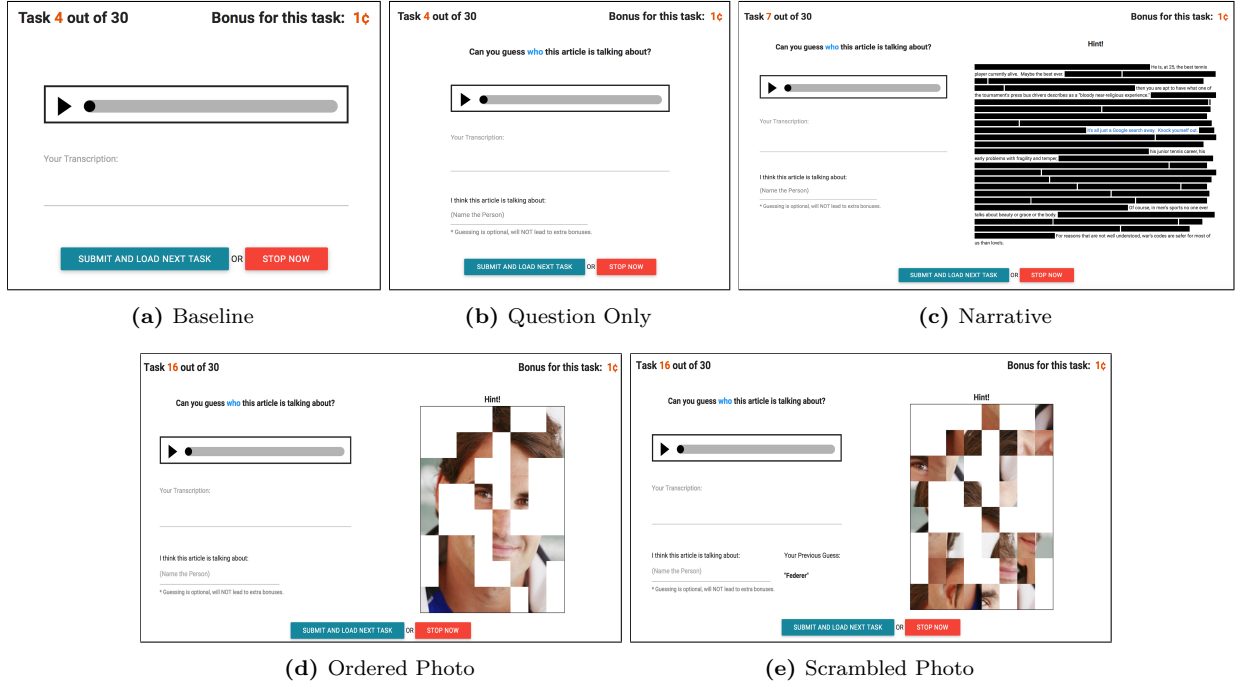


Figure 7.1: Five experimental conditions with varying curiosity interventions

to U.S. workers only. In this work, we are mostly interested in state curiosity and population-level effects; hence, we do not differentiate workers based on personal characteristics or prior experience.

7.2.3 Experimental Conditions: Control and Treatments

As mentioned previously, our curiosity-inducing stimuli consist of three design elements — information goal, gap salience and incremental reveal. By varying *which* of these three design elements are used, we created the following set of control and treatment conditions, as depicted in Figure 7.1.

- **Baseline (Control)**: Workers are *not* presented with any curiosity-inducing stimulus.
- **Question Only**: Workers are presented with a task-relevant question to induce curiosity, and the ability to guess the answer, but without any incremental reveal of information that would provide a hint to the answer.

- **Narrative:** Workers are presented with a task-relevant question, the ability to guess the answer, and a visual representation of the article in which the sentences, except the ones that the worker has transcribed, are obscured (i.e., blacked out).
- **Ordered Photo:** Workers are presented with a task-relevant question, the ability to guess the answer, and a partially obscured photo (e.g., the subject of the article) as a hint to the answer. The picture is divided into a 8×8 grid. Two more cells are revealed after each audio segment is transcribed, with the full photo revealed at the last (i.e., 30th) task.
- **Scrambled Photo:** Workers are presented with a task-relevant question, the ability to guess the answer, and a partially obscured and *scrambled* photo (e.g., the subject of the article) as a hint to the answer. The grid size and method of reveal is the same as in the ordered photo condition.

The task interface provides an audio player and a textbox to support transcription. For all conditions except the baseline, the task-relevant question is present and serves as an information goal. To make the information gap created by this question more salient, the interface includes a textbox for workers to enter guesses, and continually displays their most recent guess. We do not provide any accuracy feedback on the guesses, and only reveal the correct answer when the worker reaches the end. The information that is incrementally revealed in the narrative, ordered photo, and scrambled photo conditions serves to reduce the information gap by providing hints to the answer.

7.2.4 Research Questions and Hypotheses

Our study aims to answer three research questions (Q1—Q3).

Q1: Can crowds be motivated by curiosity? We hypothesize that curiosity interventions will affect worker retention and performance in a positive way:

[H1] Workers will complete more tasks if they are presented with a curiosity-inducing stimulus.

[H2] Workers will have a higher probability of completing all 30 tasks if they are presented with a curiosity-inducing stimulus.

[H3] Workers will have similar or better performance (in terms of work quality) if they are presented with a curiosity-inducing stimulus.

Q2: How do individuals respond differently to curiosity interventions? Lowenstein suggests that guessing draws attention to the knowledge gap and may lead to increased curiosity; hence we are particularly interested in understanding how guessing behavior (e.g., whether workers make a correct guess, incorrect guesses, or no guess at all) correlates with the effects of curiosity interventions. We hypothesize that:

[H4] Workers who make a correct guess will complete more tasks, have a higher probability of completion, and have similar or better performance (in terms of work quality) than workers who make an incorrect guess or make no guesses.

Q3: What are the interactions between task characteristics and curiosity interventions? In particular, when the article to transcribe is inherently interesting in and of itself, workers may be eager to know more about it, even *without* curiosity interventions. In contrast, if the article is not interesting, workers are likely to be indifferent, thus requiring explicit interventions to induce their curiosity. Therefore, our final prediction is:

[H5] The effect of a curiosity intervention is larger when the intervention is combined with a task that is less interesting.

[illegible]

Figure 7.2: Pilot study: Paired comparison of articles

7.2.5 Choice of Article

To study Q1—Q3, our first goal is to select a small number of articles that span a wide range in terms of how inherently interesting they are. We considered five candidate articles drawn from diverse sources — an essay about the famous tennis player Roger Federer, a health article on salt and cholesterol, a blogpost about imposter syndrome, a case study from a marketing textbook, and an excerpt from a novel.

We recruited 98 workers from MTurk to perform a series of 10 paired comparisons of the articles (as depicted in Figure 7.2) as a pilot study. In each task, workers see two articles side-by-side. Each article has all its content blacked out except for 3 randomly chosen sentences. We then ask workers to decide which of the two articles they would like to fully reveal. After 10 rounds of comparisons, we determine the article that each worker wants to reveal the most, and ask her to explain why she finds the article most interesting.

Given each worker’s full ranking of the articles, we assign each article an “interestingness” score, defined as $n_{best} - n_{worst}$, where n_{best} is the number of workers who voted the article to be the one they most want to reveal, and n_{worst} is the number of workers who voted the article to be the one they least want to reveal. Using this formulation, the interestingness of

the imposter syndrome, health and Federer articles are 12, 10 and -22, respectively, indicating that on average, the health and imposter syndrome articles are most inherently interesting, while the Federer article is the least interesting.

An analysis of the textual responses suggests that the reasons for an article being interesting fall under two broad categories: *relevance* and *intrigue*. Some workers found an article to be interesting because it was relevant to them in some way (e.g., “My family has a history of health problems so I would love to read this article,” “I used to work in marketing and am interested in online marketing and branding.”). In contrast, other workers were intrigued by the revealed sentences (e.g., “It hooked me after I read ‘raw energy’. It seemed intriguing,” “I’m curious to know what it’s all about. Who is the boy? Why is he pretending?”, “The sentence started with some mysterious voice the protagonist hears. It kind of made me want to know more about what that voice was about.”).

7.2.6 Analysis Methods

Table 7.1 summarizes the data we use in our analysis, which include worker-initiated actions during the task (e.g., quitting, guessing) as well as responses from the questionnaire. The description column describes the measurement details for the task data and the actual questions/statements associated with the questionnaire data.

In order to capture other factors that may influence retention and performance, we applied a common technique in psychological research to quantify intrinsic motivational factors, and used these factors as covariates in the analysis when appropriate. Specifically, we measure motivational factors using the Intrinsic Motivation Inventory (IMI) [Ryan, 1982], a scale that measures factors related to *enjoyment* (how much workers enjoy transcription), *competence* (how competent workers think they are at transcribing) and *effort* (how much effort workers put into the tasks). As shown in Table 7.1, workers are asked to rate on a 7-point Likert scale about how much they agree with a set of statements related to these three dimensions. For

Task Data	Description
Condition	The experimental condition the worker is assigned to
Quit Index	The number of tasks the worker completes before quitting
Error Rate	The total number of errors the worker makes, divided by the total number of words in the ground-truth transcriptions
Earliest Correct Guess	The number of tasks the worker completes before making a correct guess
Questionnaire Data	Description
Enjoyment	Mean value of Likert scale (1-7) responses for the following statements: <ul style="list-style-type: none"> · <i>This task did not hold my attention at all. (Negative)</i> · <i>While I was doing this task, I was thinking about how much I enjoyed it.</i> · <i>This task was fun to do.</i> · <i>I thought this was a boring task. (Negative)</i>
Competence	Mean value of Likert scale (1-7) responses for the following statements: <ul style="list-style-type: none"> · <i>I think I did pretty well at this task, compared to other workers.</i> · <i>After working at this task for a while, I felt pretty competent.</i> · <i>This is a task that I couldn't do very well. (Negative)</i>
Effort	Mean value of Likert scale (1-7) responses for the following statements: <ul style="list-style-type: none"> · <i>I didn't put much energy into this. (Negative)</i> · <i>I tried very hard on this task.</i> · <i>It was important to me to do well at this task.</i>
Why Quit	Did you choose to stop before reaching the end (i.e., the 30th task)? If so, why?
Why Persist	Did you stop as soon as you first thought of stopping? If not, why did you persist and continue doing more tasks?

Table 7.1: Data Summary

each dimension, we then average the workers' responses (reversing the scores for the negative statements) and use the mean value as the summary statistics for that dimension.

