Effort, Productivity, Wage Premia, and Reservation Wages: A Field Test of the Efficiency Wage Theory

Nail Hassairi

Department of Economics

University of Washington

November 3, 2016

Abstract

This paper tests the shirking and adverse selection models of efficiency wage theory by investigating the relationship between effort, wages, and reservation wages in an online labor market, using a field experiment. The experimental design controls for many of the confounders that plague other tests of the efficiency wage theory. Additionally, it allows for a clean identification of the shirking and adverse selection models individually, in order to see which one is driving the efficiency wage effect. The results suggest causal effect of wages on effort that is due to incentive not to shirk, as well as an effect that is due to the selection of more able workers. The selection effect is stronger than the incentive effect. There appears to be heterogeneity in the way workers respond to incentives with workers that aim for longer tenures responding more positively. Dynamic effects are also detected, according to which a given worker provides more effort the more of her tenure is still ahead of her.

1 Introduction

Efficiency wage theories came about as an attempt to explain a variety of puzzling labor market phenomena – non-competitive wage premia, job queues, inter-industry wage differentials, and more – by pointing to the presence of information asymmetries and modeling agents' behavior in response to these asymmetries. The theory comes in various flavors – the shirking model (Shapiro and Stiglitz [1984] and Bulow and Summers [1986]), the sorting model (Weiss [1980]), the labor turnover model (Salop [1979]), and the gift-exchange model (Akerlof [1982]) – that could all possibly contribute to its claim of a link between wages and productivity causing the occurrence of wage premia and employment rents.

The most popular version of the theory – the shirking model – is based on the idea that in a frictionless labor market unobserved worker behavior leads employers to offer wage premia to create a stake in the employment relationship. This stake (wage premium) is putting their wage above their reservation wage (outside options, competing job offers) and motivates them to perform adequately to avoid losing the job. Monitoring is an important part of the model and guarantees that workers do not shirk in equilibrium. The unemployment effect results from a cascade effect that sees other firms offering wage premia to compete for workers once the first firm offers them.

Carmichael et al. [1985] asks why people would not be willing to buy jobs if they were locked out of the primary jobs sector when they could do this possible via re-location or some other mechanism. Workers can post performance bond to eliminate the impact of information asymmetry (and lose the bond in case they are found shirking). These arguments have come to be known in the literature as the "bonding critique". Shapiro and Stiglitz [1985] counter the "bonding critique" by saying that another efficiency wage mechanism might prevent the market for jobs from emerging. The sorting model of the efficiency wage theory (originally due to Weiss [1980]) implies that should wage fall the more able workers would be put off from applying for the job (possibly in favor of a different job or self-employment).

Consequently, to evaluate the empirical relevance of the efficiency wage theory it is important to establish not only whether the general idea of wages being linked to productivity holds but more specifically what kind of behavior drives it – whether it

is in fact a combination of the behavior presented in the shirking and sorting models. The only papers to attempt the identification the sorting and shirking models within a single dataset were Cappelli and Chauvin [1991] and Esteves-Sorenson, Pohl, and Freitas [2016]. Cappelli and Chauvin [1991] did great work on discussing the shirking model and the gift-exchange model, however, their data was ill-suited for discussing the the sorting model, and they did so only as a bonus to their main analysis. Esteves-Sorenson et al. [2016] did great work discussing the role of monitoring and wage premia on performance, however, their discussion of the sorting model misses the mark by discussing performance heterogeneity of observationally distinguishable workers – an issue devoid of any information asymmetry.

My paper utilizes data from a field experiment. The experiment takes place over the course of six days, a period short enough to believe that opportunity costs and reservation wages of the workers would not change significantly due to learning. Workers' wages were varied, allowing me to estimate incentive effect of wages while learning about workers' reservation wages and estimating the impact of its heterogeneity on workers' performance. The environment in which the experiment takes place protects workers' anonymity so that there are no observable characteristics. An additional advantage of this environemnt is that it resembles the secondary job market – low wages, routine/skill-less work – which motivated the efficiency wage theory in the first place (because of the incidence of job queues). The controlled nature of the experiment and the institutional environment together allow me to succeed in separating the effects of worker heterogeneity and incentives on performance where others' attempts were less than convincing. Cappelli and Chauvin [1991] and Esteves-Sorenson et al. [2016] were either not able to observe significant variation in wages for a given worker or they could not ascertain whether this variation was induced from the supply or demand factors.

I construct a simple static model that incorporates both the heterogeneous agent model assumption implying that performance is a function of workers' acceptance/reservation wages, as well as the incentive effect of the currently offered wage. This model contains elements of revealed preference in its cost of effort function. Ability lowers the cost of effort in this function. By choosing an effort level worker reveals the impact of her ability on her effort decision. The model also endogenizes monitoring or rather the worker's belief of being found shirking. The probability of being paid is made a

function of both ability (reservation wage) and effort. I derive the optimal level of effort as a function of reservation wage, current offered wage, job characteristics and worker's experience. The optimal effort decision reflects the worker's trade-off between the benefits (higher probability of being paid) and costs of effort and its dependence on her ability. Subsequently, I estimated the parameters from this model using the experimental data. These estimated parameters allow me glean the comparative impact of ability/heterogeneity (reservation wage) and incentives (actual wage) on effort within the environment in which the experiment took place. My estimates show substantial impact of both incentives and heterogeneity on performance lending plausibility to Stiglitz's defense of efficiency wage theory against Carmichael's "bonding critique."

Section 2 describes the various efficiency wage theories and evidence brought to bear on them. Section 3 describes the experimental design, the Mechanical Turk labor market, and the structure of the data. Section 4 describes my theoretical model and identification strategy. Section 5 provides results from the analysis of the experimental data. Section 6 concludes.

2 Current Theory and Evidence

In this section I will reproduce the theory from the shirking and sorting models along with empirical evidence gathered so far pertaining to these models. Where a testable implication is exclusive to one of the models the evidence will be described along with the theory of this model, otherwise it will be described after the theoretical exposition of both models. Since in my paper I am focusing primarily on the sorting and shirking models to provide evidence on whether job queues can be eliminated by workers buying jobs, I am not discussing the remaining labor turnover and gift-exchange efficiency wage theories.

2.1 The Shirking Model

The following exposition comes from Shapiro and Stiglitz [1984]. There are N identical workers[1]. A worker's instantaneous utility is $U(w, e)$ where w is the wage received and

[1] Note the contrast with the adverse selection theory that starts from the assumption that there are different 'types' of workers.

e is the level of effort provided by the worker. The workers experience disutility from effort and positively value income[2]. The utility function is separable in wage and effort and workers are risk neutral: $U = w - e$. Effort takes on two levels; Shapiro and Stiglitz [1984] claim that continuous effort would not materially change the implications of their model. When a worker is unemployed, she receives unemployment benefits of $\bar{w}$ (and $e = 0$).

Each worker is either employed or unemployed at any point in time. There is probability b per unit time that a worker will be "separated" from her job due to exogenous reasons (other than shirking, e. g. layoffs, recession etc). Separated workers enter the unemployment pool. Workers maximize the expected present discounted utility with a discount rate $r > 0$:

$$W = E \int_0^\infty u(w(t), e(t)) exp(-rt) dt \tag{1}$$

The only choice workers make is the level of an effort level. If a worker does not shirk, she receives a wage of w and will keep her job until exogenous factors cause a separation. If she shirks she will be caught with probability q. If the worker is caught shirking she will be fired, landing in the "unemployment pool". The probability of getting a job while in the unemployment pool determines the expected length of the unemployment spell. Workers in the unemployment pool receive unemployment compensation $\bar{w}$.

