Payment Schemes in Online Marketplaces: How Do Freelancers Respond to Monetary Incentives?

March 2016

Justine Moore

Department of Economics
Stanford University
Stanford, CA, 94305
jmoore94@stanford.edu

Under the direction of
Professor John Shoven and Orie Shelef

ABSTRACT

The rise of online labor marketplaces has allowed employers to hire freelancers around the world to perform various tasks. Though effective incentive schemes are important in any labor market, they are particularly valuable in these marketplaces, as employers are unable to directly monitor their freelancers' behavior to ensure that the freelancers are spending their time productively. This presents a challenge for all employers, but especially those who pay by the hour, as they do not want to pay freelancers for time spent working on other tasks or not working at all. I conduct an experiment on Upwork, the world's largest online freelancing platform, to determine how freelancers respond to a simple monetary incentive that gives them the opportunity to earn a bonus for outperforming a benchmark. I examine how this incentive affects the quality and quantity of work, as well as how responses to the incentive differ based on a freelancer's measurable characteristics. I find that for the group of freelancers as a whole, incentives do not have a significant effect on performance. However, it seems that this is due to the fact that different sub-groups of freelancers have opposite reactions to the incentive. Freelancers with low reputation scores respond by increasing the quantity of output and improving the accuracy of their work, while freelancers with high reputation scores respond by decreasing the quantity of output, with accuracy that does not significantly increase.

Keywords: Upwork, Online Freelancing, Incentives for Employees, Hourly Wages, Performance-Based Bonuses, Reputation Scores, Job Success Scores

Acknowledgements: I would like to thank Professor Shoven, who is both my thesis advisor and my major advisor, for all of his support and advice during my time at Stanford. I thank Orie Shelef for taking the time to guide me through this process, from initially selecting my topic to writing this paper. I thank Marcelo Clerici-Arias for all of his work running the honors program, and for providing feedback on my paper.

Contents

I. Introduction

Several secular trends have contributed to an exponential growth in the online freelancing marketplace in the past several years. As Internet security allows for safer online transactions and new software creates an opportunity for employers to effectively monitor remote employees, many companies have become more comfortable with conducting business online. In addition, the proliferation of Internet users in emerging economies has allowed employees and employers in these countries to join online labor marketplaces that they previously could not access. The entrance of these employees has been crucial in encouraging employers in developed countries to participate in online marketplaces, as outsourcing tasks to developing nations allows employers to reduce their costs by hiring freelancers willing to accept low wages. There are also significant benefits for the freelancers, as working online provides flexibility that traditional jobs typically don't allow—online freelancers can work anywhere with an Internet connection and can often be employed in multiple jobs at a time. Freelancers in developing nations may be able to earn higher wages online than in the traditional labor market in their area, and freelancers in developed nations can supplement their income or even live off of their freelancing if they have specialized and highly desired skills.

Although online freelancing is becoming increasingly popular for the reasons mentioned above, employers struggle with the inability to directly monitor their freelancers. Freelancers may be working halfway around the world while their employer is sleeping, which makes it difficult for employers to track their freelancers' productivity. In this type of marketplace where close monitoring is impossible, incentives are crucial in aligning the freelancers' interests with those of the employer by motivating freelancers to

spend their time productively. Few studies have examined how to best incentivize freelancers in online marketplaces, and to the best of my knowledge, none have explored how the optimal incentive scheme may differ between freelancers with different measurable characteristics such as level of work quality, as reflected in their job success scores. Therefore, since many employers are somewhat arbitrarily selecting payment schemes for their employees without consideration of the potential impact on work quantity and quality, the online freelancing market is likely operating inefficiently. With more effective and individualized incentive schemes, employers could extract greater productivity from their freelancers without necessarily having to pay them more.

I conduct an experiment on Upwork, the world's largest online freelancing marketplace, to examine how freelancers respond to a monetary incentive. I hire freelancers to work on a simple data extraction task under one of two incentive schemes: a basic hourly wage, and an hourly wage with a piece rate bonus based on performance. The piece rate bonus was designed with the intent that the average freelancer in this incentive group would receive the same hourly wage as the freelancers in the basic hourly wage group. The freelancers are randomly assigned to one of the two incentive schemes. Both groups undergo a 30-minute training session to learn how to successfully complete the task, and then spend an hour and a half working, with their submissions recorded in individual Google forms.

I then analyze the freelancers' performance to determine whether one incentive scheme is superior in motivating freelancers to increase the quantity and quality of their work. I find that the incentive does not have a significant effect on any of our measures of performance for the group of freelancers as a whole. However, when I analyze the

interaction between the incentive schemes and a freelancer's measurable characteristics, I find that the incentive does influence freelancer behavior in specific sub-groups of freelancers. When freelancers with high job success scores are offered an incentive, their output declines, and their accuracy improves only slightly. The opposite occurs when freelancers with low job success scores are offered an incentive—both their output and the quality of their work improve significantly.

II. Literature Review

My paper builds upon prior research on both the role of reputation in online marketplaces and how various wage systems and other incentives affect work product in traditional and online marketplaces. This literature review examines both types of research and discusses how findings from prior studies are relevant to my paper.

Reputation in Online Marketplaces

Most studies on the role of reputation in online marketplaces use data from eBay, as it was one of the first e-commerce sites and it stores data from millions of transactions. In an early study using eBay data, Standifird (2001) examines the impact of a seller's positive and negative reputational ratings on the final bid price for an item, and finds that a strong positive reputation can drive up prices after the seller exceeds a certain threshold of positive comments. However, he also finds that negative ratings are significantly more influential than positive ones in determining the final price of the item. Several other studies (e.g. Houser, 2006; Wolf, 2005; Melnik, 2002) confirm the finding that sellers with good reputations receive price premiums in eBay auctions, though Jin (2006) finds that reputation is positively correlated with an increased number of bids but not with a

higher price. Livingston (2002) also concludes that sellers with positive reputations are rewarded with higher prices, though he finds that the marginal return to positive feedback is decreasing.

Cabral (2010) focuses exclusively on the effect of negative ratings, and finds that a seller's weekly sales growth rate drops from positive to negative after he receives his first negative feedback. Subsequent negative feedback arrives much more quickly, but doesn't have nearly the same impact on the sales growth rate. Cabral suggests that sellers change their behavior after receiving their first negative review, as they may be discouraged and put less effort into providing superior service for their customers, which increases the probability of receiving subsequent negative feedback. Whether or not this also occurs with online freelancers when they receive poor feedback from an employer has not been explored.

In some instances, the authors participate in the online marketplace they are studying to collect their own data on transactions, which is similar to what I do on Upwork. Resnick et al. (2006) sold matching items using both the account of an experienced eBay seller and new accounts (with no reputation scores or feedback), and find that buyers paid more to purchase the postcards from the established seller. The authors then left negative feedback on the new seller accounts before selling more items. Surprisingly, this feedback had no effect on the price the buyers were willing to pay for the postcards, even though the negative reviews constituted a significant amount of the feedback that the sellers had received. The authors therefore hypothesize that buyers may treat all new sellers as untrustworthy, regardless of whether or not they have negative feedback.

Jin (2006) also extends the typical data collection study by purchasing baseball cards from eBay sellers and hiring professionals to determine their true value. He finds that sellers who claim they have a high quality card receive a price premium, but when these "high quality" cards are evaluated by professionals, they are no better than other cards without this label. Sellers with good reputations are less likely to claim that their cards are high quality, and are also less likely to send counterfeit cards or simply default on the sale and send no card. However, they are not any more likely to send high quality cards. This suggests that a seller's reputation may be more indicative of how likely they are to "cheat" by sending a fake item or no item at all than of the quality of their items.

Most directly relevant to my experiment is a study by Pallais (2014), who hired workers to complete a task on Upwork. When the task was complete, Pallais posted either an "uninformative" comment or a detailed comment with objective information about the worker's performance on each worker's profile. Pallais then tracked the subsequent employment outcomes of workers in all three groups—the two treatment groups and the control group of workers she did not hire. She finds that workers in the uninformative comment group were significantly more likely than the workers in the control group to find subsequent work, and also requested higher wages and had higher earnings from future employers. Workers in the detailed feedback group were even more likely to be employed, requested even higher wages, and had even higher earnings. This suggests that employers believe that a job success score is a legitimate signal of a worker's quality—they are more willing to hire freelancers with more positive feedback, and are also more willing to pay them higher wages.

While these studies come to different conclusions about the relative value of various types of feedback, they all conclude that counterparties in online transactions use reputation systems to make decisions about what to purchase (or who to hire), and how much they are willing to pay. This indicates that these reputation systems provide some value in revealing information about certain qualities of a user, though there is debate about exactly what these qualities are. Therefore, we can expect that a freelancer's job success score on Upwork will be useful in revealing something about the freelancer, whether it is the quality of their work, the amount of work they are able to accomplish in a specific amount of time, or another characteristic that employers value. As a result, we might expect to see that a freelancer's job success score does provide valuable information about how a freelancer will respond to incentives.