Dependent Variables. Quit index (i.e., how many tasks the worker performed before stopping) is used as our dependent variable to understand the effects of curiosity interventions on worker retention, and error rate (i.e., the percentage of errors the worker made in the transcription tasks, as defined in Table 7.1) is used as our dependent variable to understand the effects of curiosity interventions on worker performance.

Independent Variables. The experimental condition the worker was assigned to and the article that the worker was transcribing serve as the independent variables, i.e., factors that are believed to influence worker retention and performance.

Statistical Methods. For high-level descriptions of worker retention and performance, we use descriptive statistics (e.g., mean, median) when they are appropriate. We also provide histograms and retention curves to visualize the number of workers who quit or remain after each task.

To examine the effects of curiosity interventions on how many tasks a worker completes and how well a worker performs in the tasks, we conduct one-way analysis of variance (ANOVA) or a Kruskal-Wallis test, depending on whether the residuals are normally distributed. Both tests allow us to measure whether there are any statistically significant differences across conditions in terms of the *mean* (or *median*) of the metric that is being examined.

To examine the effects of curiosity interventions on how likely a worker completes all tasks, we first use a proportion test, which allows us to measure statistically significant differences between conditions in the *proportion* of workers who completed all 30 tasks. We apply Bonferroni correction to account for the bias introduced by multiple comparisons.

By treating whether a worker completes all tasks as a binary variable, we further use a generalized linear model (GLM) [McCullagh and Nelder, 1989] with logit link function (also known as the logistic regression model) to model the probability (or odds) of completing all 30 tasks in different conditions. Specifically, to compare the completion probability in conditions with curiosity-inducing stimuli against the baseline condition, we set the baseline condition as the reference. We also control for the influences of other intrinsic motivational factors, such as enjoyment, competence and effort, in our GLM — as an ANOVA on each of these factors suggests that there is no significant difference in them across experimental conditions, we feel comfortable to include them as covariates in our regression model. The fit of each GLM is assessed graphically using the residual plots and other quantitative approaches, including the Hosmer and Lemeshow test [Hosmer and Lemeshow, 1980] and Osius-Rojek test [Osius and

Rojek, 1992], two common goodness-of-fit tests for logistic regression models.¹

7.3 Results

7.3.1 Effects of Curiosity Interventions on the Crowd

We start our analyses from answering our first research question, that is, can crowd be motivated by curiosity? Since the Federer article generated the least interest, we selected this article as the worst case scenario to understand whether curiosity interventions affect worker retention and performance. Accordingly, in our experiment, the task-relevant question is “Can you guess who this article is talking about?”, and the photo we use in the ordered photo and scrambled photo conditions is the photo of Roger Federer. We recruited 100 workers for each condition, and gathered data from a total of 500 participants. We found 4 workers who were obvious spammers and filtered out these suspicious cases, leaving 496 workers for further analysis.

Effects on Worker Retention

To understand how various curiosity interventions affect worker retention in the transcription tasks, we first plot a histogram showing the number of workers who quit after a certain number of tasks across five experimental conditions (Figure 7.3). To further illustrate the difference in retention, Figure 7.4 shows the the retention curves of each condition plotted against the baseline.

As can be seen in Figure 7.3 and Figure 7.4, in general, the majority of the workers either chose to quit when they completed fewer than half of all transcription tasks (i.e., quit before the 15th task), or kept working until they completed all tasks (i.e., quit after completing

¹Statistically significant results are reported as follows: $p < 0.001(***)$, $p < 0.01(**)$, $p < 0.05(*)$, $p < 0.1(\cdot)$.

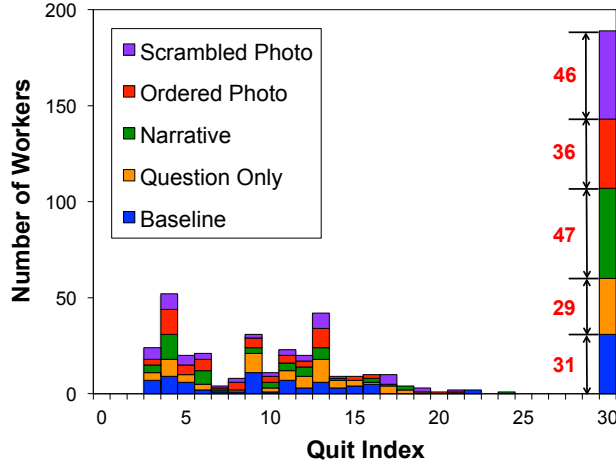


Figure 7.3: Histogram showing the number of workers who quit after completing X tasks in each experimental condition.

the 30th task). Table 7.2 further shows the median, mean, and standard deviation for the number of completed tasks in each experimental condition. The disagreement between means and medians as well as the large standard deviation again indicate the two-sided skew in the distributions of the number of completed tasks, for which measures of center are not the best characterizations. Nevertheless, we still find that workers in experimental conditions with curiosity-inducing stimuli tend to complete a larger number of tasks compared to workers in the baseline condition, which is consistent with our prediction in H1. ANOVA shows that the effects of curiosity interventions on the number of completed tasks is marginally significant, $F(4, 491) = 2.17$, $p = 0.07$.

To take a closer look, we first focus on workers who decided to quit before they were halfway through the whole HIT. We find many of them actually quit after they completed the 4th, 9th or 13th task. Interestingly, the sentences that the workers were asked to transcribe in the 5th, 10th and 14th task are among the longest and most complex sentences in the whole article, implying that workers may have decided to quit because they were deterred by the difficulty of the transcription tasks.

It appears that one of the benefits conferred by presenting curiosity-inducing stimuli on

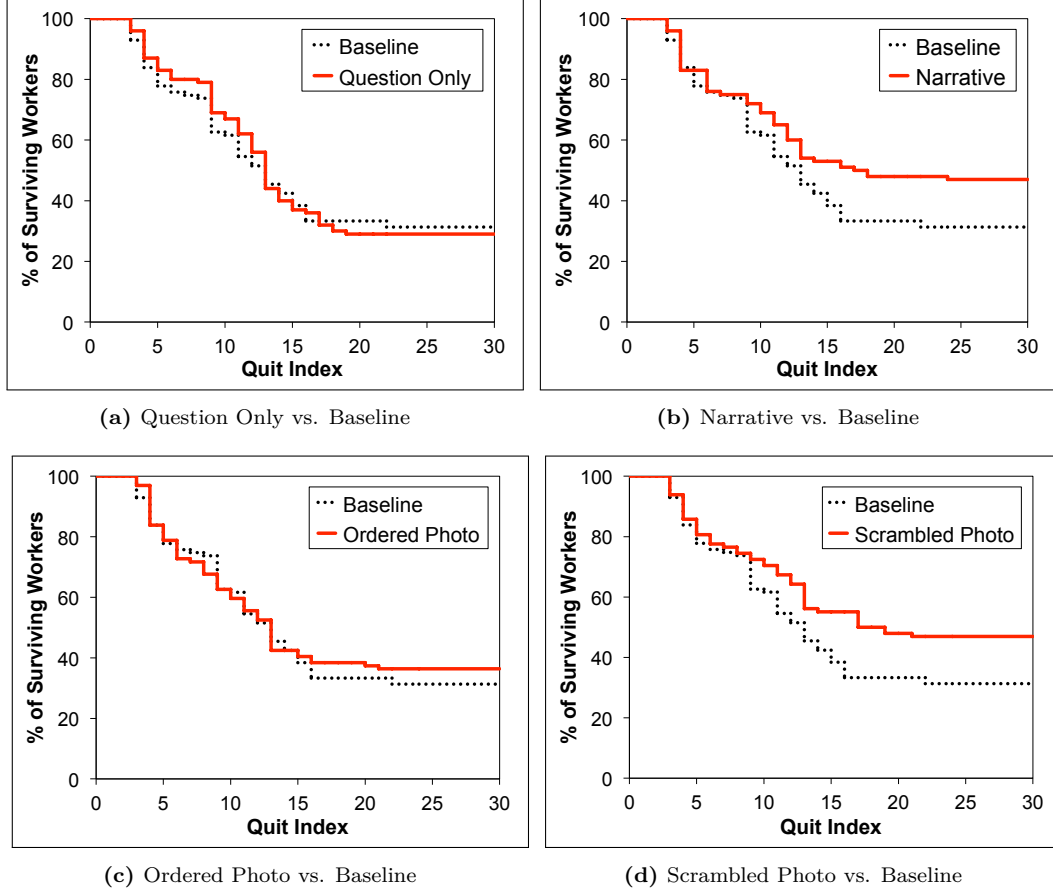


Figure 7.4: Retention curves showing the number of workers who continue working on the tasks (i.e. “survive”) after completing X tasks.

task interfaces is that it nudges workers to persist through difficult spots in the task sequence, to get to the *tipping point* where they feel the urge to complete the whole HIT. Indeed, as shown in Figures 7.4b and 7.4d, while 11% of the workers in the baseline condition quit after the 9th task, only 2–3% of workers in the narrative and scrambled photo conditions did so. The same pattern was not observed in the question only or ordered photo conditions.

Next, we shift our focus to workers who completed all tasks to understand the effects of curiosity interventions on completion (H2). Table 7.2 shows the percentage of workers who completed all 30 tasks in each condition. Compared to the baseline, most other conditions have a larger percentage of workers who completed all 30 tasks, with the question only condition being the only exception. Two-sided proportion test results suggest that

Conditions	Median	Mean (SD)	Completion Rate
Baseline	13	15.86 (10.38)	31%
Question Only	13	15.97 (9.76)	29%
Narrative	17.5	18.85 (11.14)	47%
Ordered Photo	13	16.45 (10.88)	36%
Scrambled Photo	18	19.03 (11.04)	47%

Table 7.2: Summary of quit index in different conditions.

the difference in the percentage of completion across conditions is statistically significant, $\chi^2(4, N = 496) = 12.56, p = 0.02$, yet none of the pairwise comparisons against the baseline condition is statistically significant under the multiple comparison test with Bonferroni correction.

We get a better understanding of the effects of different curiosity interventions on completion using a generalized linear model (GLM), by taking into account the influences of other intrinsic motivational factors such as enjoyment, competence and effort. Our GLM is a reasonable fit as we find no obvious pattern when the residuals are plotted against the independent variables and fitted values, and the results for both Hosmer and Lemeshow test and Osious-Rojek test also support our graphical assessment, $\chi^2(8, N = 496) = 5.88, p = 0.66$ and $z = 0.86, p = 0.39$, respectively.