By choosing the effort level the worker effectively chooses between the following income streams. V_E^S is the expected lifetime utility of an employed shirker, V_E^N as the expected lifetime utility of an employed nonshirker, and V_u is the expected lifetime utility of an unemployed worker. Shapiro and Stiglitz [1984] derive the fundamental asset equation for a shirker:

$$rV_E^S = w + (b + q)(V_u - V_E^S) \tag{2}$$

and for non-shirker:

$$rV_E^N = w - e + b(V_u - V_E^N) \tag{3}$$

[2] $\frac{\partial U}{\partial w} > 0$ and $\frac{\partial U}{\partial e} < 0$.

Shapiro and Stiglitz [1984] then go on to combine the two equations above and arrive at the no-shirking condition:

$$w \geq rV_u + \frac{(r+b+q)e}{q} \equiv \hat{w}. \tag{4}$$

This could be solved for effort (dropping b for simplicity):

$$e^* = \frac{q}{r+q}w - \frac{rV_u}{r+q} \tag{5}$$

For the purposes of testing this model it would be useful to see how effort responds to monitoring:

$$\frac{\partial e^*}{\partial q} = r\frac{w+V_u}{(r+q)^2} \tag{6}$$

and how effort responds to wages:

$$\frac{\partial e^*}{\partial w} = \frac{q}{r+q} \tag{7}$$

Consequently, the first testable implication of this model is that wages are positively correlated with effort. The magnitude of this relationship is driven by the probability of being caught shirking, q, and the worker's discount rate, r. As I will show later this implication is very similar to the assumption of labor supply behavior in the sorting model, hence most evidence regarding this implication/assumption will be shown after the presentation of the sorting model efficiency wage theory in Section 2.4.

Another testable implication of the shirking model is that monitoring is a substitute for wages for the principal. An empirical test could be this aspect of the model (Rebitzer [1995]). Rebitzer [1995] find a negative link between supervision and wages. Their approach is based on the idea that there is an isoquant along which worker productivity is constant. One can move along this isoquant by trading off monitoring for high wages or vice versa. Groshen and Krueger [1990] reach the same conclusion using hospital data on supervision and wages. Prendergast [1999] points out that this test does not differentiate between efficiency wage theory and basic agency model with no rents. Prendergast [1999] proposes a better test: "In order to test for rents, one would need to see, for example, whether higher levels of supervision within a job

increase worker turnover (since more supervisors reduce wages)." Prendergast [1999] also wonder about the source of variation in supervision and wages across firms and points out that different firms would have different returns to supervision and effort. In such a case wages and supervision could be either complements or substitutes to each other even if the efficiency wage theory holds. One also cannot guarantee that I am observing multiple different firms along the same isoquant of production. Consequently, Prendergast [1999] questions how one can form refutable hypotheses on the relationship between supervision and wages in relation to the efficiency wage theory.

Esteves-Sorenson et al. [2016] have data from Portuguese tourist sector companies and their workers. They take advantage of a Portuguese labor market institutional feature which grants workers tenure (makes workers harder to fire) after a certain apprenticeship period. This allows Esteves-Sorenson et al. [2016] a great opportunity to estimate the effect of lower monitoring on productivity and they indeed find an effect consistent with the efficiency wage theory (tenure/lower probability of being fired increased absenteeism).

2.2 The Heterogeneous Agents Model

This model is variously referred to in the literature as a 'sorting model', 'selection model', 'adverse selection model', or a 'heterogeneous agents model'. I prefer the latter name since it captures the big difference between this model and the shirking model, namely that while the shirking model assumes all agents to be the same (homogeneous) and behaving differently in different circumstances the selection model assumes that there are fundamentally different (heterogeneous) agents which drives the difference in their behavior. At the same time, I will sometimes refer to it just as a "sorting model" because it's shorter.

While moral hazard theory is based on the idea that workers are the same but behave differently under different circumstances this theory is based on the idea that workers display constant behavior defined by their ability. Effort is not discussed in Weiss [1980] and one can only assume that in this model workers always perform at their potential. As a result, monitoring plays no role in this model. Incentives do not affect agents' behavior. The pool of job applicants consists of different yet observationally

indistinguishable agents. Their difference comes from "effective labor endowments"[3].

Each worker has a labor endowment (θ) which is invariant across all jobs and an acceptance wage (w) which is an increasing function of his marginal product and is strictly greater than that marginal product. Therefore, the acceptance wage of a worker is a strictly increasing function of his labor endowment θ, so that

$$\theta = q(w) \tag{8}$$

and $q'(w) > 0$.

Weiss [1980] introduces two versions of the model, one with indistinguishable workers, another with distinguishable groups of indistinguishable workers. The former model is very stylized as it would have to assume there is no resume being collected and no interviews conducted since these would in all likelihood make the workers distinguishable from each other. Interestingly enough, however, this feature of the model fits quite neatly with how recruitment of labor actually works on Mechanical Turk which is why I choose to describe this model here.

Weiss [1980] then proceeds to describe firm behavior and the market equilibrium, which is not very relevant for my analysis as I am testing the veracity of the labor supply aspect of this model, so I will skip the exposition of those aspects of the model here. As we can see, the relationship between the effective labor endowment and acceptance wages is posited rather than explained in this model. The contribution of this model is in linking this assumption to job layoffs by modeling firm's hiring decision.

2.3 Shirking and Sorting

Cappelli and Chauvin [1991] also test the distinction between the shirking model and the selection model described in Weiss [1980]. They do this by including the wage premium from the year 1972 (the rest of the data is from 1982). Because the median worker has a tenure of 10 years in the dataset this would have been the wage that

[3]These effective labor endowments can have nothing to do with education or experience because those are visible applicant characteristics. One can then only assume (is not elaborated on in Weiss [1980]) that these effective labor endowment relate to features like teamwork, conscientiousness, initiative, creativity, leadership, responsibility, ambition, talent, communication skills etc.

was used to recruit the median worker in the dataset. My study is closely related to Cappelli and Chauvin [1991]. Like them I would like to provide evidence on the sorting and shirking models. In Cappelli and Chauvin [1991] the sorting model is an afterthought. Main emphasis is on the gift-exchange and shirking models. Using pay in 1972 is circumstancial evidence at best. If workers were being hired between 1962 and 1982 with the median being hired in 1972 than the 1972 wage would be relevant only for workers hired exactly in 1972, not 1973 or 1971. This is likely to be a very small proportion of the sample. Given this it is not surprising that Cappelli and Chauvin [1991] report a very noisy estimate for this 1972 wage in their regression and their estimate is even positive suggesting that higher wage from 1972 leads to higher disciplinary employee dismissal. In my paper the same worker was offered different wages over the course of the experiment which allows me to construct a measure of reservation wage by looking at the lowest implied hourly wage accepted. Furthermore, for every task for each task I have the wage actually offered to the worker at that exact time allowing me to closely estimate the effect of screening on wages. Finally, recruitment in the labor market where my study is conducted does not allow the employers to know anything about the workers so screening on wages is the only screening tool available allowing for a clean estimation of its effect.