Wage Systems and Incentives

Theoretical Work and Models

Much of the literature regarding the effects of financial incentives on employee performance builds upon the seminal framework developed by Holmstrom and Milgrom (1991). Their model assumes that employees have either more than one task to complete or that there are multiple elements to one task, and therefore incentives influence not only how much effort employees exert but also how they allocate time among various responsibilities. According to this model, performance-based incentives may not always be effective, particularly when performance is easily measured for one task but not for another. The authors give an example of workers producing machines. Since quantity is more easily measured than the quality of output for this task, a piece rate bonus based on

output may encourage workers to produce more at the expense of quality. The model suggests that when it becomes more difficult to measure performance in competing activities, it is less desirable to provide incentives for the activity where performance is easily measured, as workers will neglect the other activity. Holmstrom and Milgrom conclude that for some tasks, a fixed wage independent of performance is the optimal incentive scheme.

Dana (1993) explores the question of whether an hourly wage, fixed fee, or contingent fee is the optimal compensation scheme for attorneys, and concludes that the contingent fee system (in which the attorney receives a percentage of the money awarded to the client) is optimal. Contingent fees serve as a performance-based incentive, because the attorney receives nothing if the case is lost and a fixed percentage of the award if the case is won. Therefore, this system aligns the attorney's incentives with the incentives of his or her client—the attorney is motivated to win the case, not to charge as many hours as they can (which an hourly wage incentivizes them to do) or to spend as little time on the case as they can get away with (which a fixed fee incentivizes them to do).

Ritter and Taylor (1999) have a more pessimistic view of performance-based incentives. Their model suggests that low productivity workers become discouraged (and therefore perform worse) when they see high productivity workers, who typically respond dramatically to these incentives, doing significantly better. The authors use this model to predict that while a firm can increase profits by implementing an incentive scheme by which workers compete for performance-based incentives, it is de-motivating to low productivity workers. This model is particularly interesting because it suggests that there

may be differing responses to incentives between different types of workers, which I will be examining in my experiment.

Data Analysis Studies

Shearer (2004) and Lazear (1996) both analyze data from companies that switched from a fixed wage compensation system (where workers were paid a set amount to complete a job) to a piece rate compensation system (where workers were paid per unit of output). They both find that worker productivity increased by 20-30% after the switch, despite the fact that the workers were paid less per unit of output under the piece rate system. Lazear notes that some of this increase in output is due to a change in the composition of the workforce, as the piece rate wage system attracts higher quality workers and decreases turnover amongst the high quality workers who are already employed at the company. Lazear also finds that there is more variance in the productivity of workers under the piece rate system, as high ability workers are incentivized to produce more and have the capacity to do so. This suggests that in my experiment, if we assume that high ability freelancers have high job success scores, these freelancers may react more strongly to the incentive than those with lower reputation scores.

Sauermann (2014) examines the impact of performance-based incentives using data from a company that switched from an hourly wage scheme to compensation based on quality of work. He finds that after the switch, the average worker's performance improved, but the improvement was three times larger for low ability workers than for average ability workers. For high ability workers, performance-based pay had a negative effect—these workers had no reason to respond to the incentive because their

performance under the hourly system would have already qualified them for the highest level of performance pay. Interestingly, Sauermann finds that quantity of work did not drop after the performance-based pay was implemented, which seems to contradict Holmstrom and Milgrom's model. Regardless, this result suggests that workers with low reputation scores may have the most significant response to a financial incentive.

Experimental Studies

Shi (2010) and Heywood et. al (2013) conducted experiments on how workers react to a switch from an hourly wage to a piece rate wage that compensates workers per unit of output. In both experiments, the quantity of output increased significantly after the implementation of the piece rate system, and survey results indicate that high quality workers significantly preferred (and benefitted the most from) the switch. Shi finds that the quality of the work doesn't change under the piece rate system, while Heywood et. al find that quality improves for workers who are closely monitored and declines for workers who are not closely monitored. Therefore, we may expect to see that workers with high job success scores have the most significant response to the performance-based incentive. Given that Upwork freelancers are not particularly closely monitored, we may also expect to see an increase in quantity of output but a decrease in quality.

Gneezy and Rustchini (2000) conducted an experiment to determine the effect of incentives on the results of an IQ test. All of the participants were randomly sorted into one of four incentive groups, each of which was awarded a different piece rate bonus for each question answered correctly. The authors find that participants in the two highest piece rate groups scored significantly better on the IQ test than participants in the lowest piece rate group and the group that received no incentive. When they sort participants by

ability, they find that this holds true for all subgroups of participants, and that the change in performance motivated by the incentives is similar for each subgroup. If this finding holds true for my experiment, I would expect to see that a worker's job success score does not influence his or her response to incentives.

My paper also builds off findings from several studies focusing exclusively on incentives in Mechanical Turk, an online freelancing marketplace run by Amazon that typically hosts shorter-term jobs than those on Upwork. On Mechanical Turk, workers are often hired on a piece rate basis to complete small portions of a larger task. Early papers focusing on the effect of incentives on the performance of Mechanical Turk workers by Mason and Watts (2009) and Horton and Chilton (2010) suggest that the magnitude of a financial incentive has a significant effect on the quantity of output, though not necessarily on the quality. Both papers use data from experiments in which Mechanical Turk workers were randomly assigned to one of two piece rates per unit of output, and find that workers receiving a higher piece rate payment are willing to continue the task for significantly longer.

Rogstadius et. al (2011), Yin et. al (2013), and Harris (2011) all expand upon this work by adding to the complexity of the basic experiment. Rogstadius et. al (2011) explore the effects of both intrinsic and extrinsic incentives, and confirm the finding that higher pay increases the quantity of work produced but does not affect the quality. However, they find that quality of work improves when workers believe that their work product is benefiting a charity. Yin et. al (2013) find that performance-contingent financial rewards had no effect on either the quantity or quality of work, contrary to the results of previous studies, but that changing the amount of the reward did have a

significant effect on work quality and quantity—if a worker's bonus increased, he or she completed more tasks with improved accuracy. Therefore, the authors conclude that there is a powerful anchoring effect, as workers use their first payment to form their perception of a fair wage for the job, and respond accordingly when their wage increases or decreases.

Harris (2011) explored the effect of both positive and negative incentives on the performance of Mechanical Turk workers. He randomly sorted freelancers into one of four incentive schemes: baseline (no performance-based incentive), positive (base rate plus bonuses for accuracy), negative (base rate minus penalties for inaccuracy), and combined (base rate plus bonuses for accuracy and minus penalties for inaccuracies). He found that the freelancers working under the positive, negative, and combined incentive schemes produced significantly more accurate work than those working under the baseline incentive scheme. The positive incentive group had the best performance, followed closely by the combined incentive group and then the negative incentive group.

In my experiment, I attempt to determine how conventional wisdom on the power of incentives applies to online freelancers with different characteristics, particularly job success scores. Some studies suggest that a freelancer's job success score will influence his or her response to incentives, while others suggest that job success score and incentives should not interact. Intuitively, it makes sense that if a high ability worker has to exert a minimum amount of effort to keep his or her job in an hourly wage system, he or she will have greater capacity for improvement than a low ability worker when an incentive is introduced. However, this does not always hold true in experiments, which suggests that there is more work to be done in this field. In addition, my paper is one of

the first to use data from a new online marketplace—Upwork—which typically hosts longer-term jobs than Mechanical Turk. Upwork's job postings therefore more closely resemble jobs in the traditional labor market.

III. Study Design

I conducted an experiment on Upwork to explore the questions of how a financial incentive influences a freelancer's performance and how the effect of this incentive differs between freelancers. The experiment was designed so that all freelancers had exactly the same experience with the task (with the exception of the financial incentive) to ensure that differences in performance could be attributed to the incentive. The freelancers were hired on a first come, first serve basis (with the exception of candidates without a job success score, who were automatically disqualified) until I exhausted my budget, and the freelancers were then randomly sorted into one of two incentive groups. They also participated in a training session before I started tracking performance to ensure that they were starting on a relatively level playing field with regard to their understanding of how to complete the task. However, the task was specifically designed to not require any advanced skills—freelancers were only asked to be able to read and understand English, and extract data from a file.

To maintain the integrity of the experiment, it was important that the freelancers believed that they were working on a typical Upwork job, not participating in an experiment. As I've been hiring freelancers on Upwork for the past two years, mostly for real tasks related to data extraction and processing, my account looked legitimate—I had posted eight jobs and had 32 reviews from freelancers that I had previously hired. I made sure that my job posting looked similar to those for regular data extraction tasks, and that

I was not doing anything out of the ordinary that might make freelancers suspicious (e.g. paying a rate that was significantly lower or higher than the average, or making freelancers sign a release form to participate). As far as I know, none of my freelancers suspected that they were participating in an experiment.