Table 7.3 reports the results for the GLM. Here, we find that when a curiosity-inducing stimulus is present, workers are more likely to complete all 30 tasks in all cases (i.e., the estimated coefficient $\hat{\beta}$ is positive) compared to workers in the baseline condition. Holding all other explanatory variables constant, workers in the narrative (and scrambled photo) condition have a significantly higher estimated odds to complete all tasks that is $e^{0.98} = 2.65$ (and $e^{0.88} = 2.41$) times as large as that for workers in the baseline condition, while the increase of estimated odds to complete all tasks in the question only and ordered photo condition is not statistically significant. Interestingly, we also find that enjoyment and self-reported competence both have a significant impact on the likelihood of completing all the tasks — the more a worker enjoys the tasks and/or feels competent at the tasks, the

Variable	Model Parameters				
	$\hat{\beta}$	Std. Error	t	p -value	
Question Only	0.10	0.32	0.31	0.75	
Narrative	0.98	0.31	3.10	1.91×10^{-3}	**
Ordered Photo	0.47	0.32	1.49	0.14	
Scrambled Photo	0.88	0.31	2.81	4.91×10^{-3}	**
Enjoyment	0.19	0.10	1.91	0.06	.
Effort	0.02	0.11	0.21	0.83	
Competence	0.52	0.11	4.85	1.24×10^{-6}	***

Table 7.3: GLM for the probability of completing all 30 tasks in each condition, with the baseline condition being the reference.

more likely she will complete all tasks, which is quite intuitive.

Why Quit? Why Persist?

By looking into the questionnaire data, where workers explain their reasons for quitting and persisting despite initial urges to quit, we gain more insights into the diverse factors that may contribute to whether and how our curiosity interventions influence worker retention.

In general, we find a few major categories of reasons for quitting: *low payment* (“The pay is low for the time it was consuming”), *task difficulty* (“I chose to stop because there came a point where I was having difficulty with understanding some of what was being said, and I didn’t want to transcribe incorrectly”), *lack of engagement* (“I was bored and did not know who they were talking about”), and *external factors* (“I was interrupted by someone at my door”).

As for persisting to work, the major reasons cited are *payment* (“I wanted to make more money from the bonuses”), *learning* (“It was good practice, I felt I was getting better as I went on”), and a *completionist attitude* (“I don’t like leaving things half-finished”).

Importantly, we notice that when curiosity-inducing stimuli are present, many workers actually cite *curiosity* as their primary reason for continuing to work on the tasks. For example:

- “I wanted to see if I could figure out the player’s name.”
- “I persisted because I was curious about how the article would unfold.”
- “I was addicted to transcribing the next sentence to reveal the article’s subject, and once I knew who the article’s subject was, I just wanted to complete the article.”
- “I wanted to know who the article is about. It was like getting a puzzle piece and putting it all together.”
- “I thought of stopping several times, but my desire to do a good job, earn the maximum bonus, and to be frank, my curiosity kept me going.”

In other words, our curiosity interventions are shown to be effective on many workers as they become eager to find out the answer to the question, and they also take the satisfaction of their curiosity into the consideration when making the cost-benefit analysis on whether to continue or not.

Furthermore, the lessening or lack of curiosity plays a role in why workers *quit*. Some workers cited their inability to guess the answer to the question (“I had no idea who it could be and stopped caring”) or their certainty about the answer (“I thought I had already figured out who it was talking about and didn’t want to transcribe any more”) as reasons for quitting. In other words, when curiosity cannot be satisfied or if curiosity dies, workers would end up quitting. This is in line with Loewenstein’s observations [[Loewenstein, 2005](#)] — the accumulation of new information may cause the information goal (what one wants to know) to change, or the objective value of the missing information to decrease (because one can infer the answer), thus diminishing curiosity. In other words, curiosity increases with information, but curiosity may also die as the gap is dynamically re-adjusted in light of new information.

Effects on Worker Performance

The median error rates in baseline, question only, narrative, ordered photo and scrambled photo conditions are 4.29%, 3.91%, 3.55%, 3.96% and 3.54%, respectively. While workers seem to perform better in experimental conditions with curiosity-inducing stimuli, the Kruskal-Wallis test results suggest that the difference in error rate across conditions is actually not statistically significant, $\chi^2(4, N = 496) = 6.85, p = 0.14$ ². Our findings on worker performance supports hypothesis H3, that is, workers are able to maintain a high level of accuracy when presented with curiosity-inducing stimuli. Meanwhile, we also notice that the error rate in all experimental conditions is already *very low*, which implies that the space for performance improvement can be very limited. Further examination on task duration shows that each transcription task takes around 1 minute to complete on average, and there are no statistically significant difference across conditions, $F(4, 491) = 0.52, p = 0.72$.

7.3.2 Individual Differences in Reaction to Curiosity Interventions

Next, we move on to examine how do individuals respond to curiosity interventions differently. The information gap theory predicts that making guesses leads to increased curiosity. Hence, in this study, we are particularly interested in understanding the differences in worker’s guessing behavior when curiosity interventions are presented and their connection to worker retention and task performance. Specifically, we make comparisons across three groups of workers in experimental conditions with curiosity-inducing stimuli: workers who made *correct* guesses to the question, workers who made *incorrect* guesses, and workers who made no guess. Since the correct answer to the question (i.e., the word “Federer”) is revealed at the 25th sentence for the first time, we restrict our attention to guessing behavior before that task.

²An one-way ANOVA is not suitable here due to the non-normally distributed residuals.

Metrics		Correct Guess	Incorrect Guess	No Guess
Quit Index	Median	30	11	12
	Mean (σ)	22.63 (9.84)	13.60 (10.03)	12.57 (8.69)
Completion	%	62	24	15
Error Rate	Median	0.03	0.05	0.04
	Mean (σ)	0.04 (0.04)	0.06 (0.04)	0.05(0.03)

Table 7.4: Retention and performance by guessing behavior.

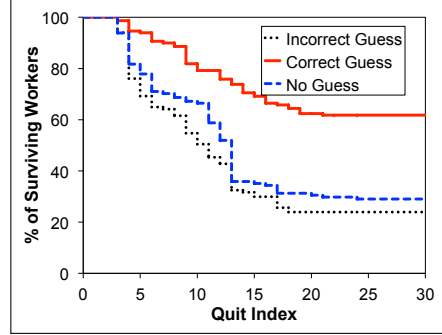


Figure 7.5: Retention curves: correct, incorrect and no guesses.

Among 397 workers, 149 workers made a correct guess, 117 made incorrect guesses and 131 did not make a guess. Table 7.4 and Figure 7.5 report the comparison across these three groups of workers. Results show that curiosity interventions affect workers who made correct guesses the most, as they complete significantly more tasks ($t(247) = 7.35, p = 2.82 \times 10^{-12}$ and $t(270) = 6.04, p = 5.16 \times 10^{-9}$ with Bonferroni correction), are significantly more likely to complete *all* tasks ($\chi^2(1, N = 266) = 36.33, p = 3.32 \times 10^{-9}$ and $\chi^2(1, N = 266) = 28.74, p = 8.30 \times 10^{-8}$ with Bonferroni correction), and are more accurate in their transcriptions ($t(116) = -14.63, p = 4.4 \times 10^{-16}$ and $t(130) = -17.15, p = 4.4 \times 10^{-16}$ with Bonferroni correction) than workers who made incorrect guesses or no guess.

A variety of reasons can account for these differences. First, we observe that correct guessers made their first guesses (median at task 1) as well as their first correct guesses (median at task 5) quite early, suggesting that they might have certain prior knowledge about the tennis player Federer and hence were more curious. Second, correct guessers self-reported higher levels of competence in the questionnaire than incorrect guessers ($t(236) = 3.15, p = 0.002$),

implying that they had more confidence in their performance than incorrect guessers. Our conjecture is that the effects of curiosity interventions can vary depending on individual worker’s level of prior knowledge as well as actual and perceived competence.

7.3.3 Interactions between Task Characteristics and Curiosity Interventions

Finally, we look into how the effects of curiosity interventions are influenced by task characteristics, such as the inherent interestingness of the tasks. Our previous experiment shows the effects of curiosity interventions when the task itself is *not* interesting (i.e., the Federer article is the least interesting article according to our pilot study). We now repeat our experiment for the health and imposter syndrome articles — the two articles that generated most interests among workers in our pilot study — to understand how the effects of our curiosity interventions may differ. In particular, we include three experimental conditions for each of these two articles: baseline, question only and narrative. The task-relevant questions we pose for the health and imposter syndrome articles are “Is salt good or bad for you?” and “What psychological condition is the article talking about?”, respectively.

We recruited 100 workers per condition per article and randomly assigned them to one of the three experimental conditions. As in the previous experiment, workers are asked to complete at least 3 tasks to get the base payment of 45 cents. After the 3rd task, workers may choose to stop at any time or complete more tasks in exchange for an extra 1-cent bonus per task. We explicitly prevent workers in our previous experiment (who worked on transcribing the Federer article) from participating in this experiment, and each worker is only allowed to take this experiment once.

Figure 7.6 illustrates the number of workers who quit after a certain number of tasks for the three articles. Visually, we can see that the proportion of workers who completed all the

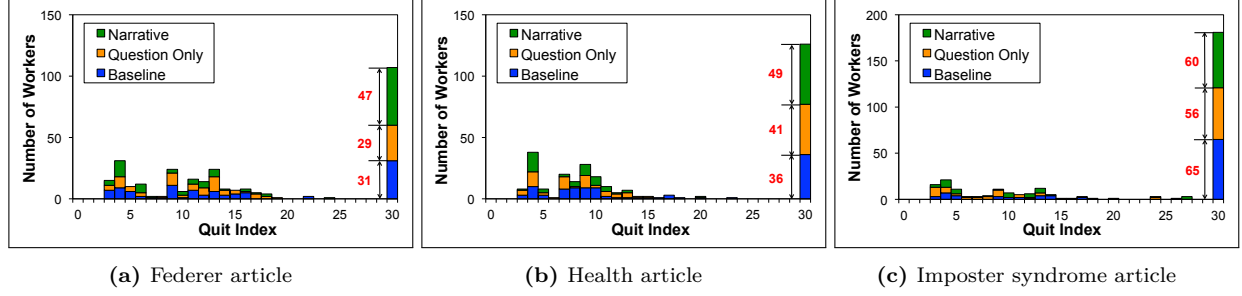


Figure 7.6: Histogram showing the number of workers who quit after completing X tasks: variation between different tasks (i.e. different articles).

tasks is much higher for the health and imposter syndrome articles. In addition, for these two inherently interesting articles, the completion rates across experimental conditions appear to be similar.

To validate our visual intuition, we combine our data in all three tasks and again create a GLM with logit link function to model the probability of completing all tasks, with the baseline condition and the Federer article as the reference. In particular, we use task (i.e., the article used for the transcription) as an independent variable, and the interaction terms between tasks and curiosity interventions are also included in the model. Significant interaction terms then imply that the effects of curiosity interventions depend on the tasks. Both visual examination on the residual plots and the two goodness-of-fit test results suggest that our model is a reasonable fit. Results for this GLM is shown in Table 7.5.