Esteves-Sorenson et al. [2016] also study the sorting model. They show that workers granted tenure have stronger background credentials (higher education etc). This has nothing to do with the original sorting model described in Weiss [1980]. The information on workers' educational credentials is easy to obtain and in fact is a standard information being collected by employers and is no source of informational asymmetry in the labor market. The argument made in Weiss [1980] is that offered wages serve as a tool to screen workers on unobservable characteristics. In this sense offering higher wages is a complement to other recruitment practices. If one could simply look at educational credentials to find higher productivity employees offering wage premia as a screening device would a wasteful activity.

In this sense, analysis in Esteves-Sorenson et al. [2016] simply finds that some firms pay more for higher education. This is not a premium wage but rather paying an appropriate wage for a worker with given skillset and is exactly the kind of identification problem that Cappelli and Chauvin [1991] warned about.

Esteves-Sorenson et al. [2016] are making another argument they claim supports the adverse selection story. They show that post-tenure shirking increases but less so than in firms that do not pay premium wage. They attribute this to the adverse selection story. However, this is perfectly consistent with the shirking model. The workers in the premium firms may have more security but they also have higher wages. These two work in opposite directions and if the impact of the premium wages is stronger than the shirking model would still predict that shirking in the premium firms would be lower than in the benchmark market-wage-paying firms even after accounting for the increased security of employment[4].

Both Esteves-Sorenson et al. [2016] and Cappelli and Chauvin [1991] simply assume that workers are receiving wage premia but the assumption that these wage premia are due to unobserved firm or worker productivity is an equally plausible assumption. My study does not suffer from this problem since the wages are randomly assigned and I do not interfere with worker self-selection at all. All workers willing to work are hired. The rest of this section describes studies investigating the efficiency wage theory broadly and whose findings could equally well be ascribed to the selection as well as the shirking effect.

2.4 Evidence on Wages and Shirking

Early anecdote on the existence of efficiency wages comes from Smith [1776] and Raff [1988]. Cappelli and Chauvin [1991] use a plant level dataset from a single large manufacturing firm with multiple plants within the same geographic area (the midwestern US). The hope is that by using this dataset they are able to avoid all the pitfalls mentioned above. They make the claim that in their data shocks to productivity would happen to all the plants in the study simultaneously. For example, all workers are covered by the same union (United Auto Workers). "Management's personnel policies on issues such as shirking and discipline are centrally controlled and are generally identical across all plants as are the union's policies for dealing with management on these issues." They find a negative relationship between the incidence of disciplinary

[4]Tenure does not mean that it's impossible to fire workers. So if a worker has a non-zero probability of being fired when caught shirking than an increase in wage would increase the stakes for this worker, how much she loses by being fired. The shirking model would predict increased effort in this case.

employee dismissals and wage premia (paid over the average wage in the area paid by other companies for comparable jobs).

2.5 Relative Wages and Output in Firm Level Data

Wadhwani and Wall [1991] use firm level panel from the UK to estimate a production function that includes unemployment and relative wages along with hired labor as determinants of efficient units of labor. They estimate positive and statistically significant elasticities of output (real sales) with respect to relative wages and unemployment in the industry as well as labor hired by the firm. This is important evidence because it shows that from the firm viewpoint it can be efficient to pay efficiency wages. Observational studies are not without their challenges, however. Cappelli and Chauvin [1991] note that it is difficult to identify and control for exogenous, nonwage factors that might affect worker productivity (differences in worker quality and the nature of their jobs). They also note that it is difficult to establish direction of causation between high relative wages and worker productivity.

Levine [1992] investigates whether it is in fact profitable for the firm to increase wage on the margin in terms of increased productivity. He finds that the elasticity of output with respect to wages is such that marginal wage increases do lead to higher productivity. Esteves-Sorenson et al. [2016] also conduct this type of analysis and find gains from paying wage premia and from recruiting workers with better employment credentials (education, experience etc. The productivity gains in their paper is measured as money not paid for paid absences of workers.

2.6 Inter-Industry Wage Differences

Under the efficiency wage theory wage premia exist that are unrelated to productivity. These premia do not exist in the competitive model. Krueger and Summers [1988] set out to test for the presence of wage premia that are not explained by regional, demographic and human capital variables. They reject the null hypothesis of no inter-industry wage differentials. They found large effects of industry affiliation on wage differentials which they attribute to non-competitive rents and efficiency wages. The exact efficiency wage mechanism in effect was not a subject of their study.

2.7 Firm Size Wage-Premia, Tournaments

Some authors have speculated that workers receive higher rents in bigger firms. Several explanations were proposed for why firm would have the rent source to be able to pay these. The firm would then share these rents to ensure effort provision in line with the shirking model. Looking at the specific labor market in law firms it has been shown that this is in fact not the case (Rebitzer and Taylor [1995]). Associates in law firms go through a tournament like "associate" period before they are either fired or promoted to a partner status. This then ensures effort provision voiding the need to provide efficiency wages. If bigger firms pay higher wages due to efficiency wage theory this would mean that big law firms should not be paying relatively higher wages. But they in fact do which means that the large firm wage premia are not caused by the efficiency wage theory (Rebitzer and Taylor [1995]). This claim, however, rests on the assumption that the tournament scheme employed in law firms fully resolves the moral hazard problem inherent in the shirking model (this is contested by Akerlof and Katz [1986]).

2.8 Issues with Empirically Assessing the Efficiency Wage Theory

This section has highlighted the empirical contributions made thus far in the quest to (in)validate the efficiency wage theories. As I have noted in the introduction various methods have their advantages and limitations. My contribution to this literature is an experimental design that allows for clean identification of the structural relationships of interest together with the advantages of conducting the experiment in the field – real workers, real labor market, real decisions). Furthermore, while a great attention has been focused on the shirking and gift-exchange models the sorting model has usually been treated as an afterthought if discussed at all despite the fact it is the confluence of sorting and shirking that provides a theory robust to potential criticism in the form of the "bonding critique". The next section will describe the institutional environment in which our field experiment took place.

3 Experimental Design

3.1 The Mechanical Turk Labor Market

Employers can post almost anything as a job on Mechanical Turk; examples include transcribing audio recordings into text, reviewing products, rewriting paragraphs, labeling images, searching for information, data entry, and answering surveys.

Amazon's Mechanical Turk is the largest and most flexible of the emerging microtask markets. Anyone can register to post jobs on Mechanical Turk and the main restriction for people looking to work is that they have to be 18 years or older. The individual tasks in a job are called HITs (Human Intelligence Tasks).[5] The suppliers of labor are "workers" and the agents demanding labor are "requesters." Mechanical Turk has over 100,000 registered workers from over 100 countries [Buhrmester, Kwang, and Gosling, 2011].

Figure 1 shows an example of available jobs on Mechanical Turk. Each job has a title and description, and the worker can preview a job before accepting it, and abort the job without penalty at any time. Workers choose jobs from the list, which can be sorted by criteria such as pay and posting date, or searched by keyword or employer name.