Introduction to Upwork

Upwork (formerly known as oDesk), had nine million registered freelancers as of May 2015, when the company last reported growth metrics. At this time, four million clients (employers) were registered on the site, three million jobs were posted annually, and over $1 billion in transactions were taking place on Upwork every year. Upwork connects clients with freelancers who work in a wide variety of fields, from graphic design to electrical engineering. The most popular category is administrative support, which had more than 625,000 freelancers as of February 2015.

Freelancers apply for jobs by submitting their Upwork profile, a bid for the job, and any other application materials requested by the client (typically a short statement about why they are interested in the job). After completing the job, which can vary in length from a few hours to over a year, the freelancer receives a reputation score from the client that reflects his or her performance in a variety of fields, including work quality, communication, and timeliness. When a freelancer has been a member of Upwork for at least three months and completed at least four jobs for three unique clients, these individual reputation scores begin to count towards an aggregate job success score, which reflects the percentage of a freelancer's jobs that resulted in a "great client experience."

As Upwork's job success score formula is proprietary, the company does not reveal exactly how it is calculated to either freelancers or clients. However, Upwork's public guide on job success scores suggests that at a high level, freelancers and clients should view the score in the following way: (successful contract outcomes – negative contract outcomes) / total outcomes. The actual calculation is significantly more complex than this, and takes into account other factors such as the length of each relationship, the reputation of the client, and the number of relationships that end without any activity or payment.

Data Collection

I collected my data by creating an Upwork client account and hiring freelancers to work on a data extraction task with a wage of $3/hour. Each freelancer was randomly assigned to work under one of two incentive schemes. The first, Incentive Scheme A, was a simple hourly wage ($3/hour) with a bonus unrelated to performance ($1/hour), for a total hourly wage of $4/hour. The second, Incentive Scheme B, was a base wage of $3/hour with an opportunity to earn a $0.15/bio bonus for each bio that the freelancer processed accurately beyond the base rate of ten accurate bios per hour. For example, if a freelancer working under Incentive Scheme B processed 15 bios accurately in an hour, they would receive $3.75. I intended for the average hourly wage earned to be approximately $4/hour in both groups to prevent any differences in performance that could result from the freelancers being paid different amounts. The pilot experiment and my past experience with hiring on Upwork suggested that a $0.15/bio bonus with a ten bio per hour base rate should have resulted in the average freelancer working under

Incentive Scheme B receiving approximately $4/hour. However, I found that the freelancers in this experiment underperformed expectations and were paid an average bonus of $0.27/hour, making their average total wage $3.27/hour.

I hired a total of 62 freelancers (31 in each incentive group), and 59 completed the task. The three freelancers who did not complete the task were working under Incentive Scheme B, so the final results are from 31 freelancers working under Incentive Scheme A and 28 freelancers working under Incentive Scheme B. Before beginning work, all of the freelancers spent half an hour in a training session to learn how to complete the task, and were paid a base rate of $3/hr during this time. After completing training, each freelancer worked for an hour and a half under their incentive scheme. The assigned task was pulling educational information (e.g. college degree, college name, year degree was received) out of biographies of hedge fund managers, and every freelancer received the same set of biographies to ensure that differences in performance could not be attributed to differences in the task. This task required very few skills other than basic reading comprehension and the ability to submit a Google form. Freelancers were not asked to do any further research on the hedge fund managers or consult any additional resources other than the biographies that were provided to them. All of the freelancers also received a detailed set of instructions and a list of common mistakes to avoid, which I compiled from errors that other freelancers made during the pilot experiment.

I tracked freelancer performance by having them each use their own Google form to submit the biographical information. This form, along with Upwork's time tracker, allowed me to see how long it took each freelancer to process each set of biographies. The tracker told me how much time the freelancer logged to process a given set of files,

and provided screenshots of the freelancer's screen every ten minutes. The Google form allowed me to see when the freelancer submitted each entry, down to the exact second of submission. I also automated the process of checking each freelancer's submissions for accuracy by using an Excel formula to compare their submissions with an answer key.

Accuracy was measured through two metrics: the number of bios processed accurately per hour, and the number of accuracy points a freelancer received out of the total. For a bio to count as "accurate," a freelancer must have filled in every field of the Google form correctly. I was concerned that this measure might yield extremely poor accuracy scores for freelancers who often made small mistakes on one field of the form (such as spelling the university name incorrectly) but had otherwise perfect submissions. Therefore, I calculated a second measure of a freelancer's accuracy score: the number of fields that he or she filled in correctly out of the total number of fields. For example, if a freelancer correctly identified the manager's university name and degree type but made a typo in the "year received" field, he or she still received two out of three accuracy points for that biography.

Empirical Strategy

A. Do incentives matter?

After confirming that the two groups of freelancers are sufficiently similar in all measurable characteristics, I run several regressions to answer my first question: How do incentives affect performance in terms of both quantity and quality of work? I regress my outcome variables against the treatment dummy and my controls:

$$Bios = \beta_0 + \beta_1 I^{treatment} + \beta_2 JobsCompleted + \beta_3 PreferredHourlyWage + \beta_4 JobSuccessScore$$

$$AccurateBios = \alpha_0 + \alpha_1 I^{treatment} + \alpha_2 JobsCompleted + \alpha_3 PreferredHourlyWage + \alpha_4 JobSuccessScore$$

$$PercentCorrect = \gamma_0 + \gamma_1 I^{treatment} + \gamma_2 JobsCompleted + \gamma_3 PreferredHourlyWage + \gamma_4 JobSuccessScore$$

$I^{treatment}$ is a dummy variable that equals 0 if the freelancer is working under Incentive Scheme A (basic hourly wage) and equals 1 if the freelancer is working under Incentive Scheme B (hourly wage with the potential to earn a bonus). $Bios$ represents the number of bios that the freelancer processed during the experiment (excluding bios processed during training), so β_1 is indicative of the effect of the incentive on quantity of output. $AccurateBios$ represents the number of bios processed with 100% accuracy, so α_1 is a measure of the effect of the incentive on both quantity and quality of output. $PercentCorrect$ is another measure of accuracy—it is calculated by taking the points that the freelancer received divided by the total number of points possible for the bios that the freelancer processed. Therefore, γ_1 is a measure of the effect of the incentive on quality of output.

B. Do different people respond to incentives differently?

I then try to answer my next question—do incentives affect different freelancers in different ways? Here, I can quantify differences in freelancers as the differences in

observable characteristics pulled from each freelancer's profile (job success score, jobs worked, and preferred hourly wage).

I add interaction variables between incentive and the various observable characteristics to my regressions and again regress these variables against my three outcome variables (bios, correct bios, and percent correct). I will use "OutcomeVariable" to represent these three outcome variables from now on:

$$OutcomeVariable = \beta_0 + \beta_1 JobSuccessScore + \beta_2 JobsCompleted + \beta_3 PreferredHourlyWage + \beta_4 I^{treatment} + \beta_5 I^{treatment} * JobSuccessScore + \beta_6 I^{treatment} * JobsCompleted + \beta_7 I^{treatment} * PreferredHourlyWage$$

I run an additional set of regressions for job success score, as Upwork provides guidelines regarding different categories of job success scores. According to Upwork's website, any score at or above 90% is "excellent," while a score at or below 75% could result in the freelancer struggling to win new clients. Therefore, I've bucketed the freelancers into three different bins by job success score: high (90-100%), medium (76-89%), and low (60-75%). None of the freelancers I hired had a job success score below 60%.

Table 4.1: Job Success Score Bins Using Upwork Categories

Bin	Mean Job Success Score	Freelancers in Incentive Scheme A	Freelancers in Incentive Scheme B	Total Freelancers
High	0.964	17	14	31
Medium	0.855	7	8	15
Low	0.689	7	6	13

I then regress the outcome variables on interactions between these bins and the incentive and my control variables to determine if the effect of the incentive differs by bin. Here, I leave out a dummy for medium job success score (JSS) under Incentive Scheme A so that it serves as the baseline.

$$OutcomeVariable = \beta_0 + \beta_1 JobsCompleted + \beta_2 PreferredHourlyWage + \beta_3 I^{highJSS} * I^{incentiveA} + \beta_4 I^{highJSS} * I^{incentiveB} + \beta_5 I^{mediumJSS} * I^{incentiveB} + \beta_6 I^{lowJSS} * I^{incentiveA} + \beta_7 I^{lowJSS} * I^{incentiveB}$$

After running the regressions using the bins provided by Upwork, I reclassify the bins in an effort to have approximately equal numbers of freelancers in each bin. Using Upwork's classification of a high job success score, more than half of freelancers fall in the high job success score category. Therefore, I suspect that the medium job success score group may not be an accurate baseline, as all of the freelancers in that group have job success scores below the median. The new bins are the following: high (95-100%), medium (87-94%), and low (60-86%).