As expected, compared to workers performing the less interesting task (i.e., transcribing the Federer article), workers in more interesting tasks are more likely to complete all tasks, even *without* explicit curiosity interventions (i.e., the estimated coefficient $\hat{\beta}$ is positive for both the imposter syndrome article and the health article), and this difference is statistically significant for the task of transcribing the imposter syndrome article, $t(881) = 4.27, p = 1.93 \times 10^{-5}$. Furthermore, while introducing curiosity interventions improves the probability of completion in the question only and narrative conditions for the Federer article (i.e., in both cases, the estimated coefficient $\hat{\beta}$ is positive), almost all the estimated coefficients for interaction

Variables	Model Parameters				
	$\hat{\beta}$	Std. Error	t	p -value	
Question Only	0.09	0.32	0.27	0.79	
Narrative	0.95	0.31	3.07	2.16×10^{-3}	**
Health Article	0.12	0.32	0.38	0.71	
Imposter Article	1.33	0.32	4.27	1.93×10^{-5}	***
Question Only \times Health Article	0.35	0.44	-0.79	0.43	
Narrative \times Health Article	-0.29	0.43	-0.68	0.50	
Question Only \times Imposter Article	-0.18	0.44	-0.42	0.68	
Narrative \times Imposter Article	-1.10	0.43	-2.53	0.01	*
Enjoyment	0.19	0.07	2.50	0.01	*
Effort	0.01	0.08	0.17	0.87	
Competence	0.47	0.08	6.02	1.76×10^{-9}	***

Table 7.5: Generalized linear model: 3 articles

terms between tasks and curiosity interventions are *negative*, showing that the curiosity interventions are more effective for inherently uninteresting tasks, which is consistent with our hypothesis H5.

7.4 Discussion

In this chapter, through an experimental study, we highlight that stimulating curiosity *can* be an effective way to incentivize the on-demand, crowd workers. Specifically, we find a close relationship between curiosity and worker retention and performance — given curiosity interventions, workers completed more tasks, while maintaining a high level of performance. Furthermore, we find that workers who made a correct guess complete a significantly larger number of tasks with significantly higher quality than those who made incorrect guesses or no guess. Finally, the effects of our curiosity interventions also depend on the characteristics of the tasks; namely, the effects are larger when the interventions are introduced to tasks that are less interesting.

Why Are Some Interventions More Effective?

Some interventions (e.g., narrative and scrambled photo) seem to be more effective than other interventions (e.g., question only and ordered photo) in incentivizing crowd workers. In the question only condition, while we create an information goal and make the information gap salient by allowing workers to make guesses, we do *not* explicitly provide any information to help workers answer the question. Thus, the fact that question only is not as effective as other interventions indicates that simply setting an information goal, without providing the means for workers to satisfy their curiosity in the interim, is not an effective approach.

A related question is why the ordered photo condition is not that effective. We found that workers in the ordered photo condition attempted to make their first guesses *and* figured out the answer to the question much earlier than workers in other conditions – the medians are 2, 5 and 4.5 for the number of tasks completed before the first guess, and 5, 21 and 24 for the number of tasks completed before the first correct guess for the ordered photo, narrative and scrambled photo conditions respectively. There are also more workers who made a correct guess (68%) in the ordered photo condition than any other conditions (i.e., 43%, 58% and 54% for the question-only, narrative and scrambled photo condition). That is to say, compared to other conditions, workers in the ordered photo condition are given too much information that enables them to satisfy their curiosity early on, making the effects of the curiosity interventions there virtually not any different from the question only condition, where workers are not given any information at all. This reflects the subtlety in designing curiosity interventions as information can be a double-edged sword: Too little information is not enough for inducing curiosity, yet too much information could satisfy one’s curiosity too soon and diminish curiosity. In general, both personal characteristics (e.g., prior knowledge) and beliefs (e.g., how certain one is about the answer) can influence the effectiveness of curiosity interventions.

Design Space and Generalizability

Our curiosity interventions consist of three design elements — a question that is answerable by the information contained in a series of tasks (which serves as the information goal), a mechanism for eliciting guesses (which increases the salience of the information gap), and an incremental reveal of information (which closes the information gap as workers do more tasks). This conceptual approach is quite general, and is applicable to any setting in which the employer/requester has some basic knowledge about the task data. For example, the question can be about (i) a feature/property shared by some or all tasks, e.g., the neighborhood that a set of images depict, (ii) the global picture of how the individual tasks fit together, e.g., a design task where the identity of the larger system is not revealed until all the submodules are completed, (iii) some global statistics computed from individual data points gathered from the tasks, e.g., a counting task where participants assess the number of flowers on a herbarium specimen, each of which contributes a data point towards testing the hypothesis “flowering time is becoming earlier over time due to effects of climate change.” The idea is to obscure by hiding or scrambling certain information that is incrementally revealed as tasks are being completed, and by making the obfuscation salient such that people notice and become curious about the missing information.

Our work demonstrates that information gap theory can be operationalized to affect the behavior of crowd workers. However, the generalizability of the specific designs that we explored in this study, i.e., the use of scrambled photo and obscured article, is limited to transcription tasks and situations where the answer to the question has a visual representation (e.g., the article is about Roger Federer).

In practice, the design elements are *knobs* that can be tuned — the particular choice of questions, feedback mechanisms for responding to guesses, or the frequency of information reveal all have subtle impacts on the extent to which workers feel curious. Additional *knobs* include: (1) questions that reveal a different *type* of information, such as social comparison

statistics (e.g., “how much do your transcriptions agree with other workers”); (2) the number of curiosity stimuli to present, that is, whether we create one primary information gap or bite-sized information gaps that are routinely revealed and satisfied (e.g., giving workers a new puzzle to solve when they guessed the answer correctly); (3) the amount and complexity of information to present (e.g., varying the number of cells to reveal and the extent to which the photo is scrambled); and (4) whether or not to provide regular accuracy feedback about the guesses to further increase the attention to the knowledge gap. It is thus an interesting future work to develop techniques that automatically create curiosity-inducing stimuli by tuning design parameters so that we may not only induce but also sustain curiosity, or even adapt curiosity interventions to account for individual differences in knowledge and interest.

Ethical Implications

Prior work has explored non-monetary mechanisms to motivate workers on paid on-demand work platforms like crowdsourcing markets, e.g., by offering micro-diversions [Dai et al., 2015] or providing an altruistic purpose for the task [Chandler and Kapelner, 2013, Shaw et al., 2011, Ariely et al., 2008, Rogstadius et al., 2011]. On the one hand, one can argue that these intrinsic motivators (e.g., enjoyable activities during breaks, a meaningful purpose, the desire to read the entire article) serve as *extra payment*, that is, workers are rewarded with a valuable experience. On the other hand, these mechanisms for increasing the intrinsic motivation of *extrinsically motivated workers* raise ethical concerns — workers may, unknowingly, be doing more work for less pay. In this study, we chose task-relevant questions as curiosity stimuli because we want workers to feel a sense of engagement with the task at hand. Nevertheless, curiosity interventions would find a more natural home in volunteer-based on-demand work settings such as citizen science projects, where the goal is to stoke people’s curiosity about science in addition to collecting data to facilitate discoveries.

7.5 Acknowledgements

The work in this chapter was produced in collaboration with Edith Law, Joslin Goh, Kevin Chen, Michael Terry and Krzysztof Z. Gajos. I contributed in designing and conducting the experiment, analyzing the results and the paper writing. Portions of this chapter previously appeared in the CHI publication *Curiosity Killed the Cat, but Makes Crowdwork Better* [Law et al., 2016].

Chapter 8

Conclusion

The rise of the on-demand economy in the past few years has led to dramatic changes in our society—it creates new business patterns which enable an efficient and direct matching between supply and demand; it pushes the boundaries of the modern computing and in particular, artificial intelligence technologies; it expands scientific discovery and expedites scientific advancement remarkably. While the on-demand economy has demonstrated its practical importance and potential with its rapid growth in numerous domains around the globe, scientific understandings and rigorous design principles for it are only in their infancy. Many people still perceive the on-demand economy as a black-box approach to soliciting labor from a crowd of workers in an on-demand manner, without much idea about how it works or how it can work better.

This dissertation opens up the black box of on-demand economy, both to obtain a fundamental understanding of what happens behind the scenes, and to explore effective interventions and techniques to make it better. Investigations in this dissertation are conducted through a particular platform—Amazon Mechanical Turk, which is one of the leading and most widely used on-demand crowdsourcing platforms.

8.1 Summary of Contributions

In Chapters 2, 3 and 4, I focus on understanding the on-demand economy of today, with an emphasis on understanding who the crowd of on-demand workers are and how they behave in work. Chapter 2 investigates the temporal dynamics of the crowd of on-demand workers, not only in terms of how the demographic composition of the crowd changes over time but also with respect to the variations in the economic behavior, cognitive abilities and styles, and personality that workers who are available at different times of day exhibit. These results are especially relevant to scientific researchers who conduct crowd-based studies on on-demand platforms, as the observed temporal dynamics implies that when researchers launch studies on these platforms at different times, they may in fact approach to sub-populations of subjects with significantly different characteristics and may even obtain different results for their studies. Chapter 3 reveals a substantial communication network hidden inside the crowd of on-demand workers. Being able to recover this communication network helps us to depict a typical working scene for a significant portion of on-demand workers, which is in stark contrast to stereotype impressions: instead of working independently, these workers intensively interact with a large number of “co-workers” through both online discussion forums and other one-on-one channels so that they are effectively working in a collaborative community. Importantly, such communication may confer some informational advantage to workers allowing them to find out valuable work earlier, as well as providing emotional support to help workers through the ups and downs in their daily work. Chapter 4 challenges the common perception of on-demand work being fully flexible, and identifies worker’s desire for more flexibility in on-demand work, especially more freedom to decide how to allocate time within an individual task and across different tasks. In fact, with more flexibility being provided in the on-demand work, workers are able to efficiently schedule their workload so as to work at their own pace, which leads to higher levels of engagement and performance. Furthermore, the majority

of on-demand workers are willing to forego some financial compensations for the ability to control their own time in on-demand work, implying the significant values that workers attach to the flexibility of on-demand work.