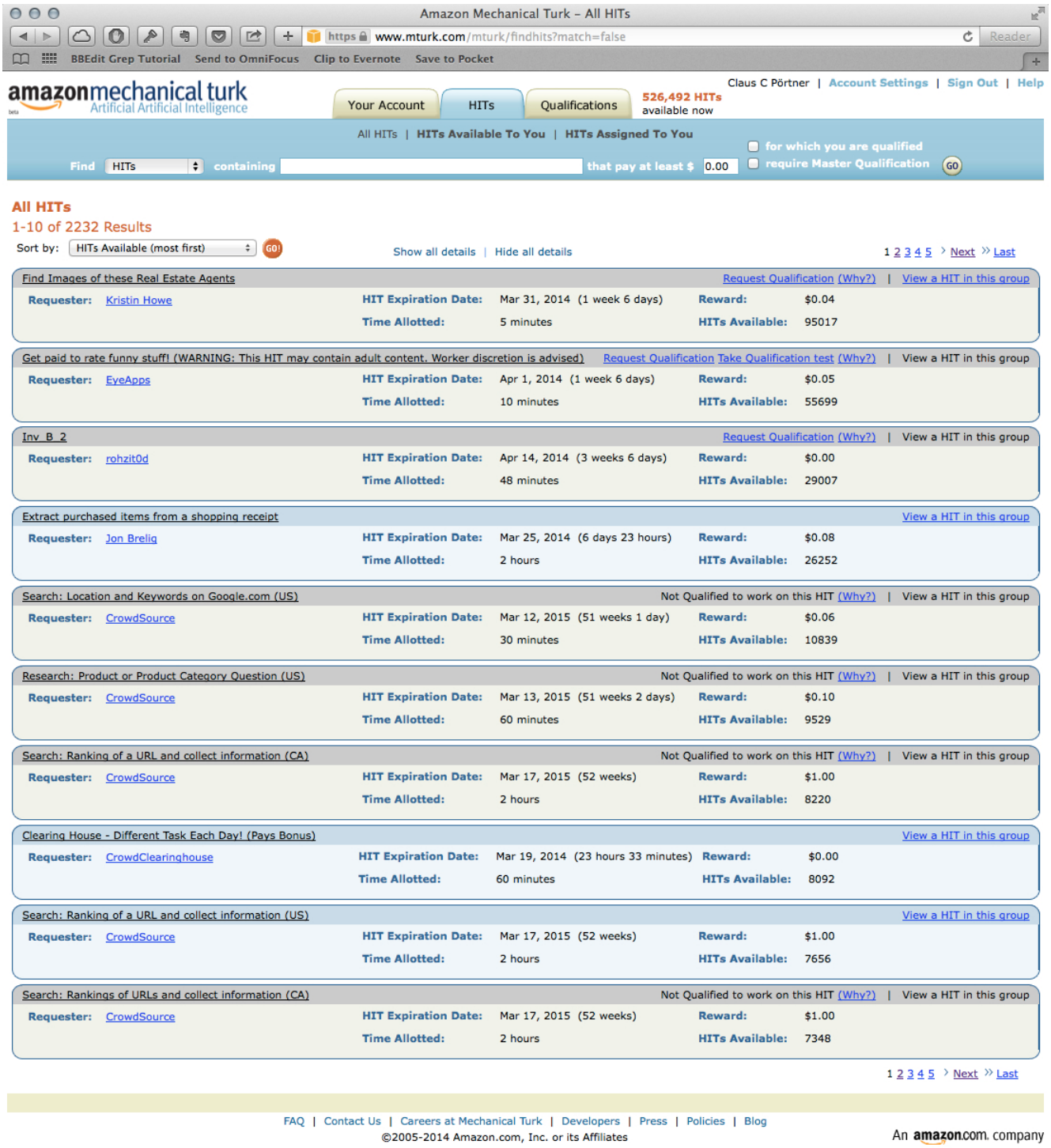
Work is paid per task, and although the corresponding hourly wage may not be typical of the overall US labor market, it will be close for workers on the current U.S. minimum wage.[6] [7] There are generally between 5,000 to 30,000 tasks completed each day [Ipeirotis, 2010]. Workers communicate on 3rd-party web forums, share tips, and discuss jobs and employers (see, for example, www.turkernation.com). Requesters can reject HITs for subpar work. Having HITs rejected has negative consequences for workers because requesters can exclude workers with high rejection rates [Horton, 2011].

[5] The tagline for Amazon's Mechanical Turk is "Artificial Artificial Intelligence" to emphasize that these are jobs that are done by people.

[6] The tax implications of working on Mechanical Turk are unclear, but Amazon does collect tax identification number from workers from both US and other countries.

[7] Appendix A contains complementary graphs of our dataset. Figure 13 shows that a substantial proportion of workers made around $5 per hour but the figure also shows that some workers were able to make as much as $25 per hour.

Figure 1: Listing of jobs on Mechanical Turk



The worker demographics has been studied by posting surveys to Mechanical Turk itself [Ipeirotis, 2008]. United States account for 46% of workers, with 34% in India, and 19% in other countries. Mechanical Turk workers are similar to the Internet population, although slightly more female, slightly younger, and more likely to be single and with smaller families. Many report having Master's or Ph.D. degrees, and the income distribution closely follows the distribution for the overall U.S. population.

Mechanical Turk is clearly not like "off-line" labor markets. There are no explicit contracts, no set working hours, no commuting, and clothing is entirely optional. Is it, however, similar to the market for freelance or independent contractor work, which rapidly is becoming more and more important in the US economy. A recent estimate is that there are 17.7 million independent workers, making close to $ 1.2 trillion in total income in 2013 and these numbers are been increasing over time [MBO Partners, 2013].[8] Most importantly, Mechanical Turk attracts people actively looking for work, rather than being a sample of undergraduate students participating in a lab experiment. These features make Mechanical Turk closer to a standard neoclassical labor market and well suited for experiments.

3.2 Image Tagging Job

The data for my paper comes from an experiment conducted to demonstrate the existence of compensating wage differentials (Pörtner, Hassairi, and Toomim [2015]). The image tagging job was chosen because it had advantages for the research questions posed therein but also because it is relatively familiar to workers on Mechanical Turk and simple to explain.[9] Within the job four job characteristics and the pay offered are randomized. The experiment uses a full factorial design [Fisher, 1935]. Experimental conditions are created by systematically varying the levels of each job characteristics

[8] There is, however, substantial uncertainty about these numbers since the Bureau of Labor Statistics does not directly count these types of employment.

[9] A subset of other possible jobs that were considered are: reading and categorizing text, searching keywords on Google, answering simple questions about images, such as whether a computer was present, scoring articles, providing summaries of articles, and creating chapter/time stamps for different videos. Most were rejected because they did not allow for implementation of varying job characteristics without substantially changing the length of time required to finish the task.

and pay, so all possible combinations are covered. The main benefit of this approach is efficiency; fewer workers are required to achieve the same level of statistical power as other approaches (see, for example, Wu and Hamada 2011 and Collins, Dziak, Kugler, and Trail 2014). With a factorial design one can estimate main effects of the various job characteristics without having to run individual experiments for each job characteristics, by "recycling" observations.[10]

Once a worker clicks on our job in the list of available jobs data collection begins. To ensure that job characteristics are not systematically related to the time of day, we listed all the possible combinations in random order. Each arriving worker is automatically assigned the next combination in this list. We observe whether the worker accepts the job and, if so, how many HITs are performed.

We act as a regular employer on Mechanical Turk. Worker is not informed that the offered jobs are part of an experiment and is always presented with the same set of circumstances based on their unique worker ID number assigned by Mechanical Turk. We did not inform workers that they were part of an experiment to rule out an observer effect, where workers change behavior in response to being part of an experiment. Workers do know that their output is monitored, but this monitoring is identical across experiments and conditions and akin to what one would find in any job. The experiments were conducted exclusively through computers ruling out any experimenter bias.