Table 4.2: Job Success Score Bins Using Revised Categories

Bin	Mean Job Success Score	Freelancers in Incentive Scheme A	Freelancers in Incentive Scheme B	Total Freelancers
High	0.984	9	11	20
Medium	0.908	11	8	19
Low	0.736	11	9	20

I then run the same regressions as above using the new bins in an attempt to compare the interactions between job success categories and the incentive using a more accurate baseline.

C. How do incentives influence payment?

Finally, I attempt to determine how the magnitude of payment differs between freelancers in both incentive groups to examine how the incentive influences payment. Table 4.3 below contains information regarding the total amount paid to freelancers in each incentive group, the cost per freelancer, and productivity statistics involving payment (bios processed per dollar and correct bios processed per dollar). As the table illustrates, the average freelancer working under Incentive Scheme A received a total of $1.14 more than the average freelancer working under Incentive Scheme B. As a result, the average freelancer working under Incentive Scheme B processed more bios (and more correct bios) per dollar spent.

Table 4.3: Productivity by Incentive Group

Incentive	Total Cost	Cost / Freelancer	Total Bios Processed	Total Correct Bios	Bios Processed/$	Correct Bios Processed/$
Incentive Scheme A	$186.00	$6.00	1169	249	6.28	1.34
Incentive Scheme B	$136.05	$4.86	1092	245	8.03	1.80

I regress the treatment dummy against three variables relating to payment: hourly wage (the amount that the freelancer was actually paid, not their preferred hourly wage), bios processed per dollar spent, and correct bios per dollar spent to determine if these variables differ significantly between the incentive groups.

$$Wage = \beta_0 + \beta_1 I^{treatment}$$

$$BiosPerDollar = \alpha_0 + \alpha_1 I^{treatment}$$

$$CorrectBiosPerDollar = \gamma_0 + \gamma_1 I^{treatment}$$

IV. Data

As I was able to conduct my own experiment, the data I have collected is relatively close to what I would consider the ideal dataset to study the question of how incentives and job success score interact to influence freelancer performance. In an ideal world, I could have controlled the outside factors in each freelancer's life, such as how many other jobs they held while completing my task, to ensure that any differences in performance could not be attributed to those factors. This became particularly relevant when flooding in Chennai, India in December 2015 caused several freelancers to lose Internet connection for over a week, which delayed their completion of the task. In addition, my sample size is relatively small (59 freelancers completed the experiment) due to budgetary limits. With a larger budget, I would have been able to hire significantly more freelancers, and an increased sample size would have likely allowed me to draw more definitive conclusions from my results. Please see Appendix 3 for further ideas on how to improve subsequent studies.

As previously mentioned, I was careful to ensure that all freelancers had the same experience during the hiring and training process to prevent any differences in performance that could be attributed to differences in onboarding. All freelancers saw the same job posting, which advertised a data extraction task that paid $3.00/hour. Freelancers were hired on a first-come, first-serve basis between November 18, 2015 and December 1, 2015, with the exception of freelancers with no job success score, who were not hired. All freelancers received the same scripted message with instructions for the training period, as well as a link to the personal Google form they used to submit their work.

After the 30-minute training period, I gave each freelancer feedback on their work. All of the freelancers received between two and four (depending on the number of errors in the training submissions) pieces of feedback taken from a repository of feedback generated based on common mistakes in the pilot version of the experiment. For example, a common piece of feedback was "Make sure not to leave any blank spaces before or after anything you type into a field—the software that checks your responses will mark that as incorrect." The end of each message varied by incentive group. Freelancers working under Incentive Scheme A were informed that their wage was being increased to $4/hour, and freelancers working under Incentive Scheme B were informed that they had the chance to earn a bonus and were given an example of how the bonus system worked.

A total of 59 freelancers (out of the 62 we originally hired) completed the experiment. All three of the freelancers who quit were working under Incentive Scheme B—one quit before training began with no explanation, and two quit after completing training (one stopped responding to emails, and the other quit because he wanted a job where he could more quickly log a significant number of hours). Most freelancers completed both rounds of the experiment within a week, but a few freelancers were delayed for various reasons, including the flooding in India. The first freelancer completed the experiment on November 20, 2015, and the last freelancer completed the experiment on December 21, 2015.

Full demographic information for the freelancers is provided in Appendix 1, but it is important to note that the plurality of freelancers live in India (27.59%), followed by Bangladesh (22.41%), and Philippines (20.69%). The average number of hours worked

by each freelancer prior to being hired for my task was 1976.33. Approximately 37% of the freelancers had worked more than 1000 hours on other Upwork jobs, and 56% had worked more than 500 hours on other Upwork jobs. Only 29% of the freelancers had worked fewer than 100 hours on Upwork, and nearly a fourth of these freelancers had only worked fixed wage jobs before (and therefore had no record of hours worked). In addition, 73% of freelancers had previously worked at least 10 jobs on Upwork, and the average number of jobs worked before being hired for my task was 56.20. Therefore, the majority of the freelancers were relatively experienced, having worked hundreds (or even thousands) of hours on tens (or even hundreds) of jobs.

V. Results and Discussion

A. Do incentives matter?

First, I confirm that the randomization was successful by testing the difference in means between the groups for all measurable characteristics using standard t-tests. I find that for all three of the measurable characteristics pulled from each freelancer's Upwork profile (job success score, jobs worked, and preferred hourly wage), the freelancers in Incentive Group A and Incentive Group B were not significantly different.

Table 5.1: Test of Randomization

	Incentive Group A			**Incentive Group B**			**Difference**	
	Mean	SD	N	Mean	SD	N	Mean	P-Value
Job Success Score	0.87	0.12	31	0.88	0.11	28	0.01	0.74
Jobs Worked	42	54	31	72	132	28	29	0.28
Preferred Hourly Wage	5	1.8	31	6.2	4.1	28	1.2	0.15

Now that I have confirmed that the controls do not differ significantly between the incentive groups, I regress the outcome variables against the incentive dummy alone. I find that the incentive has a positive effect on all of the outcome variables—the number of bios completed increases by 1.3, the number of correct bios increases by 0.7, and percent correct increases by 4.7 percentage points—but none of these effects are statistically significant.

Table 5.2: Do Incentives Matter?

	Bios Completed	**Correct Bios**	**Percent Correct**
Incentive	1.290 (4.393)	0.717 (2.551)	0.047 (0.068)
Constant	37.710 (3.026)	8.032 (1.757)	0.550 (0.047)
Observations	59	59	59
R-squared	0.0015	0.0014	0.0084

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

This suggests that overall, the incentive that we provided did not motivate the freelancers to significantly improve their performance. Though this result may seem somewhat surprising, as previous studies suggest that monetary incentives are usually successful in motivating workers to increase the quantity (if not also the quality) of tasks completed, there are several reasons that could explain why our experiment yielded different results.

The first explanation is that the monetary value of the incentive that we provided was simply not significant enough to motivate a notable increase in performance. Receiving $0.15 per additional bio processed beyond the threshold represented a wage increase of 5% per bio over the initial hourly wage. Though this may have motivated some freelancers, it's possible that the incentive was not large enough to motivate the freelancers who would have to exert a significant amount of effort to cross the threshold of accurate bios in order to qualify for a bonus. Therefore, the incentive may have been rendered ineffective because it was simply not large enough to motivate the majority of freelancers to change their behavior.

Another possible explanation is that the incentive had opposite effects on sub-groups of freelancers within the incentive group, and that these effects canceled each other out. If this occurred, the overall effect might be zero even if certain groups of freelancers (e.g. those with particularly high or low reputation scores) had a significant response to the incentive. We will further examine this explanation when attempting to answer the question of whether or not the incentive affected different groups of freelancers differently.

Finally, it's possible that another incentive was at work—the desire to have a high job success score. The job success score system is new to Upwork, and it is evident from Upwork's freelancer support forums that many freelancers are confused about exactly how it is calculated. Therefore, our experiment may have been picking up on the fact that all of the freelancers, even those not offered a monetary incentive, were exerting their maximum amount of effort for my job in hopes of boosting their job success score. If that were the case, we would expect to see no improvement in performance when freelancers

were offered an incentive, as the freelancers would already be operating at maximum effort and would have no capacity to improve.

B. How do incentives affect different types of freelancers differently?

Next, I use the observable characteristics gleaned from each freelancer's Upwork profile to determine whether the incentive affected different freelancers differently. I regress each of the outcome variables against the treatment dummy, the controls (job success score, jobs worked, and preferred hourly wage), and interactions between each of the treatment dummy and each of the controls.