In Chapters 5, 6 and 7, I study possible improvements to the design of on-demand work to enable a more efficient and sustainable on-demand economy in the future. I approach this problem from the perspective of devising effective incentives for on-demand work. Chapter 5 shows how financial incentives can be used in a most effective way to elicit high performance from on-demand workers given the presence of certain psychological biases of workers. For example, in a sequence of tasks of the same type, the increase of incentive magnitude over subsequent tasks matters more for motivating workers compared to the absolute magnitude of incentive in each individual task, as workers may anchor their perception of appropriate payment levels on the incentive that they first receive in the sequence. In settings where tasks of different types interleave with each other, placing financial incentives at the switching points where task types change leads to the largest improvement in worker performance as it helps to reduce the switch cost to a large degree. Chapter 6 highlights an algorithmic framework to guide requesters to dynamically decide whether and when to offer extra monetary rewards in a session of on-demand tasks, in order to encourage high-quality work while taking the financial cost of rewards into consideration. The two major building blocks of this algorithmic framework is a quantitative model to predict work quality under monetary intervention and an online planning algorithm to make near-optimal decisions under uncertainty. The feasibility of algorithmically controlling financial incentives in an on-demand work environment is also showed on real on-demand task sessions for the first time. Finally, Chapter 7 explores the potential of incorporating curiosity as a new type of intrinsic motivator in the on-demand work through clever designs of task interfaces. Inspired by the information gap theory of curiosity, three key elements of this task interface design—information goal, gap salience and incremental information reveal—are proposed to create synergy between completing the

on-demand work and satisfying one’s curiosity. It is showed that interfaces employing these curiosity-inducing design elements lead to improved worker engagement without degrading performance, indicating curiosity as an effective means to motivate on-demand workers.

8.2 Connections between Chapters

While each chapter in this dissertation presents one or more independent studies, there are many notable connections between chapters, both within each of the two components of the dissertation and between them. Connecting perspectives and results in different chapters together, therefore, provides unique opportunities for examining the on-demand economy from a more integrated view, which can help us to better summarize our existing findings as well as identify interesting directions for further study.

For example, taking results in Chapters 2 and 4 together leads to a more comprehensive view of how flexible workers are in the on-demand economy. On the one hand, the observed temporal dynamics of on-demand workers in Chapter 2, especially in terms of their demographic compositions at different times of day, is a clear indicator suggesting that on-demand workers have *some* degree of flexibility. On the other hand, results in Chapter 4 show that the level of flexibility on-demand workers have is *not enough*, especially in terms of the ability to control their own time within each task. As an analogy, current on-demand workers are like employees in traditional jobs who enjoy an extreme version of flextime, so they can determine when to work and how long to work at their own will. However, once they start to work, each task that they work on imposes a tight deadline on them, making it not uncommon for workers to “rush for deadlines” as they have little freedom to schedule the work in a way that they find most comfortable. Results in these two chapters, therefore, urge us to think about what levels of flexibility can be afforded with this new form of on-demand work, and to what degree we can free workers from the “micro-management” in the work and provide them with

more control of their own work. One important element to consider in working towards this goal of empowering on-demand workers with more flexibility is to cultivate among workers a “habit” of leveraging the flexibility. To that end, encouraging workers to communicate with one another, either through online forums or other one-on-one channels (Chapter 3), may be helpful, because workers can share with each other their experience and best practices on how to efficiently schedule work given the extra flexibility provided in the work.

As another example, one may conjecture that social interactions among on-demand workers may have played a role in shaping the temporal dynamics of the crowd when considering Chapters 2 and 3 together. One possible scenario is that a small group of workers has a “collective” working schedule so that each worker of the group works at the same time, possibly in the same physical space. As different groups may have different schedules (e.g., an east-coast group has a different schedule than a west-coast group), it is natural to expect that on-demand workers who are available at different times of day are partly composed of different subsets of these worker groups. For workers who decide their own working schedules, being able to communicate with other workers through online forums can also give them a sense of virtual social community. As a result, these workers may tend to work at times when the other workers that they are familiar with are also online, again contributing to the temporal dynamics of the crowd. Moreover, the observed temporal differences in worker’s economic behavior may also be related to the connections among workers. Take worker’s incentivized decisions in the public goods game as an example. One possibility is that how much a worker is willing to put in the public account in a game is partly decided by how much the worker cares about the welfare of other workers at that time, which can be largely influenced by the degree of familiarity among workers of that time. Even more, it is also possible that workers of certain time slots can actually initiate intensive discussions among themselves about their strategies in the game so as to coordinate their actions and generate more social welfare in the game. So clearly, connecting Chapters 2 and 3 together opens up a

set of new directions to explore for verifying each of these conjectures.

Chapters in the second component of the dissertation, the incentive design for on-demand work, are also interconnected. The empirical understandings of worker’s reactions to financial incentives (Chapter 5) provide the basis for the algorithmic framework of dynamically controlling financial incentives in on-demand work (Chapter 6). One of the most important insights that experimental studies in Chapter 5 offer is that the worker performance in a task is not only decided by the incentive in this task, but also incentives in some of its surrounding tasks in the work session. Such insight guides us to quantitatively characterize the effects of financial incentives on workers in the context of a task workflow, rather than for individual tasks in Chapter 6. Different from the extrinsic, financial incentives discussed in Chapters 5 and 6, Chapter 7 focuses on the design of intrinsic motivation like curiosity, but it still reveals some similar observations. We find that when designed appropriately, incentives—be it extrinsic or intrinsic—can be used to motivate higher effort and performance from the *population* of on-demand workers (Chapters 5 and 7). In the meantime, it is also worthwhile to note that *individuals* respond to incentives in different ways (Chapters 6 and 7). For example, as shown in Chapter 6, some workers can consistently submit high quality work in a task even without extra financial incentives while others need the additional rewards to perform well; and in Chapter 7, it is found that curiosity interventions are most effective for workers who can make correct guesses to the question that is presented on the task interface and intended to induce curiosity. These observations on the heterogeneity in worker’s reaction to various incentives suggest that in the future, instead of seeking for an one-size-fits-all incentive to motivate all workers, perhaps a more effective approach is to identify what motivates one the most for each worker and provide a personalized motivation, tailored to the needs of each individual.

Results in the first (i.e., understanding crowd behavior, Chapters 2, 3, 4) and second component (i.e., design extrinsic and intrinsic incentives, Chapters 5, 6 and 7) of the

dissertation mutually influence each other as well. More specifically, various characteristics of the crowd behavior that are identified in the dissertation suggest new perspectives to consider when designing incentives for on-demand work in the future.

For instance, the observed temporal dynamics of the crowd in Chapter 2 offers a new angle for examining the incentive design in on-demand work: Are the effects of various incentives the same throughout a day? Can it be more effective to adopt different incentive mechanisms at different times of day? For example, consistent trends have been identified in the change of people’s mood throughout a day¹. Thus, it can be interesting to investigate whether and how interactions between the emotion of on-demand workers and incentives vary at different times of day.

Knowledge on social interactions among workers (Chapter 3) presents both opportunities and challenges for the incentive design of on-demand work. On the one hand, knowing that there are connections between workers opens a set of new possibilities in structuring incentives. One possibility is to reward workers not only for their own contribution in a task but also for routing tasks to other workers with necessary skill sets or information [Zhang et al., 2012]. Another option is that instead of decomposing complex work into micro-tasks and assign them to multiple (and possibly unrelated) workers, requesters may allow a group of interconnected workers to accept the complex work together and reward them as a team. On the other hand, since workers can talk to each other, cautions should be used when providing incentives to individual workers in different ways (e.g., the dynamic bonus policy presented in Chapter 6). This is because workers may feel being treated unfairly if they compare with each other the rewards they receive, and workers may even collectively reverse-engineer the underlying mechanism for incentive provision and attempt to game the system. It is therefore important to take into account the ethical justification and strategy-proofness of a

¹See <http://www.ccs.neu.edu/home/amislove/twittermood/> for an example.

personalized incentive mechanism when designing it.

Lastly, Chapters 5, 6 and 7 explore how effective extrinsic and intrinsic incentives can be devised given the current common practice of task design, that is, a number of tasks are organized into a work session while workers don't have much control over their working time in the session. As showed in Chapter 4, granting workers with more flexibility in each task of the session leads to higher worker engagement and performance and thus may lower the requirement for additional extrinsic or intrinsic motivation needed in the work. However, with more flexibility, workers may switch back and forth among many different work sessions to optimally schedule their workload, and the possible interference between work sessions leaves the effects of extrinsic and intrinsic incentives in one particular work session unclear. Therefore, further research is needed to thoroughly study whether flexibility and incentives complement or impede each other, and how requesters can effectively incentivize workers in a session while taking worker's desire for more flexibility into consideration.

8.3 Future Directions

Taken together, chapters in this dissertation present a close examination at the on-demand economy from two perspectives—understanding crowd behavior and designing effective incentives. While findings in this dissertation provide fresh views and redesign ideas, they have only scratched the surface of a new area of study that calls for a comprehensive understanding of on-demand economy as well as a thorough exploration of its vast design space, and plenty of questions remain open.

First, this dissertation mainly approaches the problem of understanding the on-demand economy from the perspective of understanding the workers (i.e., the supply of labor) in it, using a particular on-demand platform—Amazon Mechanical Turk—as an example. A more comprehensive view of the on-demand economy, however, should span all different parties

in it, including workers, requesters (or customers, i.e., the demand of labor) and platforms which can come with different formats ranging from online labor marketplaces to mobile apps. Therefore, there are many interesting research questions one can ask about requesters and platforms in the on-demand economy. For example, who are the requesters in the on-demand economy? Why do they choose to use on-demand labor? How much do they rely on the on-demand labor? How do they allocate their work to and interact with workers? Answering these questions will lead to a more precise picture of the demand of labor in on-demand economy. In the meantime, various platforms in the on-demand economy have been observed to play different roles: some platforms (e.g., Uber) actively match the supply and demand, while others simply provide a common virtual space where demand can meet with supply. The latter can be further divided into subcategories depending on whether the exchange of labor is driven by the demand (e.g., Amazon Mechanical Turk, where requesters post tasks and workers accept) or centered around the supply (e.g., Fiverr, where freelancers list their skills and customers search for the ones suitable for their needs). It is thus interesting to develop a taxonomy of on-demand platforms and to understand the commonalities and differences between different platforms: How does the setup of a platform shape the behavior of workers and requesters on it? What kind of on-demand work fits a particular type of platform well? What are the unique challenges for each type of platforms?