Requestors can only contact workers they have paid in the past. We therefore paid all new workers a \$0.25 "bonus" as shown in Figure 2. We do this only the first time a worker looks at one of our jobs; otherwise the worker is taken straight to the regular job. The bonus allows us to register workers who do not submit the actual work. The bonus may make workers feel an obligation to work, which would inflate the number who do at least one HIT and the number of HITs performed. This is not a concern here since the new worker bonus does not vary systematically across the different conditions and I am only interested in the differences between conditions.

[10] It is also, in principle, possible to estimate interaction effects between different job characteristics, although my experiments were not powered to do that. I have little in the way of theoretical prediction to suggest what characteristics these interactions should have and even relatively larger interaction effects between job characteristics would require sample sizes that I considered unlikely to achieve.

Figure 2: Letters to Prisoners Experiment—New Worker "Bonus"

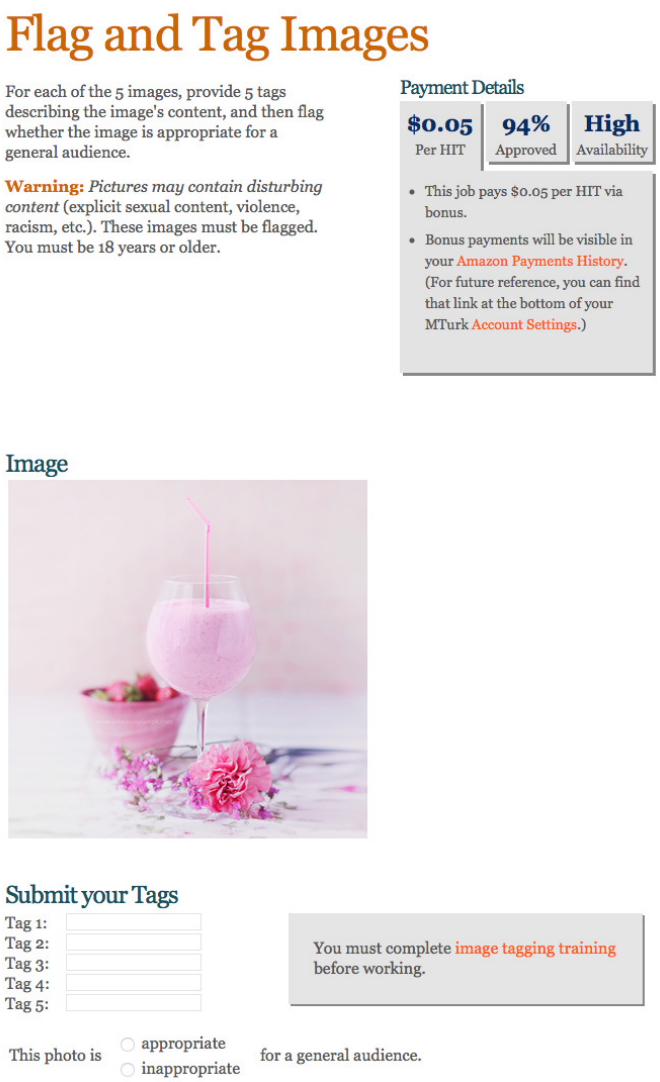
Mechanical Turk allows requestors to require skills and "certifications" of workers. Our only requirement is that the computer accessing the HITs must be in the US. This allows us to estimate consistent wage responses while achieving a sufficient sample size. It is possible to circumvent the location restriction through the use of proxy servers, but Amazon requires that workers provide a US tax ID number if they use a computer that appears to be in the US, which significantly limits the usefulness of using a proxy server to access Mechanical Turk.

The image tagging job is similar to other tagging jobs on Mechanical Turk, where requestors have workers go through images before deciding which ones to license. Once a worker clicks on the job, our software selects and displays five pictures. For each image we ask the worker to provide five tags or keywords, *in addition to clicking a radio button indicating whether each of the image is appropriate for a general audience*[11]. Figure 3 shows part of the page presented once a worker accepts the HIT, including one image.

We change the job's agreeableness by varying the number of disagreeable images. These attributes were central to Pörtner et al. [2015] in its own right, while for this paper their usefulness is in having workers rate the images based on their appropriateness for sensitive audiences. There are six levels in the experiment, corresponding to 0, 1, 2, 3, 4, or 5 disagreeable pictures per HIT. In our data disagreeableness is

[11]These radio buttons are central to the empirical analysis in this paper and their use for the purpose of measuring quality of worker's output is detailed in Section 3.4.

Figure 3: Image Tagging Experiment Page View

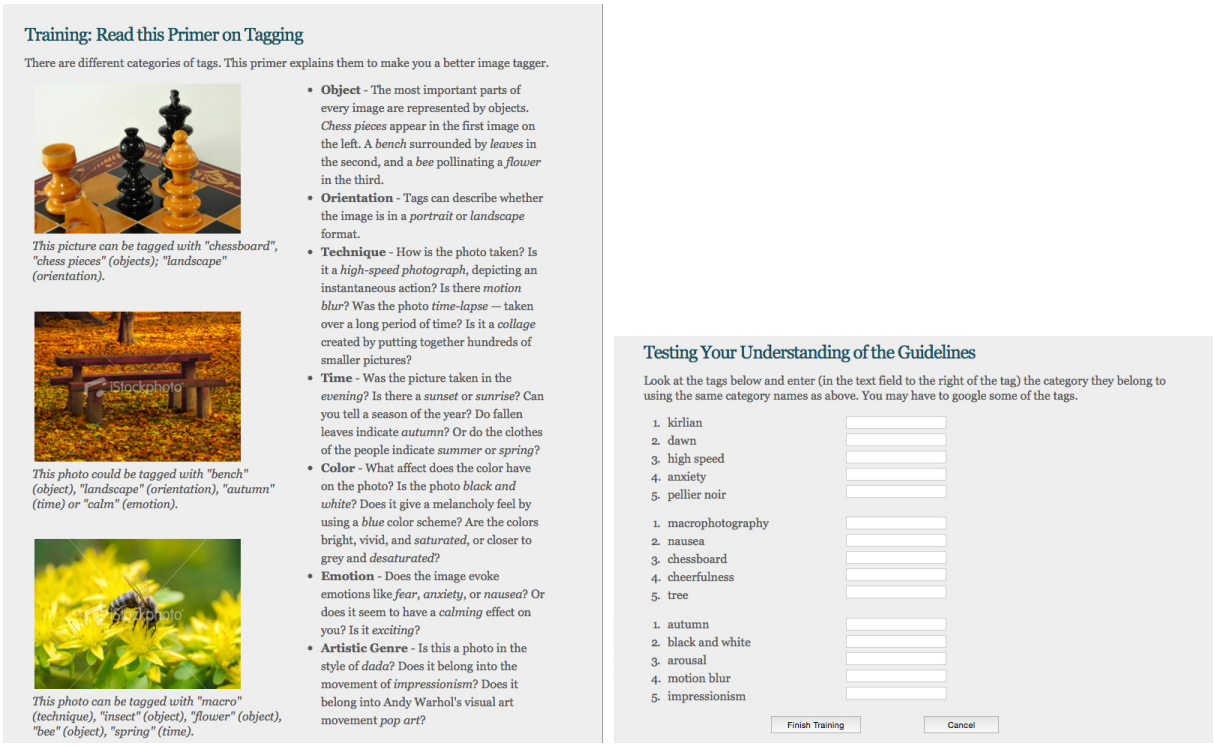


expressed as a ratio between 0 and 1. The number of disagreeable pictures do not change between HITs, but the ordering is randomly allocated, so that a worker with, say, one disagreeable image per HIT (20%) may see that as, for example, the first image on one page and as the third on the next. The agreeable images cover a wide variety of topics such as garden pictures, nature, travel photo, food, and animals. We have a collection of 5921 of these pictures. The disagreeable images were identified using Google Image search terms and then we deleted false positives.[12] This process