I find that in this regression, three variables have a statistically significant effect on the number of bios completed—Job Success Score (p-value of 0.022), Incentive (p-value of 0.058), and the interaction between Job Success Score and Incentive (p-value of 0.015). Both Job Success Score and Incentive have a positive effect, and the interaction has a negative effect. None of the controls or interactions had a significant effect on any of the other outcome variables, though the interaction between incentive and jobs completed has a relatively strong negative effect on the number of correct bios (decrease of 35.51). Therefore, we can conclude that out of all of the quantifiable characteristics we tested, only a freelancer's job success score has a significant effect on his or her response to an incentive.

Table 5.3: Regressions with Interactions between Incentive and Freelancer Characteristics

	Bios Completed	**Correct Bios**	**Percent Correct**
Job Success Score	59.674** (25.246)	4.257 (15.486)	-0.271 (0.409)

Jobs Completed	-0.049 (0.057)	0.021 (0.035)	-0.001 (0.001)
Hourly Wage	-1.632 (1.655)	-0.674 (1.015)	0.000 (0.027)
Incentive	65.055* (33.599)	30.575 (20.609)	0.113 (0.544)
Incentive * Job Success Score	-93.524** (37.207)	0.021 (0.035)	0.034 (0.602)
Incentive * Jobs Completed	0.075 (0.062)	-35.510 (22.823)	0.000 (0.001)
Incentive * Hourly Wage	2.597 (1.833)	(-0.030) (0.038)	-0.014 (0.030)
Constant	-4.057 (23.333)	6.793 (14.312)	0.814 (0.378)
Observations	59	59	59
R-squared	0.1876	0.0935	0.1090

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

While Job Success Score and Incentive both have a positive effect on the number of bios completed, the interaction between the two has a significantly negative effect, which is somewhat puzzling. This suggests that offering freelancers an incentive has a large positive effect on freelancers with low job success scores, and an even larger negative effect on freelancers with high job success scores. This lends credence to my theory that the overall effect of the incentive is negligible because opposite effects on different sub-groups of freelancers cancel each other out.

However, it is still somewhat surprising that the incentive has a significant negative effect on bios completed for freelancers with high success scores. This could be because freelancers with high job success scores focus on the fact that they need to process the bios accurately to receive a bonus (and therefore their speed decreases significantly). It's also possible that the bonus is less valuable to freelancers with high job success scores. Because these freelancers likely have more opportunities to earn higher salaries in other jobs, they might have preferred to multitask and do work for other jobs while billing me instead of work harder on my task to earn the bonus.

Regressions with Upwork Bins

I then run regressions using the dummy variables for the bins of freelancers generated from Upwork's definitions of high, medium, and low job success scores. I regress my outcome variables against my controls and an interaction variable between each bin and each incentive scheme (with the exception of Medium Job Success Score/Incentive Scheme A, which serves as my baseline).

I find that compared to this baseline, the High Job Success Score/Incentive Scheme A group processed an average of 10 more bios, and the Medium Job Success Score/Incentive Scheme B and Low Job Success Score/Incentive Scheme B groups each processed an average of 11 more bios. The High Job Success Score/Incentive Scheme B and Low Job Success Score/Incentive Scheme A groups processed slightly fewer bios than the baseline (0.5 and 2 fewer, respectively). However, none of the bin/incentive combinations are statistically significantly different from the baseline in terms of number of bios completed.

Compared to the baseline, all of the groups outperformed in terms of number of correct bios processed. The High Job Success Score/Incentive Scheme A group and the Low Job Success Score/Incentive Scheme A group processed an average of 6.8 and 8 more bios, respectively, though these differences were not statistically significant. The Low Job Success Score/Incentive Scheme B group processed significantly more correct bios (average of 16.23, with a p-value of 0.003) than the baseline.

All of the groups also outperformed the baseline in terms of percent correct, and all of the differences (with the exception of Medium Job Success Score/Incentive Scheme B) were statistically significant. In descending order, the average percent correct was higher than the baseline for Low Job Success Score/Incentive Scheme B (average of 34.8 percentage points higher, p-value of 0.016), Low Job Success Score/Incentive Scheme A (average of 26.8 percentage points higher, p-value of 0.053), High Job Success Score/Incentive Scheme B (average of 22.5 percentage points higher, p-value of 0.059), and High Job Success Score/Incentive Scheme A (average of 19.1 percentage points higher, p-value of 0.095).

Table 5.4: Regressions using Upwork Bins

	Bios Completed	**Correct Bios**	**Percent Correct**
Jobs Completed	0.014 (0.024)	0.003 (0.014)	0.000 0.000
Hourly Wage	0.576 (0.740)	-0.085 (0.415)	-0.01 (0.011)
High JSS * Incentive Scheme A	9.818 (7.495)	6.788 (4.204)	0.191** (0.112)
High JSS * Incentive Scheme B	-0.549	3.289	0.225**

	(7.751)	(4.348)	(0.116)
Medium JSS * Incentive Scheme B	10.496 (8.686)	3.942 (4.873)	0.169 (0.130)
Low JSS * Incentive Scheme A	-2.218 (9.000)	7.959 (5.048)	0.268* (0.135)
Low JSS * Incentive Scheme B	10.942 (9.335)	16.228*** (5.237)	0.348** (0.140)
Constant	29.391 (7.440)	2.799 (4.174)	0.449 (0.112)
Observations	59	59	59
R-squared	0.1336	0.1914	0.1838

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

The fact that the Low Job Success Score/Incentive Scheme B group performed significantly better than the baseline in terms of both correct bios and percent correct suggests that workers with low job success scores do respond to the incentive by improving the quality of their performance. We can conclude that for this experiment, the incentive motivated workers in the Low Job Success Score group to outperform workers with a higher job success score by a statistically significant amount in terms of both the number of bios they processed correctly and the overall percent correct. This suggests that workers with low job success scores may have the most capacity for improvement when offered an incentive, a finding that contradicts Lazear's conclusion that high ability workers have the most capacity to improve.

All of the other groups (with the exception of Medium Job Success Score/Incentive Scheme B) outperformed the baseline in terms of number of correct bios and percent correct. Therefore, this data could suggest that the freelancers with a job success score in the medium range may underperform compared to freelancers with higher or lower job success scores in terms of both of these outcome variables, even when offered an incentive. However, it could also suggest that the Medium Job Success Score group doesn't serve as a true baseline, which is one of the reasons why I then re-categorize the freelancers into modified bins. I raise the minimum to qualify for the "High" bin to 95% from 90%, and also raise the minimum to qualify for the "Medium" bin to 87% from 76%. In comparison to the old bins, the new "High" and "Medium" bins cover a narrower distribution of job success scores, while the "Low" bin covers a much broader distribution.

Regressions with Modified Bins

When I run the same regressions using my modified bins (in an attempt to establish a more accurate baseline), I find that none of the groups are statistically significantly different from the baseline in terms of number of bios processed, though the High Job Success Score/Incentive Scheme A group processed an average of 9.2 more bios, and the High Job Success Score/Incentive Scheme B group processed an average of 4.1 fewer bios. Interestingly, the trend is flipped for the low job success score groups—the Low Job Success Score/Incentive Scheme A group processed an average of 6.8 fewer bios, while the Low Job Success Score/Incentive Scheme B group processed an average of 4.2 more bios.

I find that the Low Job Success Score/Incentive Scheme B group still significantly outperforms the baseline for number of correct bios (by an average of 10.293, p-value of 0.013), and the High Job Success Score/Incentive Scheme A group also outperformed the baseline for number of correct bios (by an average of 8.363, p-value of 0.054). The High Job Success Score/Incentive Scheme B and Low Job Success Score/Incentive Scheme A groups also slightly outperform, by 2 and 2.9 bios, respectively, but these differences are not statistically significant.

None of the groups were statistically significantly different from the baseline in terms of percent correct. However, the High Job Success Score/Incentive Scheme B group outperformed by an average of 9.1 percentage points, and the Low Job Success Score/Incentive Scheme B group outperformed by an average of 14.7 percentage points. As the medium job success score groups no longer underperform all of the other groups in terms of percent correct, these modified bins may provide a more accurate baseline than the bins provided by Upwork.