Perhaps an even broader question to ask is with this rapid growth of on-demand economy, where are we going next? Without a doubt, the on-demand economy has commenced a transformation of work which affects the types of work we do and the very basics of how work is organized. Over ten years ago, [Malone \[2004\]](#) argued in his book *The Future of Work* that the development in information technology had pushed down the communication cost dramatically, which enabled businesses and organizations to adopt a “decentralized” structure with loose hierarchies and more democracies, leveraging outsourcing to scale, and setting up internal markets for information aggregation. I would like to argue that with the rise of

on-demand economy, we may envision a future where work is not necessarily associated with businesses or organizations. Much higher liquidity may be observed in various aspects in the future of work, such as for whom people work, of whom “businesses and organizations” are composed, and even the change of an individual’s role between being the supply or the demand of labor. More specifically, instead of working for one particular employer for years, workers in the future may switch between employers at a much faster pace, or even work for multiple employers at the same time. Likewise, businesses and organizations, which are formed (probably in a virtual sense through digital approaches) to fulfill some goals of the requesters of labor, may no longer keep the same structures and groups of workers all the time, but adaptively adjust and update according to the needs of work. Work may be presented in the form of “projects.” A small group of workers who are familiar with each other’s expertise and get along well may work as a team and work on projects from different requesters together from time to time. Requesters may assemble and reassemble their workforces from both individual workers and worker teams to fulfill the goal of each project. Moreover, a worker in one project could easily start her own “business” and act as a requester, hiring other workers to help her complete the project. In some sense, the future of work may share many similarities with how the film industry works today, but only with much shorter cycles and more geographically distributed, digitally coordinated, and rapidly adjusting.

There are many challenges in realizing such a vision. For workers to smoothly transit from one work to another, a reliable, cross-platform reputation system is needed to signal the skills and quality of a worker. Computational methods can be helpful in helping workers reduce the search cost for finding suitable work (e.g., through task recommendation), quickly navigate the new work (e.g., get familiar with their responsibilities and their “co-workers” in one project), and effectively improve skills or expand skill sets over time. Innovative mechanisms are required for supporting the development of workers at different stages of their “careers” (e.g., a new worker with very few reputation records vs. a well-established worker with

excellent reputation) as well as worker’s transition between on-demand work and traditional jobs. On the other hand, software and tools need to be developed to allow requesters to monitor the status of all their projects or quickly make changes to their management in each project as needed (e.g., [Retelny et al., 2014, Valentine et al., 2017]). For requesters who are in their “startup” stage to get a quick start, modules of work and templates of workflows from other requesters with similar goals could be provided.

But perhaps one of the most central problems that requesters concern is how to design the work in a way to attract and retain workers, especially in this future work environment where workers can easily come and go. To this end, it can be particularly valuable to revisit numerous topics on work design in industrial psychology, organization psychology and management literature. Lessons from these literature can provide guidance on how various concepts and theories for work design in the traditional economy can be applied in the on-demand work, as well as how the limitations of traditional models can be reduced and what kind of new models can be afforded by the new form of work. At the core of on-demand work design, however, lies the deep concern for *humanity* and respect for *human values*. The ultimate question requesters may want to ask in thinking about the design of their work is what matters the most to their potential workers *as people*. Findings in this dissertation have provided some answers, such as social interactions, flexibility and autonomy, and motivation. Many others are awaited to be explored, including recognition, creativity, feeling of competence, and sense of purpose.

And for on-demand platforms as well as the entire society, special attention should be paid to a number of ethical and policy challenges raised by the on-demand economy to ensure a sustainable future of work. Notable ones include guarantees of fair income for on-demand workers, improvement in the transparency of work (e.g., what’s the actual purpose for a particular on-demand task?), clarification on intellectual properties (e.g., who owns the intellectual property for software developed by the crowd?), protection in privacy (e.g., how

user data on on-demand platforms are stored?), and the reinvention of a social security and benefit system that adapts to the new ways of working.

For computer scientists, there could not be a better time to contribute their wisdom in shaping the future of work. First of all, the rise of on-demand economy offers an unprecedented opportunity for data science research. For example, computational social scientists can use large-scale online experimentation to *collect the right data* through on-demand platforms to understand worker behavior and study work design. Results of these experiments may further reflect certain human needs, values and biases that can be generalized to many other contexts. Meanwhile, the huge amount of data accumulated in on-demand economy over the years also allow data scientists to *reason and learn from the existing data*, such as identifying distinctive behavior patterns, analyzing organizational structures and processes, and providing recommendations and decision supports for workers and requesters. Moreover, by eliciting human intelligence through on-demand platforms, computer scientists have achieved remarkable progress in artificial intelligence, which may eventually free humans from boring, repetitive work and leave them with more creative, challenging tasks. To facilitate progress towards this goal, better knowledge is required for fully utilizing the “human intelligence”—what else humans can help with AI other than contributing their basic human knowledge and judgment like determining whether a cat is included in a picture? Promising candidates include human intuition [Kim et al., 2017], human expectations, and human values [Bonnefon et al., 2016]. And when computers can really complete some tasks on behalf of humans, one may envision a future in which artificial intelligence and human labor are integrated seamlessly—the work in the future may be jointly completed by a huge cluster of computers, some on-demand labor, and perhaps only a small group of in-house employees who are experts in certain domains. By that time, coordinating computers with humans of mixed expertise will present new challenges to computer scientists, such as how to automatically divide the labor among different parties, how to control the workflow in an

adaptive way and how to effectively share information in a team of machines and humans.

In summary, the advent of on-demand economy has brought us with many changes, and there are yet more to come. It is estimated that in 2015, 8% of Americans have earned money by serving as on-demand labor on online platforms [Smith, 2016a]. It is predicted that following the current growth rate, by 2027, nearly 1 in 3 American adults will transition to online platforms to support themselves with on-demand work [Suri and Gray, 2016]. If this is real, we'd better to be prepared now, starting from peeking into the black box of on-demand economy to understand how it works, and to explore how it can work better.

Bibliography

- J. Stacy Adams. Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, 67(5):422, 1963.
- Icek Ajzen. *Attitudes, personality, and behavior*. McGraw-Hill Education (UK), 2005.
- George A. Akerlof. Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4):543–569, 1982.
- George A. Akerlof and Janet L. Yellen. The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics*, 105(2):255–283, 1990.
- Alan Allport and Glenn Wylie. Task-switching: Positive and negative priming of task-set. In G. Humphreys, J. Duncan, and A. Treisman, editors, *Attention, Space, and Action: Studies in Cognitive Neuroscience*, pages 273–296. Oxford University Press, Oxford, 1999.
- D. A. Allport, E. A. Styles, and S. Hsieh. Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà and M. Moscovitch, editors, *Attention and Performance XV*, pages 421–452. MIT Press, Cambridge, MA, 1994.
- Teresa M. Amabile, William DeJong, and Mark R. Lepper. Effects of externally imposed deadlines on subsequent intrinsic motivation. *Journal of Personality and Social Psychology*, 34(1):92, 1976.
- Ofra Amir, David G. Rand, et al. Economic games on the internet: The effect of \$1 stakes. *PloS One*, 7(2):e31461, 2012.
- S. Anderson. Applying curiosity to interaction design: Tell me something I don’t know, 2009. Last Retrieved January 7, 2016 from <http://johnnyholland.org/2009/08/curiosity-and-interaction-design/>.
- Antonio A. Arechar, Gordon T. Kraft-Todd, and David G. Rand. Turkling overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. 2016.
- D. Ariely, E. Kamenica, and D. Prelec. Man’s search for meaning: The case of Legos. *Journal of Economic Behaviour and Organization*, 67(3):671–677, 2008.

- D. Ariely, U. Gneezy, G. Loewenstein, and N. Mazar. Large stakes and big mistakes. *Review of Economic Studies*, 76(2):451–469, 2009.
- Miriam Lueck Avery, Cindy Baskin, Rod Falcon, Eri Gentry, Alex Goldman, Ben Hamamoto, Sara Skvirsky, and Kathi Vian. Voices of workable futures: People transforming work in the platform economy. 2016.
- Brian P. Bailey and Joseph A. Konstan. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, 22(4):685–708, 2006.
- Boris B. Baltes, Thomas E. Briggs, Joseph W. Huff, Julie A. Wright, and George A. Neuman. Flexible and compressed workweek schedules: A meta-analysis of their effects on work-related criteria. *Journal of Applied Psychology*, 84(4):496–513, 1999.
- Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2):241–251, 2001.
- Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, pages 212–221. ACM, 2004.
- Yoshua Bengio and Paolo Frasconi. An input output HMM architecture. *Advances in Neural Information Processing Systems*, pages 427–434, 1995.
- Yoshua Bengio and Paolo Frasconi. Input-output HMMs for sequence processing. *Neural Networks, IEEE Transactions on*, 7(5):1231–1249, 1996.
- Y. Benkler. Coase’s Penguin, or, Linux and the Nature of the Firm. *Yale Law Journal*, 112(3):367–445, 2002.
- Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. Evaluating online labor markets for experimental research: Amazon. com’s Mechanical Turk. *Political Analysis*, 20(3):351–368, 2012.
- D. Berlyne. A theory of human curiosity. *British Journal of Psychology*, 45:180–191, 1954a.
- D. Berlyne. An experimental study of human curiosity. *British Journal of Psychology*, 45:256–265, 1954b.
- D. Berlyne. *Conflict, Arousal and Curiosity*. McGraw-Hill, London, 1960.
- Truman F. Bewley. Fairness, reciprocity, and wage rigidity. *Behavioral Economics and Its Applications*, pages 157–188, 2007.

- Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.
- Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21:375–395, 1999.
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.
- Samuel Bowles. Policies designed for self-interested citizens may undermine “the moral sentiments”: Evidence from economic experiments. *Science*, 320(5883):1605–1609, 2008.
- Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. Using gamification to inspire new citizen science volunteers. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, pages 18–25. ACM, 2013.
- Jonathan Bragg and Daniel S. Weld. Optimal testing for crowd workers. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, 2016.
- Adrian Bridgwater. Machine learning needs a human-in-the-loop, 2016. Last Retrieved March 7, 2016 from <https://www.forbes.com/sites/adrianbridgwater/2016/03/07/machine-learning-needs-a-human-in-the-loop>.
- Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.
- Ronald S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2): 349–99, September 2004.
- Ronald S. Burt. Second-hand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management Journal*, 50:119–148, 2007.
- Colin F. Camerer and Robin M. Hogarth. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1-3):7–42, 1999.
- A. Caplin and J. Leahy. Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, pages 55–79, 2001.
- Logan S. Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z. Strolovitch. Intertemporal differences among MTurk worker demographics. 2017.
- D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behaviour and Organization*, 90:123–133, 2013.

- Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A. Ratliff. Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7):1131–1139, 2015.
- G. B. Chapman and E. J. Johnson. The limits of anchoring. *Journal of Behavioral Decision Making*, 7(4):223–242, 1994.
- Alain Cohn, Ernst Fehr, and Lorenz Goette. Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8):1777–1794, 2014.
- C. Randall Colvin. “Judgable” people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology*, 64(5):861, 1993.
- S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and Foldit Players. Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760, August 2010.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- Giovanni Costa. Shift work and occupational medicine: An overview. *Occupational Medicine*, 53(2):83–88, 2003.
- Sergio Currarini, Matthew O. Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77:1003–1045, 2009.
- Peng Dai, Daniel Sabey Weld, et al. Decision-theoretic control of crowd-sourced workflows. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- Peng Dai, Jeffrey M. Rzeszutarski, Praveen Paritosh, and Ed H. Chi. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 628–638. ACM, 2015.
- Edward Deci, R. Koestner, and R.M Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):692–700, 1999.
- Edward L. Deci and Richard M. Ryan. Motivation, personality, and development within embedded social contexts: An overview of self-determination theory. *The Oxford Handbook of Human Motivation*, pages 85–107, 2012.
- Xuefei Deng, KD Joshi, and Robert D. Galliers. The duality of empowerment and marginalization in microtask crowdsourcing: Giving voice to the less powerful through value sensitive design. *MIS Quarterly*, 40(2):279–302, 2016.

- Xuefei Nancy Deng and KD Joshi. Is crowdsourcing a source of worker empowerment or exploitation? Understanding crowd workers' perceptions of crowdsourcing career. 2013.
- Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*, pages 238–247. ACM, 2015.
- Pinar Donmez, Jaime G. Carbonell, and Jeff G. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *SDM*, volume 2, page 1. SIAM, 2010.
- Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1013–1022. ACM, 2012.
- Martin Dufwenberg and Georg Kirchsteiger. Reciprocity and wage undercutting. *European Economic Review*, 44(4):1069–1078, 2000.
- David Easley and Arpita Ghosh. Incentives, gamification, and game theory: An economic approach to badge design. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pages 359–376. ACM, 2013.
- David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
- Diana Farrell and Fiona Greig. Paychecks, paydays, and the online platform economy: Big data on income volatility. *JP Morgan Chase Institute*, 2016.
- Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.
- L. Festinger. A theory of social comparison processes. *Human Relations*, 7:117–140, 1954.
- Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th International Conference on World Wide Web*, pages 333–343. ACM, 2015.
- Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001.
- Torsten Hothorn Frank Bretz and Peter Westfall. *Multiple Comparisons Using R*. Chapman & Hall/CRC, 2011.
- Shane Frederick, George Loewenstein, and Ted O'donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002.

- Bruno S. Frey and Reto Jege. Motivation crowding theory: A survey of empirical evidence. *Journal of Economic Surveys*, 15(5):589–611, 2001.
- Yihan Gao and Aditya Parameswaran. Finish them!: Pricing algorithms for human computation. *Proceedings of the VLDB Endowment*, 7(14):1965–1976, 2014.
- William W. Gaver, Jacob Beaver, and Steve Benford. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 233–240. ACM, 2003.
- Sabine AE Geurts and Evangelia Demerouti. Work/non-work interface: A review of theories and findings. *The Handbook of Work and Health Psychology*, 2:279–312, 2003.
- Sam J. Gilbert and Tim Shallice. Task switching: A PDP model. *Cognitive Psychology*, 44(3):297–337, 2002.
- Duncan S. Gilchrist, Michael Luca, and Deepak Malhotra. When $3 + 1 > 4$: Gift structure and reciprocity in the field. *Management Science*, 2016.
- U. Gneezy and A. Rustichini. Pay enough or don’t pay at all. *The Quarterly Journal of Economics*, 115(3):791–810, 2000.
- Uri Gneezy and John A. List. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384, 2006.
- Daniel G. Goldstein, Siddharth Suri, R Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. The economic and cognitive costs of annoying display advertisements. *Journal of Marketing Research*, 51(6):742–752, 2014.
- L. Gollub. Information on conditioned reinforcement. *Journal of the Experimental Analysis of Behaviour*, pages 361–372, 2001.
- Mary L. Gray and Siddharth Suri. The humans working behind the AI curtain, 2017. Last Retrieved January 9, 2017 from <https://hbr.org/2017/01/the-humans-working-behind-the-ai-curtain>.
- Mary L. Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. The crowd is a collaborative network. In *The 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2016.
- Catherine Grevet, Jennifer Mankoff, and Scott D. Anderson. Design and evaluation of a social visualization aimed at encouraging sustainable behavior. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–8. IEEE, 2010.
- Neha Gupta, David Martin, Benjamin V. Hanrahan, and Jacki O’Neil. Turk-life in India. In *The International Conference on Supporting Groupwork (Group)*, 2014.

- J Richard Hackman and Edward E. Lawler. Employee reactions to job characteristics. *Journal of Applied Psychology*, 55(3):259, 1971.
- J Richard Hackman and Greg R Oldham. Development of the job diagnostic survey. *Journal of Applied Psychology*, 60(2):159, 1975.
- J. Richard Hackman and Greg R. Oldham. Work redesign. 1980.
- David J. Hardisty, Katherine F. Thompson, David H. Krantz, et al. How to measure time preferences: An experimental comparison of three methods. *Judgment and Decision Making*, 8(3):236, 2013.
- Christopher Harris. You’re 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)*, pages 15–18, 2011.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- E. Jeffrey Hill, Alan J. Hawkins, Maria Ferris, and Michelle Weitzman. Finding an extra day a week: The positive influence of perceived job flexibility on work and family life balance. *Family Relations*, 50(1):49–58, 2001.
- Paul Hitlin. Research in the crowdsourcing age, a case study, 2016. Last Retrieved July 11, 2016 from <http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/>.
- Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pages 359–376. ACM, 2014.
- 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*, pages 419–429. ACM, 2015.
- Tin Kam Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.
- Charles A. Holt, Susan K. Laury, et al. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, 2002.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

- John J. Horton, David G. Rand, and Richard J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.
- John Joseph Horton and Lydia B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, pages 209–218. ACM, 2010.
- David W. Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*, 9(10):1043 – 1069, 1980.
- Eric Huang, Haoqi Zhang, David C. Parkes, Krzysztof Z. Gajos, and Yiling Chen. Toward automatic task design: A progress report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85. ACM, 2010.
- 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*, pages 621–630. ACM, 2013.
- Connor Huff and Dustin Tingley. “Who are these people?” evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648, 2015.
- Intuit. How the on-demand economy is reshaping the 40-hour work week, 2017. <http://investors.intuit.com/press-releases/press-release-details/2016/How-the-On-Demand-Economy-Is-Reshaping-the-40-hour-Work-Week/default.aspx>.
- Panagiotis G. Ipeirotis. Demographics of Mechanical Turk. Technical Report CeDER-10-01, March 2010.
- Panos Ipeirotis. Demographics of mechanical turk: Now live! (April 2015 edition), 2015. <http://www.behind-the-enemy-lines.com/2015/04/demographics-of-mechanical-turk-now.html>.
- Shamsi T. Iqbal and Eric Horvitz. Disruption and recovery of computing tasks: Field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 677–686. ACM, 2007.
- 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)*, 2013.
- Patricia A. Jacobs and Peter AW Lewis. Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36, 1983.

- G.D. Jenkins Jr, A. Mitra, N. Gupta, and J.D. Shaw. Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, 83(5):777, 1998.
- Li-Jun Ji, Zhiyong Zhang, and Richard E. Nisbett. Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology*, 87(1):57, 2004.
- Oliver P. John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2(1999): 102–138, 1999.
- S. Jones. Curiosity and knowledge. *Psychological Reports*, 45:639–642, 1979.
- Kerry Joyce, Roman Pabayo, Julia A. Critchley, and Clare Bambra. Flexible working conditions and their effects on employee health and wellbeing. *The Cochrane Library*, 2010.
- Hyun Joon Jung and Matthew Lease. A discriminative approach to predicting assessor accuracy. In *Advances in Information Retrieval*, pages 159–171. Springer, 2015a.
- Hyun Joon Jung and Matthew Lease. Modeling temporal crowd work quality with limited supervision. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015b.
- Hyun Joon Jung, Yubin Park, and Matthew Lease. Predicting next label quality: A time-series model of crowdwork. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.
- Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.
- Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Fairness and the assumptions of economics. *Journal of Business*, pages S285–S300, 1986.
- Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives*, 5(1): 193–206, 1991.
- Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- M. Kang, M. Hsu, I. Krajbich, G. Loewenstein, S. McClure, J. Wang, and C. Camerer. The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, 20(8):963–973, 2009.

- David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, pages 1953–1961, 2011.
- Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. Control and interference in task switching — A review. *Psychological Bulletin*, 136(5):849, 2010.
- Joseph Kim, Christopher J Banks, and Julie A Shah. Collaborative planning with encoding of users’ high-level strategies. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.
- Peter Kinnaird, Laura Dabbish, Sara Kiesler, and Haakon Faste. Co-worker transparency in a microtask marketplace. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1285–1290. ACM, 2013.
- Jon Kleinberg, Siddharth Suri, Éva Tardos, and Tom Wexler. Strategic network formation with structural holes. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC)*, 2008.
- K. Koffka. *Principles of Gestalt Psychology*. Harcourt-Brace, New York, 1935.
- William Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- K. Langer. Suspenseful Design: Engaging Emotionally with Complex Applications through Compelling Narratives. Master’s thesis, University of Waterloo, Waterloo, Ontario, 2014.
- Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4098–4110. ACM, 2016.
- Edward P. Lazear. Performance pay and productivity. *The American Economic Review*, 90(5):1346–1361, 2000.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Semi-local affine parts for object recognition. In *British Machine Vision Conference (BMVC’04)*, pages 779–788. The British Machine Vision Association (BMVA), 2004.
- Christopher Lin, Mausam, and Daniel Weld. Dynamically switching between synergistic workflows for crowdsourcing. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI ’12, 2012.
- J. Litman. Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition and Emotion*, 19(6):793–814, 2005.