[12] The Google Image search terms included topics such as amputations, autopsy, broken limbs, gangrene, and larvas to name a few. All pictures are publicly available online.

is, of course, open to cultural biases in what is considered disagreeable, but certain responses are more likely biological responses and we aim at those. The conclusions in Pörtner et al. [2015] show that workers were willing to pay substantially to avoid working on those images. The stock of disagreeable images consists of 1131 pictures. Not all of these images are equally disagreeable and we did not attempt to rank them in any way. This does introduce some amount of measurement error in that workers with the same observed level of disagreeableness may see slightly different actual levels of disagreeableness. This variation should, however, be completely random and therefore only make the estimated standard errors larger.

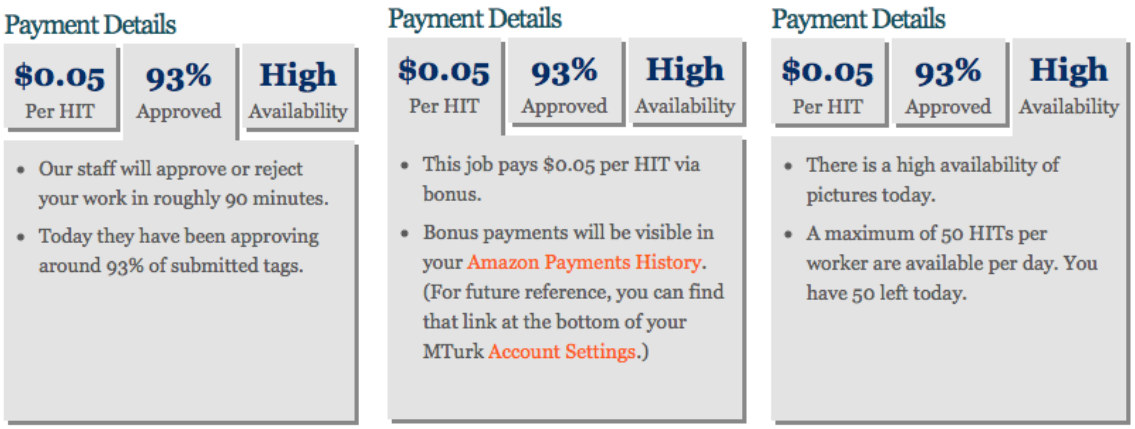
Figure 4: Image Tagging Experiment—Training and Test



Cost of learning is another job attribute that featured prominently in Pörtner et al. [2015]. In this paper, we will control for the possible difference in behavior due to this job attribute, but it will not be of importance to the main investigation. Cost of learning is difficult to capture in a setting where the tasks themselves are relatively short and simple. We need to vary the cost of learning without making the job itself easier or harder or otherwise fundamentally changing the job. We solved this by including

a "training component" with or without a "test." Everybody was asked to read a description of different categories of tags and examples of each. Those selected for the "training" condition got 15 questions to answer, where they were asked to categorize a set of tags based on what they had just read. Workers could not go on until they had answered all correctly. Workers not selected for "training" were asked to click a button indicating that they had read and understood the content. Figure 4 shows the guidelines and the test questions.

Figure 5: Image Tagging Experiment—Approval rate, pay, and availability



The probability of success is captured by our "approval" rate for tags[13]. Figure 5 shows an example. Because the experiment was designed to run over multiple days the actual number was drawn from a uniform distribution with the mean approval rate equal to either a low, 56%, or a high, 93%, approval rate depending on which was randomly assigned to the worker. This was to ensure that the worker did not see exactly the same number over multiple days when the expectation would be that there would be some variation over time. We paid everybody for all work irrespectively of the assigned approval rate. Furthermore, we never rejected HITs. This is probably responsible for our low estimates of monitoring's impact on quality of work submitted by the workers. Many workers worry that rejecting HITs may hurt their access to future jobs, because some requestors restrict access to job by requiring a certain acceptance rate.

[13]In Pörtner et al. [2015] this was just another job attribute, however, in this paper it takes on a special significance as a measure of monitoring, a crucial aspect of the efficiency wage theory.

The final part of the experiment is the pay offered. Workers were randomly assigned to a pay per five images tagged, equal to 25 tags, of between \$0.05 and \$0.50 in \$0.05 increments. Figure 5 shows an example of pay and availability. All workers could work up to 50 HITs per day. This limit was implemented to ensure that we did not run out of money.

The experiment ran over six days in 24 hour segments starting at 07.58 GMT. A worker would see one set of conditions during each 24 hour period and then after 07.58 GMT the job conditions and pay would be randomized anew. The randomization did not take into account previous job characteristics or pay. We choose 07.58 GMT because that was the time of the day where there were the fewest number of workers on Mechanical Turk. This set-up allows us to determine the minimum wage the workers are willing to work fors, as well as to see what is the incentive effect of increase the wage above this minimum.

3.3 Data Structure

Our data is collected at the HIT level. A HIT is very short, not more than a minute's worth of work (see Section 3.1 for details on the HIT concept and Figure 6 for a histogram of time spent on HIT). For every HIT I have information about worker ID, quality of work, wage offered, job attributes and time spent on the HIT. I can also infer additional information by aggregation. For example, for a given HIT I have information on how many HITs given worker has already submitted before they started working on that HIT. Similarly, I have information on total HITs submitted over the course of the experiment and how many HITs will the worker submit before leaving our experiment altogether. While quality varies on a HIT by HIT basis our treatment variables vary on a day by day basis. This raises a question how should I aggregate the data. I have opted for a worker-HIT panel since one can use the following information collected at the HIT level:

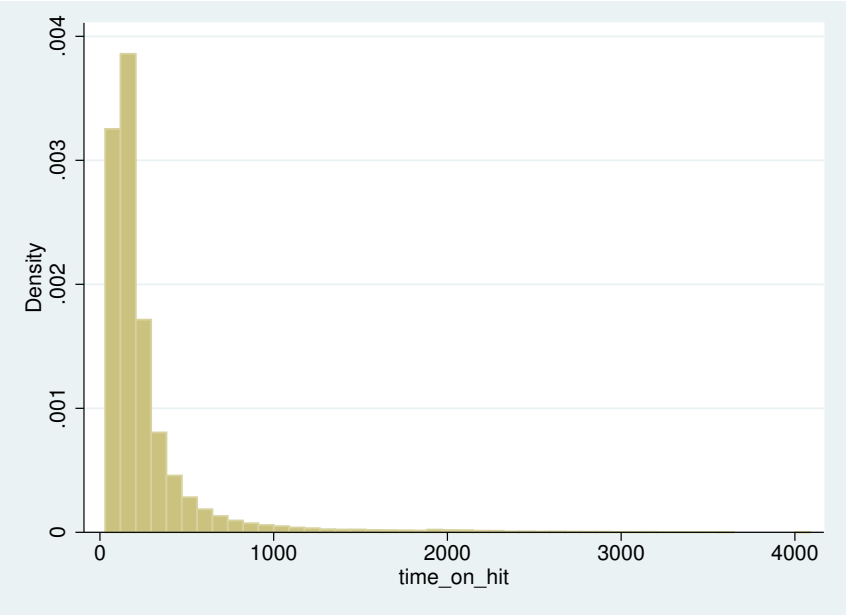
1. HIT count as a measure of experience (or job fatigue),
2. HITs left as a measure of how much longer a worker desires to work on my HITs (assuming rational expectations that are on average correct) – this will reveal whether incentive effects vary with tenure,
3. HIT quality,
4. time spent on HIT.