Table 5.5: Regressions using Modified Bins

	Bios Completed	**Correct Bios**	**Percent Correct**
Jobs Completed	0.009 (0.026)	0.004 (0.014)	0.000 0.000
Hourly Wage	0.476 (0.750)	-0.033 (0.421)	-0.01 (0.012)
High JSS * Incentive Scheme A	9.174 (7.545)	8.363** (4.236)	0.033 (0.118)
High JSS * Incentive Scheme B	-4.113 (3.235)	2.078 (4.004)	0.091 (0.111)

Medium JSS * Incentive Scheme B	3.235 (8.361)	-0.844 (4.694)	0.016 (0.131)
Low JSS * Incentive Scheme A	-6.831 (7.189)	2.845 (4.036)	0.018 (0.112)
Low JSS * Incentive Scheme B	4.247 (7.550)	10.293*** (4.239)	0.147 (0.118)
Constant	34.712 (6.536)	4.603 (3.670)	0.601 (0.102)
Observations	59	59	59
R-squared	0.1244	0.1914	0.1072

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

These results seem to support the interpretation that most workers respond to the incentives, but the responses are different for different subgroups of workers (and in aggregate, cancel each other out). Workers with low job success scores seem to respond to the incentive by significantly increasing their output (as measured by number of bios completed)—we can see that the coefficient in front of the interaction variable between the incentive and the group switches from negative to positive when a low job success score worker switches from Incentive Scheme A to Incentive Scheme B. However, the increase in quantity doesn't appear to come at the expense of quality. Workers in the Low Job Success Score/Incentive Scheme B group significantly outperform the baseline in terms of correct bios, while workers in the Low Job Success Score/Incentive Scheme A group only slightly outperform. The Low Job Success Score/Incentive Scheme B group also has the highest outperformance in terms of percent correct, though this outperformance is not statistically significant.

Workers with high job success scores appear to respond to the the incentive differently—it seems that they either focus too heavily on the quality aspect of the incentive (but are not very successful in improving their accuracy) or are simply demotivated by the incentive. Output (bios processed) declines with the introduction of the incentive, as the coefficient in front of the interaction variable between the incentive and the group switches from positive to negative when a high job success score worker switches from Incentive Scheme A to Incentive Scheme B. Unsurprisingly, the number of correct bios also appears to decrease—the High Job Success Score/Incentive Scheme A group significantly outperforms the baseline in terms of correct bios, but the High Job Success Score/Incentive Scheme B group only slightly outperforms. However, it doesn't appear that quality (as measured by percent correct) declines overall. Quality may even slightly increase, as the High Job Success Score/Incentive Scheme B group outperforms the baseline by 9.1 percentage points while the High Job Success Score/Incentive Scheme A group only outperforms the baseline by 3.3 percentage points, on average.

C. How do incentives influence payment?

I then regress the incentive dummy against the cost-related variables—hourly wage, bios processed per dollar spent, and correct bios per dollar spent. I find that the freelancers working under Incentive Scheme B were paid an average of $0.76 less than the freelancers working under Incentive Scheme A, a difference that is statistically significant (p-value less than 0.00001). There is also a statistically significant difference between the incentive groups in terms of bios processed per dollar—freelancers working under Incentive Scheme B processed an average of 1.78 more bios per dollar spent than

freelancers working under Incentive Scheme A. There was not a significant difference between the groups in terms of correct bios processed per dollar.

Table 5.6: Regressions with Cost Variables

	Wage	**Bios Per Dollar**	**Correct Bios Per Dollar**
Incentive	-0.761*** (0.107)	1.783** (0.843)	0.249 (0.405)
Constant	4.000 (0.074)	6.285 (0.580)	1.339 (0.279)
Observations	59	59	59
R-squared	0.4704	0.0728	0.0066

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

As the average number of bios completed is approximately equal between the two incentive groups but the average worker in Incentive Group B earns a significantly lower wage per hour, it's unsurprising that the average worker in Incentive Group B was paid significantly less per bio processed. Freelancers in Incentive Group B processed an average of 1.80 bios correctly per dollar spent, compared to 1.34 correct bios per dollar for freelancers in Incentive Group A, but this difference is not statistically significant.

VI. Conclusion

This paper examines a relatively new phenomenon—online labor marketplaces—and how employers in these marketplaces can best compensate online freelancers to optimize productivity. I hire 59 freelancers on Upwork, the world's largest online freelancing marketplace, to complete a simple data extraction task. Each freelancer

is randomly assigned to one of two incentive groups: approximately half receive a financial "gift" unrelated to performance (in the form of an increase in wage), and the other half receive a performance-based incentive. Through this experiment, I aim to answer three questions: "How do incentives influence performance, in terms of both quality and quantity?", "How do incentives affect different freelancers differently?", and "How do incentives influence payment?".

I find that the incentive does not have a statistically significant effect on any of the outcome variables used to measure a freelancer's performance (including variables measuring both quantity and quality of work). However, I cannot conclude that financial incentives are ineffective in motivating freelancers to improve performance. Instead, I consider the possibility that this incentive was simply not large enough or that it had opposite effects on different sub-groups of freelancers. If the latter were the case, the overall effect of the incentive could be zero even if the incentive had significant effects on particular groups of freelancers, such as those with low job success scores. It is worth noting that while the incentive didn't have a positive overall effect on performance, it did allow us to pay the freelancers working under Incentive Scheme B a significantly lower rate for each bio processed.

I then attempt to determine whether the effect of the incentive differs based on a freelancer's observable characteristics, measured here as job success score, number of jobs previously worked, and preferred hourly wage. I find that while both a freelancer's job success score and the incentive have a positive effect on number of bios completed, the interaction between the two has a significant negative effect. Although this result seems somewhat counterintuitive, I speculate that freelancers with high job success

scores either focus more on accuracy (which results in a decrease in speed) or find the incentive less motivating because they have opportunities to earn higher wages in other jobs.

I also sort the freelancers into three categories by job success score (low, medium, and high). I find that freelancers with a low job success score underperform the baseline in terms of correct bios when not offered an incentive, but outperform the baseline when they are offered an incentive. This trend flips for freelancers with high job success scores—they outperform the baseline when not offered an incentive, and underperform when offered an incentive.

In terms of number of correct bios, freelancers with a low job success score working under Incentive Scheme B and freelancers with a high job success score working under Incentive Scheme A outperform the baseline. This seems to support the conclusion that freelancers with low job success scores are highly motivated by the incentive—and the number of bios that they process correctly increases. However, freelancers with high job success scores do not significantly improve their performance when offered incentives. In fact, while freelancers with high job success scores outperform the baseline in terms of correct bios with no incentive, this outperformance disappears when they have a chance to earn the performance based-incentive.

The low and high job success score groups outperform the baseline in terms of percent correct under both incentive schemes. The outperformance is slightly higher (by three percentage points) for freelancers with high job success scores when they are offered an incentive compared to when they are not offered an incentive, which at least suggests that the incentive does not result in a decrease in work quality for this group. For

freelancers with low job success scores, the difference in outperformance between the incentive group and no incentive group is more significant (eight percentage points), indicating that the incentive motivates these freelancers to improve their accuracy by a non-negligible amount.

Though my results seem to suggest that financial incentives do affect performance for specific sub-groups of freelancers, more research needs to be done before attempting to generalize my conclusions to a broader range of workers. As a result of my limited budget, I was only able to hire around 60 freelancers, which is a fairly small sample size. To make a definitive conclusion about whether this type of financial incentive is effective in improving performance, more research needs to be done with a larger number of freelancers. In addition, future research building off this paper could focus on how different amounts and types of financial incentives affect performance differently to answer the question of how to optimize freelancer performance while minimizing costs.

VII. References

Cabral, L. and Hortacsu, A. 2010. "The Dynamics of Seller Reputation: Evidence from eBay," *Journal of Industrial Economics*, 58(1): pp. 54–78.

Dana, J. and Spier, K. 1993. "Expertise and Contingent Fees; The Role of Asymmetric Information in Attorney Compensation," *Journal of Law, Economics, & Organization*, vol. 9: pp. 349-67.

Harris, C. 2011. "You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks." ACM International Conference on Web Search and Data Mining.

Heywood, J., Siebert, S. and Wei, X. 2013. "The Consequences of a Piece Rate on Quantity and Quality: Evidence from a Field Experiment." Institute for the Study of Labor. Discussion Paper No. 7660.

Holmstrom, B., and Milgrom, P. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, vol. 7: pp. 24-52.

Horton, J. and Chilton, L. 2010. "The Labor Economics of Paid Crowdsourcing," *Proceedings of the 11th ACM Conference on Electronic Commerce*.

Horton, J. and Golden, J. 2015. "Reputation Inflation: Evidence from an Online Labor Market." New York University. Working Paper, February.

Houser, D. and Wooders, J. 2006. "Reputation in Auctions: Theory and Evidence from eBay," *Journal of Economics & Management Strategy*, vol. 15: pp. 353-369.

Jin, G. Z. and Kato, A. 2006. "Price, Quality and Reputation: Evidence from an Online Field Experiment," *The RAND Journal of Economics*, 37(4): pp. 983-1005.