- J. Litman, T. Hutchins, and R. Russon. Epistemic curiosity, feeling-of-knowing, and exploratory behavior. *Cognition and Emotion*, 19(4):559–582, 2005.
- Michael L. Littman, Anthony R. Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 1995.
- Lennart Ljung. *System Identification*. Springer, 1998.
- Edwin A. Locke and Gary P. Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705, 2002.
- Edwin A. Locke, Karyll N. Shaw, Lise M. Saari, and Gary P. Latham. Goal setting and task performance: 1969–1980. *Psychological Bulletin*, 90(1):125, 1981.
- G. Loewenstein. The psychology of curiosity: A review and reinterpretation. *Cognition and Emotion*, 19(6):793–814, 2005.
- T. Malone. Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 4: 333–369, 1981.
- Thomas W. Malone. The future of work: How the new order of business will shape your organization, your management style and your life. 2004.
- Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? Predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- Gloria Mark, Daniela Gudith, and Ulrich Klocke. The cost of interrupted work: More speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 107–110. ACM, 2008.
- A. Markey and G. Lowenstein. Curiosity. In R. Pekrun and L. Linnenbrink-Garcia, editors, *International Handbook of Emotions in Education*. Rutledge, New York, 2014.
- Jennifer Marlow and Laura A. Dabbish. The effects of visualizing activity history on attitudes and behaviors in a peer production context. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 757–764. ACM, 2015.
- David Martin, Benjamin V. Hanrahan, Jacki O’Neil, and Neha Gupta. Being a turker. In *The 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2014.
- Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, March 2012.
- Winter Mason and Duncan J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

- U. Mayr and R. Kliegl. Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 26(5):1124–1140, 2000.
- Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172. ACM, 2013.
- Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 2 edition, 1989.
- Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
- S. Menon and D. Soman. Managing the power of curiosity for effective web advertising strategies. *Journal of Advertising*, 31(3):1–14, 2002.
- Anthony J. Mento, Robert P. Steel, and Ronald J. Karren. A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes*, 39(1):52–83, 1987.
- W. Metzger. *Laws of Seeing*. MIT Press, Cambridge, MA, 2006.
- N. Miller. Liberalization of basic S-R concepts: Extensions to conflict behavior, motivation and social learning. *Psychology: The Study of a Science*, pages 196–292, 1959.
- N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The Peer-Prediction Method. *Management Science*, 51:1359–1373, 2005.
- Phyllis Moen, Erin L. Kelly, Wen Fan, Shi-Rong Lee, David Almeida, Ellen Ernst Kossek, and Orfeu M. Buxton. Does a flexibility/support organizational initiative improve high-tech employees’ well-being? Evidence from the work, family, and health network. *American Sociological Review*, 81(1):134–164, 2016a.
- Phyllis Moen, Erin L. Kelly, Shi-Rong Lee, J Michael Oakes, Wen Fan, Jeremy Bray, David Almeida, Leslie Hammer, David Hurtado, and Orfeu Buxton. Can a flexibility/support initiative reduce turnover intentions and exits? Results from the work, family, and health network. *Social Problems*, page spw033, 2016b.
- Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, 2003.
- F. Naylor. A state-trait curiosity inventory. *Australian Psychology*, 16:172–183, 1981.
- Sander Nieuwenhuis and Stephen Monsell. Residual costs in task switching: Testing the failure-to-engage hypothesis. *Psychonomic Bulletin & Review*, 9(1):86–92, 2002.

- Hylco H. Nijp, Debby GJ Beckers, Sabine AE Geurts, Philip Tucker, and Michiel AJ Kompier. Systematic review on the association between employee worktime control and work-non-work balance, health and well-being, and job-related outcomes. *Scandinavian Journal of Work, Environment & Health*, pages 299–313, 2012.
- Gerhard Osius and Dieter Rojek. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(140):1145 – 1152, 1992.
- Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5:411–419, 2010.
- Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.
- Yubin Park, Carlos Carvalho, and Joydeep Ghosh. Lamore: A stable, scalable approach to latent vector autoregressive modeling of categorical time series. In *AISTATS*, pages 733–742, 2014.
- G. Pluck and H. Johnson. Stimulating curiosity to enhance learning. *GESJ: Education Sciences and Psychology*, 2(19):1512–1801, 2011.
- Ioana Popescu and Yaozhong Wu. Dynamic pricing strategies with reference effects. *Operations Research*, 55(3):413–429, 2007.
- D. Prelec. A Bayesian truth serum for subjective data. *Science*, 306:462–466, 2004.
- Robert D. Pritchard and Michael I. Curts. The influence of goal setting and financial incentives on task performance. *Organizational Behavior and Human Performance*, 10(2): 175–183, 1973.
- M. Raddick, G. Bracey, P. Gay, C. Lintott, C. Cardamone, P. Murray, K. Schawinski, A. Szalay, and J. Vandenberg. Galaxy zoo: Motivations of citizen scientists. *Astronomy Education Review*, 12(1), 2013.
- David G. Rand. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299:172–179, April 2011.
- Al M Rashid, Kimberly Ling, Regina D Tassone, Paul Resnick, Robert Kraut, and John Riedl. Motivating participation by displaying the value of contribution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 955–958. ACM, 2006.
- Vikas C. Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13:491–518, 2012.

- Katharina Reinecke and Krzysztof Z. Gajos. Labyrinthwild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1364–1378. ACM, 2015.
- Peter J. Rentfrow, Samuel D. Gosling, Markus Jokela, David J. Stillwell, Michal Kosinski, and Jeff Potter. Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, 105(6):996, 2013.
- Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S. Bernstein. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User Interface Software and Technology*, pages 75–85. ACM, 2014.
- Robert D. Rogers and Stephen Monsell. Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2):207, 1995.
- J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *AAAI Conference on Weblogs and Social Media*, 2011.
- Joel Ross, Lilly Irani, M. Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI’10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM, 2010.
- Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) ’12, pages 217–226, 2012.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- R. Ryan. Control and information in the interpersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 42:450–461, 1982.
- Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54–67, 2000a.
- Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68, 2000b.

- Jeffrey M. Rzeszotarski and Aniket Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 13–22. ACM, 2011.
- Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We are Dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 2015.
- Dario D. Salvucci and Niels A. Taatgen. *The Multitasking Mind*. Oxford University Press, 2010.
- Dario D. Salvucci, Niels A. Taatgen, and Jelmer P. Borst. Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1819–1828. ACM, 2009.
- Mehrnoosh Sameki, Danna Gurari, and Margrit Betke. Predicting quality of crowdsourced image segmentations from crowd behavior. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- Neil Savage. Gaining wisdom from crowds. *Communications of the ACM*, 55(3):13–15, 2012.
- Noam Scheiber. How uber uses psychological tricks to push its drivers’ buttons, 2017. <https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html>.
- F. Schmitt and R. Lahroodi. The epistemic value of curiosity. *Education Theory*, 8:125–148, 2008.
- Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.
- A. Segal, R. Simpson, Y. Gal, V. Homsy, M. Heartwood, K. Page, and M. Jirotko. Improving productivity in citizen science through controlled intervention. In *WWW*, pages 331–337, 2015.
- Avi Segal, Ya’akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. In *Proceedings of the 25th International Conference on Artificial Intelligence*, 2016.
- Aaron D. Shaw, John J. Horton, and Daniel L. Chen. Designing incentives for inexperienced human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 275–284. ACM, 2011.

- J. Shirk, H. Ballard, C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, and R. Bonney. Public participation in scientific research: A framework for intentional design. *Ecology and Society*, 17(2):29–48, 2012.
- Kristen M. Shockley and Tammy D. Allen. When flexibility helps: Another look at the availability of flexible work arrangements and work–family conflict. *Journal of Vocational Behavior*, 71(3):479–493, 2007.
- P. J. Silvia. Curiosity and motivation. In R. M. Ryan, editor, *Oxford Handbook of Motivation*. Oxford University Press, New York, 2014.
- Yaron Singer and Manas Mittal. Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1157–1166. International World Wide Web Conferences Steering Committee, 2013.
- Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1167–1178. International World Wide Web Conferences Steering Committee, 2013.
- Aaron Smith. Gig work, online selling and home sharing, 2016a. <http://www.pewinternet.org/2016/11/17/gig-work-online-selling-and-home-sharing/>.
- Rebecca Smith. Flexibility and the on-demand economy. 2016b.
- Mark Snyder. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of Personality*, 51(3):497–516, 1983.
- Cheri Speier, Joseph S. Valacich, and Iris Vessey. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2):337–360, 1999.
- C. Spielberger and L. Starr. Curiosity and exploratory behavior. *Motivation, Theory and Research*, pages 221–243, 1994.
- Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, September 2015.
- J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643, 1935.
- Siddharth Suri. Technical perspective: Computing with the crowd. *Commun. ACM*, 59(6):101–101, May 2016. ISSN 0001-0782. doi: 10.1145/2927926. <http://doi.acm.org/10.1145/2927926>.

- Siddharth Suri and Mary L. Gray. Spike in online gig work: Flash in the pan or future of employment?, 2016. <https://socialmediacollective.org/2016/11/17/spike-in-online-gig-work-flash-in-the-pan-or-future-of-employment/>.
- Rob Tieben, Tilde Bekker, and Ben Schouten. Curiosity and interaction: Making people curious through interactive systems. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, pages 361–370. British Computer Society, 2011.
- Arthur N. Turner and Paul R. Lawrence. *Industrial Jobs and the Worker*. Harvard Univ., 1965.
- A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.
- Melissa A. Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S. Bernstein. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3523–3537. ACM, 2017.
- Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.
- L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.
- Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.
- Luis von Ahn. Duolingo: Learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 1–2. ACM, 2013.
- Jing Wang and Panagiotis Ipeirotis. Quality-based pricing for crowdsourced workers. 2013.
- Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer-Verlag, Berlin, 2003.
- Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.
- Duncan J. Watts and Steven H. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393(6684):440–442, 1998.
- Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

- Aaron Wilson, Margaret Burnett, Laura Beckwith, Orion Granatir, Ledah Casburn, Curtis Cook, Mike Durham, and Gregg Rothermel. Harnessing curiosity to increase correctness in end-user programming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 305–312. ACM, 2003.
- Glenn Wylie and Alan Allport. Task switching and the measurement of “switch costs”. *Psychological Research*, 63(3-4):212–233, 2000.
- Ming Yin and Yiling Chen. Bonus or not? Learn to reward in crowdsourcing. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- Ming Yin and Yiling Chen. Predicting crowd work quality under monetary interventions. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.
- Ming Yin, Yiling Chen, and Yu-An Sun. The effects of performance-contingent financial incentives in online labor markets. In *Conference on Artificial Intelligence, AAAI ’13*, pages 1191–1197, 2013.
- Ming Yin, Yiling Chen, and Yu-An Sun. Monetary interventions in crowdsourcing task switching. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.
- 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*, pages 1293–1303. International World Wide Web Conferences Steering Committee, 2016.
- M. Zeiler. Schedules of reinforcement. In W.K. Hong and J.E.R. Staddon, editors, *Handbook of Operant Behaviour*. Prentice Hall, Englewood Cliffs, NJ, 1977.
- Haoqi Zhang, Eric Horvitz, Yiling Chen, and David C Parkes. Task routing for prediction tasks. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 2*, pages 889–896. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- Haiyi Zhu, Robert Kraut, and Aniket Kittur. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 935–944. ACM, 2012.
- M. Zion and I. Sadeh. Curiosity and open inquiry learning. *Journal of Biological Education*, 41:162–8, 2007.
- Kathryn Zyskowski and Kristy Milland. Crowdworking visibility: An ethnography of a discussion board for digital labor. Unpublished Manuscript, 2015.