Since I wanted to be able to control for learning and fatigue with the job attributes which does vary on a HIT by HIT basis and conceivably could impact provided effort, I have decided to model worker's behavior on a HIT basis as a decision on how much effort a worker provides on a single HIT. I have performed analysi on worker-day level as well, however, the results are not very illuminating and I suspect the results are obscured by un-captured HIT-by-HIT dynamics. Section 5 provides more details on the respective results obtained from the worker-HIT and worker-day models.

3.4 Output/Productivity Data

As described in Subsection 3.2 the job I offered to Mechanical Turk workers involved tagging images. A worker is presented with an image and is asked to:

Figure 6: Histogram of Time on HIT Data for submitted HITs only



1. flag whether the image is not suitable for children and sensitive audiences.
2. provide five tags that best describe the image,

My quality measure is based on the former, the indicator of suitability for sensitive audiences. To prevent disagreeable images from being seen by an audience that could be hurt by seeing them I have asked workers to flag images as appropriate or inappropriate (See Figure 3 for a screenshot). Since the original experiment was designed to price job attributes the images were in fact selected into a treatment and control group in which the treatment contained "inappropriate" images while the control group contained neutral ones. Consequently, I had the "true" assessment of inappropriateness of these images from the treatment assignment. My measure of productivity then consists of a count of how many times workers agreed with our judgement of inappropriateness of these images. I refer to this measure in my regression analysis variously as "correct appropriatness ratings", "correct_ratings" or simply "ratings".

3.5 Effort vs Output

The shirking model makes predictions about effort, while the sorting model makes predictions about output (since effort is abstracted from in that model). Since I am

trying to test both hypotheses within a single theoretical and regression model it would follow that I need both a measure of effort, and a measure of output. How does effort differ from output? Is there a one-to-one relationship between effort and output?

Some authors question this claiming that more effort does not always result in more output (Ariely, Gneezy, Loewenstein, and Mazar [2009]). At the same time, these same authors limit their findings to high stakes. As stakes get higher, their findings imply, the cognitive system is impaired and workers' output no longer responds to the higher effort exerted. I will ignore this consideration, as on Mechanical Turk the stakes could not be smaller (our priciest HIT was offered at $.50).

Effort and output will be taken as one and the same in my model and data. Ability will not contribute directly to the marginal product but rather will act through its effect on the cost of effort and through its impact on beliefs about workers' probability of success, and the extent to which they need to exert effort to be successful.

It has also been suggested that sometimes not paying at all is better than paying low wages (Gneezy and Rustichini [2000]). This would make more sense in the Mechanical Turk environment, however, on close examination it is also unlikely to be affecting my results (we have not tried to pay $0 wages) – the research shows that this is the case in context where an activity otherwise considered an honor activity is transformed into "mere" work and thus stripped of its social prestige. Since the experiment is conducted within an established labor market and we offered jobs under similar conditions as other requesters, this is unlikely to affect my results.

3.6 Measure of Acceptance Wages

Weiss [1980] builds his model around the idea that prices (wages) have two roles; one is a screening role, and the other is the usual allocating role in which prices are equal to marginal products. In his model one price plays these two roles and the contradictory demands of these two roles distort markets. In my experiment, therefore, I have attempted to decouple acceptance wage from actual wages so I can study these roles separately.

As described in Section 3, wages varied every day (See Figure 5) and on a given day workers could do up to 50 HITs. Workers did not have to work every day; the choice

on which day(s) to work was left to workers. I have looked at all wages that given worker worked for over the course of the experiment. I have taken the minimum of these accepted wages and used it as 'minimum wage accepted', a proxy for opportunity cost/reservation wage/acceptance wage. This minimum accepted wage will play the screening role described in Weiss [1980] while the actual wage paid for a given HIT would play the incentive role of wages implied by the shirking model.

Since our wages are experimentally determined (randomized) our wages do not really communicate much in a way of information about relative scarcity of resources as wages normally are supposed to do according to the neoclassical competitive model.

4 Identification Strategy

4.1 Disciplining Device: Dismissal vs Non-Payment

In our experiment we do not fire workers; the disciplinary tool of choice is not paying them for a particular job. In our experiment we made no reference to workers being fired, however, it is understood on the platform that workers who engage in blatant disregard of the job instructions in a systematic way are excluded by employers. As a result, workers may be worried about being fired despite the fact that it is not the policy in the job being offered for this experiment. In summary, to use the language of Shapiro and Stiglitz [1984], my disciplining device was non-payment (for more details see Section 3.2), rather than dismissal.

4.2 The Framework

The Mechanical Turk worker is assumed to maximize a utility function that incorporates the probability of HIT rejection (due to poor quality), the utility from income/consumption and disutility of effort. Our framework nests the heterogeneous model (Weiss [1980]) by including the minimum accepted wage/reservation wage (proxy for ability) in the probability of rejection function and cost of effort function, as well as the shirking model (Shapiro and Stiglitz [1984]), by including effort as both a direct disutility and a positive contributor to the probability of being paid. Experience is also included in the model as a control and to indicate possible findings regarding the labor turnover model (Salop [1979]). Weiss [1980] also does not really explain why should acceptance wage be an increasing function of labor endowment. While this is a starting assumption in his model I will construct a model of rational maximizing agents which will allow us to see whether this assumption is actually a plausible one.

4.3 Model Specifications

1. The probability of rejection function in the model is chosen to be a simple exponential function which is not very standard. I performed some simulations that suggest that the functions behave fairly similar to each other except that the exponential is more prone to exhibit an interior solution than the logistic for some

parametrizations (based on my data).

2. Risk-neutrality is assumed for the workers (as was the case in Shapiro and Stiglitz [1984]).
3. No constraint is assumed for the optimization problem. No direct constraint for effort. No functional constraint for effort (cognitive capacity or effort being constraint by ability).
4. Static model is assumed for the behavior on the HIT. Our empirical results may lead me to the conclusion that a dynamic model would provide further insight on my question of interest.
5. I will take the quality measure of workers' output and use it as a measure of effort (Section 3.5 offers justification for doing so).
6. Ability enters the cost of effort function to allow for the possibility that more able workers can deliver the same level of effort with lower cost/disutility.
7. I assume that reservation wage is a static concept (see Section 5.7 for more detailed discussion of this assumption).
8. Job attribute can affect agent's optimization problem through cost of effort. I am assuming that it is less costly to provide effort on a job with positive job attributes, ceteris paribus.