Kerr, S. 1995. "On the Folly of Rewarding A, While Hoping for B," *The Academy of Management Executive*, 9(1): pp. 7-14.

Lazear, E. 1996. "Performance Pay and Productivity," *American Economic Review*, 90(5): pp. 1346-1361.

Livingston, J. 2002. "How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions," *Review of Economics and Statistics*, 87(3): pp. 453-465.

Mason, W. 2009. "Financial Incentives and 'The Performance of Crowds.'" Proceedings of the ACM SIGKDD Workshop on Human Computation.

Melnik, M. I. and Alm, J. 2002. "Does a Seller's Reputation Matter? Evidence from eBay Auctions," *Journal of Industrial Economics*, 50(3): pp. 337-349.

Pallais, A. 2014. "Inefficient Hiring in Entry-Level Labor Markets," *American Economic Review*, 104(11): pp. 3565-99.

Resnick, P., Zeckhauser, R., Swanson, J. and Lockwood, K. 2006. "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics*, vol. 9: pp. 79-101.

Ritter, J. and Taylor, L. 1999. "Low Powered Incentives," Federal Reserve Bank of St. Louis. Working Paper, May.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M. 2011. "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets." Conference Paper, January.

Sauermann, J. 2014. "The Heterogeneous Effect of Bonus Pay on Performance Outcomes: Evidence from Personnel Data." Beiträge zur Jahrestagung des Vereins für Socialpolitik. Conference Paper No. D02-V3.

Shaw, A. D., Horton, J.J. and Chen, D.L. 2011. "Designing Incentives for Inexpert Human Raters." Berkman Center for Internet & Society. Research Publication No. 2011-02, March.

Shearer, B. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment," *Review of Economic Studies*, 71(2): pp. 513–34.

Shi, L. 2010. "Incentive Effect of Piece-Rate Contracts: Evidence from Two Small Field Experiments," *B. E. Journal of Economic Analysis and Policy,* 10 (1).

Standifird, S. 2001. "Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings," *Journal of Management*, vol. 27: pp. 279-295.

Wolf, J., and Muhanna, W. 2005. "Adverse Selection and Reputation Systems in Online Auctions: Evidence from eBay Motors," ICIS 2005 Proceedings. Paper No. 67, December.

Ying, M., Chen, Y., and Sun, Y. 2013. "The Effects of Performance-Contingent Financial Incentives in Online Labor Markets," AAAI 2013 Proceedings, July.

VIII. Appendices

Appendix 1

Table 1.1: Geographic Locations of Hired Freelancers

Geographic Location	Frequency	% of Total
Bangladesh	13	22.41%
India	16	27.59%
Pakistan	8	13.79%
Philippines	12	20.69%
Serbia	2	3.45%
Other (includes all countries with <2 freelancers)	7	12.07%

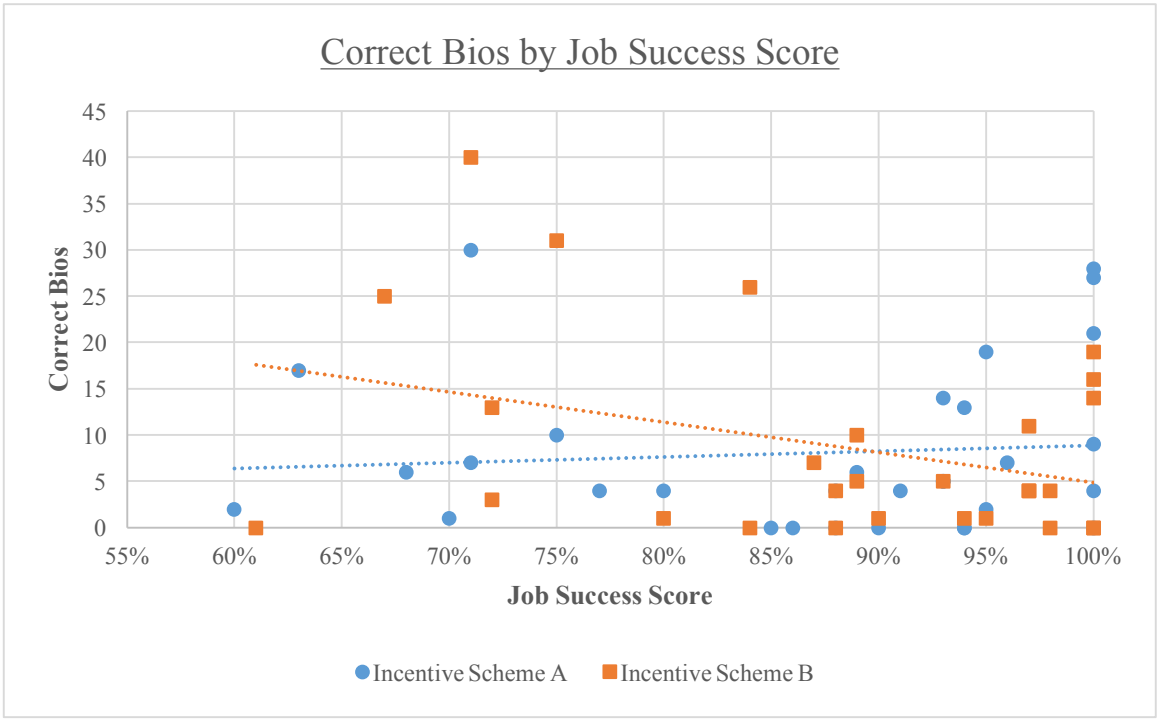
Table 1.2: Upwork Experience of Hired Freelancers

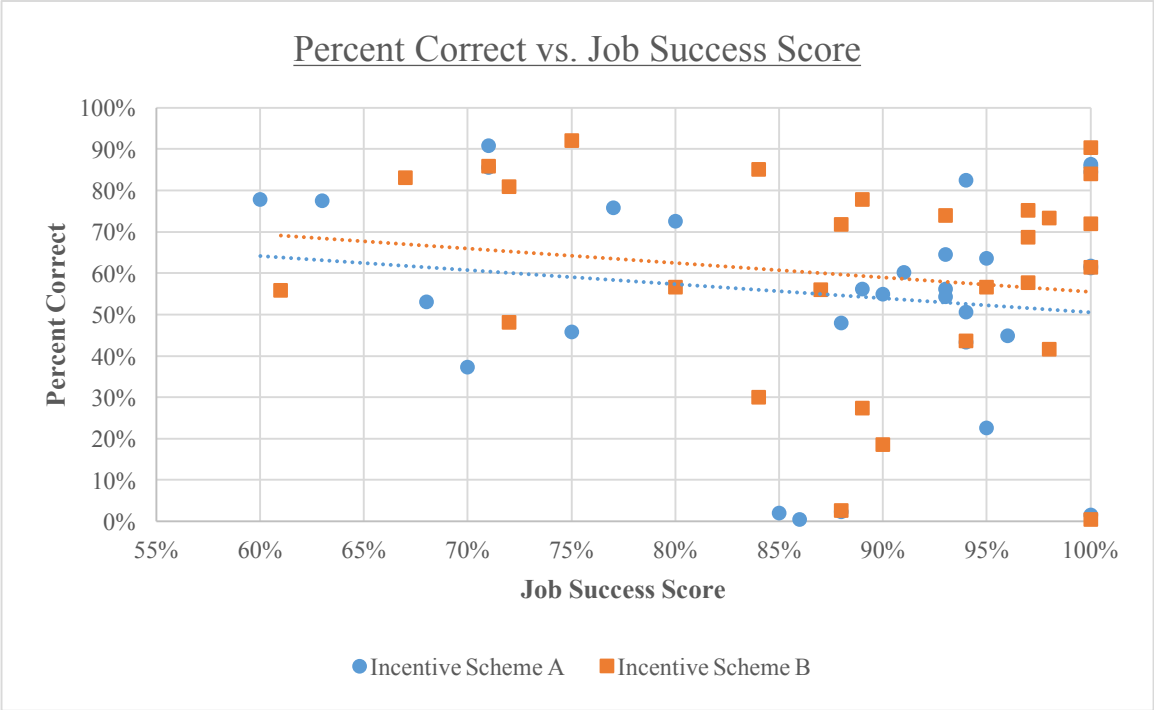
Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Hourly Rate*	59	$5.55	$3.16	$3.00	$22.00
Jobs Worked	59	56.20	99.07	2	655
Hours Worked**	55	1976.33	3307.01	1	17290

*The hourly rate is the preferred rate listed on the freelancer's profile, which is supposed to represent the minimum wage that the freelancer will accept. This may or may not be the same as the rate that the freelancer bids for a particular job.
**The observations for this variable do not include freelancers who had only worked fixed wage jobs in the past (and therefore had no record of hours worked).

Appendix 2

Additional Graphs




Percent Correct vs. Job Success Score
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
55%
60%
65%
70%
75%
80%
85%
90%
95%
100%
Job Success Score
Incentive Scheme A
Incentive Scheme B

Appendix 3

Suggestions for Future Studies

After completing this experiment and analyzing my data, I reflected on what I would do differently in subsequent studies to more precisely answer the questions of how incentives influence performance and how the effect of incentives differs between freelancers. If I were able to redo my experiment with a more substantial budget, I would like to make the following changes:

1. Hire more freelancers.

With a budget of $500, I could only afford to hire approximately 60 freelancers to work for a total of two hours (including 30 minutes of training) each. With a larger budget, I could hire more freelancers with varying job success scores and other characteristics, which would more accurately represent the broader population of freelancers on Upwork. Ideally, each freelancer would also be hired to work for a longer period of time (e.g. three hours instead of one and a half), so I could be more confident that differences in performance could be attributed to the incentive and not to the fact that some freelancers may take longer to acclimate to the task than others. In addition, having performance data over a longer period of time could allow me to study whether the effect of the incentive varies over time—for example, if the freelancer seems to care more about earning a bonus in the first hour versus the third hour of working.

2. Make the threshold to qualify for a bonus lower.

As I mentioned earlier in the paper, I structured the incentive payment in hopes that the average hourly wage for freelancers working under Incentive Scheme B would be $4/hour. However, this did not turn out to be the case—the average hourly wage for

freelancers working under Incentive Scheme B was $3.27. If I were to redo this experiment, I would want the average hourly wages to be the same for both incentive groups to ensure that differences in performance were not due to differences in wages, so I would need to find a way to raise the average wage for freelancers working under Incentive Scheme B. I could do this by lowering the threshold to qualify for a bonus, raising the amount of the bonus, or both.

Given that 79% of the freelancers working under Incentive Scheme B did not qualify for any bonus, I would first want to focus on lowering the threshold. Among the full group of freelancers, the average number of correct bios per hour was 5.6. Therefore, assuming that I raised the amount of the bonus slightly (e.g. $0.20 per correct bio beyond the threshold), I would need to lower the threshold to around one correct bio per hour to expect an average wage of $4/hour for freelancers working under Incentive Scheme B. Alternatively, I could raise the amount of the bonus quite significantly (e.g. $0.50 per correct bio beyond the threshold), and lower the threshold to around three correct bios per hour. However, I would need to further consider these proposals (and potentially run small pilots to test them) before deciding on one. It's possible that lowering the threshold significantly to a level that seems more achievable would motivate the freelancers enough that raising the amount of the bonus wouldn't be necessary.

3. **Use a different freelancing website or wait until Upwork's job success score system is more established.**

As the job success score was so new to Upwork when I conducted my experiment, and the formula for calculating the score is not public, it's possible that the freelancers were extremely apprehensive of the system. This could have significantly affected the freelancers' behavior. I hypothesized earlier in the paper that if the

freelancers didn't fully understand how the job success score was calculated and wanted to establish a high score in the new system, they might have exerted their maximum effort without the incentive. As a result, the effect of the incentive could have been severely diminished. To eliminate this issue in a future experiment, I would likely want to wait until freelancers are more comfortable with the new job success score system and it is less likely to significantly affect their behavior. Alternatively, I could switch to a different online freelancing marketplace with a more established reputation system. Upwork seems to be the ideal platform for this experiment given the fact that it is by far the most popular online freelancing marketplace, and there might be something different about the freelancers that choose to use another website. For this reason, I would likely want to wait until freelancers become more comfortable with the job success score system on Upwork instead of trying to find another online freelancing platform to use for the experiment.

Appendix 4

Full Freelancer Performance Results

The table below contains the full performance metrics for each freelancer. In the second column, Incentive, 0 means that the freelancer was working under Incentive Scheme A ($4/hour wage), and 1 means that the freelancer was working under Incentive Scheme B ($3/hour base wage with the opportunity to earn a performance-based bonus). The next four columns include information regarding the freelancer's characteristics before beginning the experiment. All of the columns following the final characteristic column ("Hourly Wage," which contains the freelancer's preferred hourly wage) contain information regarding the freelancer's performance in the experiment.

"Files Completed" refers to the number of full files that the freelancer was able to process (bios were contained within files, and files had varying number of bios).

Freelancer	Incentive	Job Success Score	Jobs Worked	Hours Worked	Hourly Wage	Files Completed	Bios Completed	Correct Bios	Points Received	Possible Points	Percent Correct
1	0	93%	3	13	$10.00	12	38	5	163	300	54%
2	0	88%	4	1	$3.00	22	58	4	161	336	48%
3	0	60%	6	0	$5.00	3	12	2	74	95	78%
4	0	70%	6	0	$6.00	3	23	1	66	177	37%
5	0	63%	7	160	$3.33	21	54	17	241	311	77%
6	0	75%	7	407	$3.33	8	12	10	114	249	46%
7	0	100%	7	539	$4.44	17	42	21	233	270	86%
8	0	100%	7	900	$3.33	24	57	0	5	331	2%
9	0	100%	9	0	$3.00	15	39	4	158	256	62%
10	0	85%	9	117	$3.33	24	57	0	5	331	2%
11	0	91%	11	41	$8.89	18	43	4	165	274	60%
12	0	94%	12	77	$3.33	22	60	0	231	533	43%

13	0	68%	14	71	$3.33	6	23	6	77	145	53%
14	0	100%	14	587	$4.44	18	48	9	177	289	61%
15	0	71%	17	3,237	$3.50	10	31	7	187	206	91%
16	0	93%	21	697	$3.50	6	22	5	86	153	56%
17	0	80%	23	2,996	$4.44	4	22	4	111	153	73%
18	0	96%	25	41	$6.67	16	41	7	115	256	45%
19	0	77%	25	5,525	$8.00	5	19	4	94	124	76%
20	0	90%	28	352	$5.56	6	21	0	79	144	55%
21	0	100%	30	3,653	$3.33	20	50	27	262	305	86%
22	0	94%	33	686	$5.56	18	48	0	152	300	51%
23	0	71%	33	1,118	$6.00	21	54	30	273	319	86%
24	0	93%	35	609	$3.33	20	50	14	197	305	65%
25	0	86%	78	3,138	$6.67	18	27	0	1	179	1%
26	0	95%	93	2,323	$4.00	8	27	2	42	186	23%
27	0	100%	105	13,818	$5.56	22	57	28	284	336	85%
28	0	94%	117	2,226	$5.00	7	25	13	146	177	82%
29	0	95%	149	3,117	$5.00	22	57	19	177	278	64%
30	0	89%	181	2,097	$6.00	15	41	6	144	256	56%
31	0	88%	203	5,180	$7.00	10	11	0	2	87	2%
32	1	75%	2	7	$7.78	19	49	31	276	300	92%
33	1	97%	2	16	$4.44	8	26	4	119	206	58%
34	1	94%	5	24	$4.44	9	31	1	94	215	44%
35	1	98%	6	48	$3.00	5	11	0	99	238	42%
36	1	100%	8	705	$5.00	7	25	0	123	200	62%
37	1	61%	9	259	$5.00	11	32	0	143	256	56%
38	1	100%	10	0	$15.00	42	92	0	3	554	1%
39	1	71%	12	5	$4.00	22	61	40	405	472	86%

40	1	67%	18	1,993	$5.56	17	48	25	211	254	83%
41	1	89%	21	47	$8.00	13	34	10	210	270	78%
42	1	80%	21	1,494	$4.00	12	34	1	127	224	57%
43	1	72%	24	750	$3.33	9	28	13	157	194	81%
44	1	84%	26	65	$9.00	17	48	26	246	289	85%
45	1	87%	27	137	$5.00	23	60	7	198	353	56%
46	1	97%	28	868	$5.56	11	40	11	206	300	69%
47	1	95%	29	3,593	$3.33	12	34	1	127	224	57%
48	1	72%	30	1,429	$3.33	5	42	3	130	270	48%
49	1	97%	50	6,211	$3.33	6	23	4	115	153	75%
50	1	100%	54	17,290	$5.56	8	28	16	168	186	90%
51	1	100%	64	548	$3.33	10	31	14	173	206	84%
52	1	84%	71	161	$5.00	15	33	0	46	153	30%
53	1	88%	72	1,741	$4.00	37	67	0	9	336	3%
54	1	100%	85	548	$3.32	17	42	19	208	289	72%
55	1	98%	88	9,760	$10.00	5	18	4	91	124	73%
56	1	93%	116	2,366	$6.00	6	26	5	131	177	74%
57	1	88%	143	119	$10.00	8	18	4	89	124	72%
58	1	90%	328	268	$22.00	17	43	1	51	274	19%
59	1	89%	655	4,520	$5.00	25	68	5	111	405	27%