4.4 Utility Function

The utility function has the following general form:

$$U(w, e, \bar{w}, J) = \big(1 - b(e, p, n, \bar{w})\big)g(w) - h(e, \bar{w}, J) \tag{9}$$

where $b(e, p, n, \bar{w})$ is the probability of not receiving payment (work being judged as subpar) and $g()$ and $h()$ specifying how exactly wages and effort affect utility. e stands for the effort level, p stands for the advertised probability of success (quality standards), n is a number of HITs done by the work up until now, w is the current wage, J is a job attribute and $\bar{w}$ is worker's reservation wage or opportunity cost. In particular, I choose the following specification[14]:

[14]For reasons of tractability and comparability with Esteves-Sorenson et al. [2016].

$$\max_{e} E\left[1 - \exp(\alpha + \gamma p + \theta \bar{w} + \mu n + \omega e)\right] w - \delta \exp(\psi e + \zeta \bar{w} + \phi J) \tag{10}$$

4.5 Structural Parameter Assumptions

The following behavioral assumptions are consistent with the spirit of assumptions in Shapiro and Stiglitz [1984] and Weiss [1980], however, this claim is somewhat speculative since my model is more structural and comprehensive than either of them.

$$\theta < 0 \tag{11}$$

$$\omega < 0 \tag{12}$$

$$\psi > 0 \tag{13}$$

$$\zeta < 0 \tag{14}$$

$$\phi > 0 \tag{15}$$

These assumptions say that

1. higher ability leads to higher belief that worker's own work will be accepted,
2. more effort leads to lower chance of negative assessment (rejection) of worker's output,
3. more effort leads to more disutility,
4. ability decreases cost of effort, and
5. job disamenity increases the cost of effort

4.6 Equilibrium Effort Provision by Workers

Maximizing the above utility and solving for optimal effort yields:

$$e^* = \frac{\log\left(-\frac{\omega}{\delta\psi}\right) + \alpha}{\psi - \omega} + \frac{1}{\psi - \omega}\log(w) + \frac{\gamma}{\psi - \omega}p + \frac{\theta - \zeta}{\psi - \omega}\bar{w} + \frac{\mu}{\psi - \omega}n + \frac{-\phi}{\psi - \omega}J \tag{16}$$

4.7 Implication of the Shirking Model

The testable implication of the shirking model is that holding monitoring level constant higher wage leads workers to exert more effort:

$$\frac{1}{\psi - \omega} > 0 \tag{17}$$

This would be guaranteed by the above direct assumptions on structural parameters since $\psi > 0$ and $\omega < 0$. However, I will only be able to identify this fraction rather than the actual structural parameters ψ and ω individually.

4.8 Effort Substitute for Ability

If I were to run the reduced form Equation 16 with my data and found that $\frac{\theta-\zeta}{\psi-\omega} < 0$ this would imply that either

1. $\theta < \zeta$ and $\psi > \omega$; or
2. $\theta > \zeta$ and $\psi < \omega$

The latter option is inconsistent with the Structural Parameter Assumptions 11 since ψ is supposed to be positive while ω is supposed to be negative so it would be hard for the former to be smaller than the latter. So starting from the Structural Parameter Assumptions I am lead to believe that $\theta < \zeta$ and $\psi > \omega$. Since θ and ζ are both negative this really means $|\theta| > |\zeta|$, in other words θ has a stronger effect than ζ which means that higher ability is not as important in lowering agent's cost of effort as it is in increasing the agent's belief (since everyone actually got paid) about her chance of having her work accepted and paid for. In combination with the fact that there is a substitution between effort and ability this would raise the question whether there is a potential role for overconfidence. If agent overestimates her ability given that the ability does not seem to lower cost of effort too much she might decide to substitute ability (perceived) for effort and lower her effort while thinking that she is maintaining her chance of having her work accepted at an equal level.

4.9 Effort Complement for Ability

On the other hand, if it turns out that $\frac{\theta-\zeta}{\psi-\omega} > 0$ I would reach the opposite conclusion, i. e. ζ is the stronger force than θ. This would mean that ability primarily works through lowering the cost of effort and the agent's beliefs about her own ability are not significantly distorting her effort decision.

5 Results

In Section 4 I derived an expression for an optimal amount of effort as a function of the model variables and parameters. This section will provide an estimation of this equilibrium effort function using the experimental data. Equation 16 becomes:

$$corectRatings^* = \frac{\log\left(-\frac{\omega}{\delta\psi}\right) + \alpha}{\psi - \omega} + \tag{18}$$

$$\frac{1}{\psi - \omega}\log(wage) + \tag{19}$$

$$\frac{\gamma}{\psi - \omega}SuccessRate + \tag{20}$$

$$\frac{\theta - \zeta}{\psi - \omega}log(MinAcceptedWage) + \tag{21}$$

$$\frac{\mu}{\psi - \omega}HITsDone + \tag{22}$$

$$\frac{-\phi}{\psi - \omega}Disagreeable + \epsilon \tag{23}$$

This translates into the following reduced form version:

$$correctRatings^* = \beta_0 + \tag{24}$$

$$\beta_1 \log(wage) + \tag{25}$$

$$\beta_2 SuccessRate + \tag{26}$$

$$\beta_3 log(MinAcceptedWage) + \tag{27}$$

$$\beta_4 HITsDone + \tag{28}$$

$$\beta_5 Disagreeable + \epsilon \tag{29}$$

5.1 Censored Regression

There were 5 images in every HIT and worker was asked to use 5 radio buttons to indicate whether the image is appropriate for a sensitive audience. My measure – correct ratings – enumerates how many times worker's judgement aligned with the the researchers' (our) judgement. Given the way this variable is constructed one may worry about upper or lower censoring or both. Histograms in Figure 7, Figure 8, Figure 9 and Figure 10 show that upper censoring at the value of 5 (5 agreements between workers' and our judgements of image inappropriateness). I have performed both regular and censored regressions in Appendix B and it turns out that the results differ greatly which necessitates using the censored model in my main analysis.

Figure 7: Histogram of Correct Appropriateness Ratings for submitted HITs

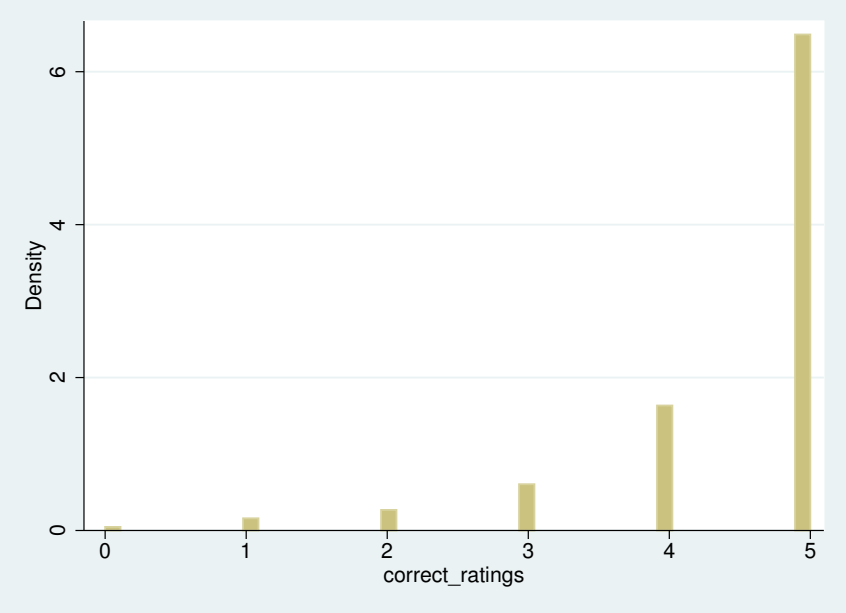
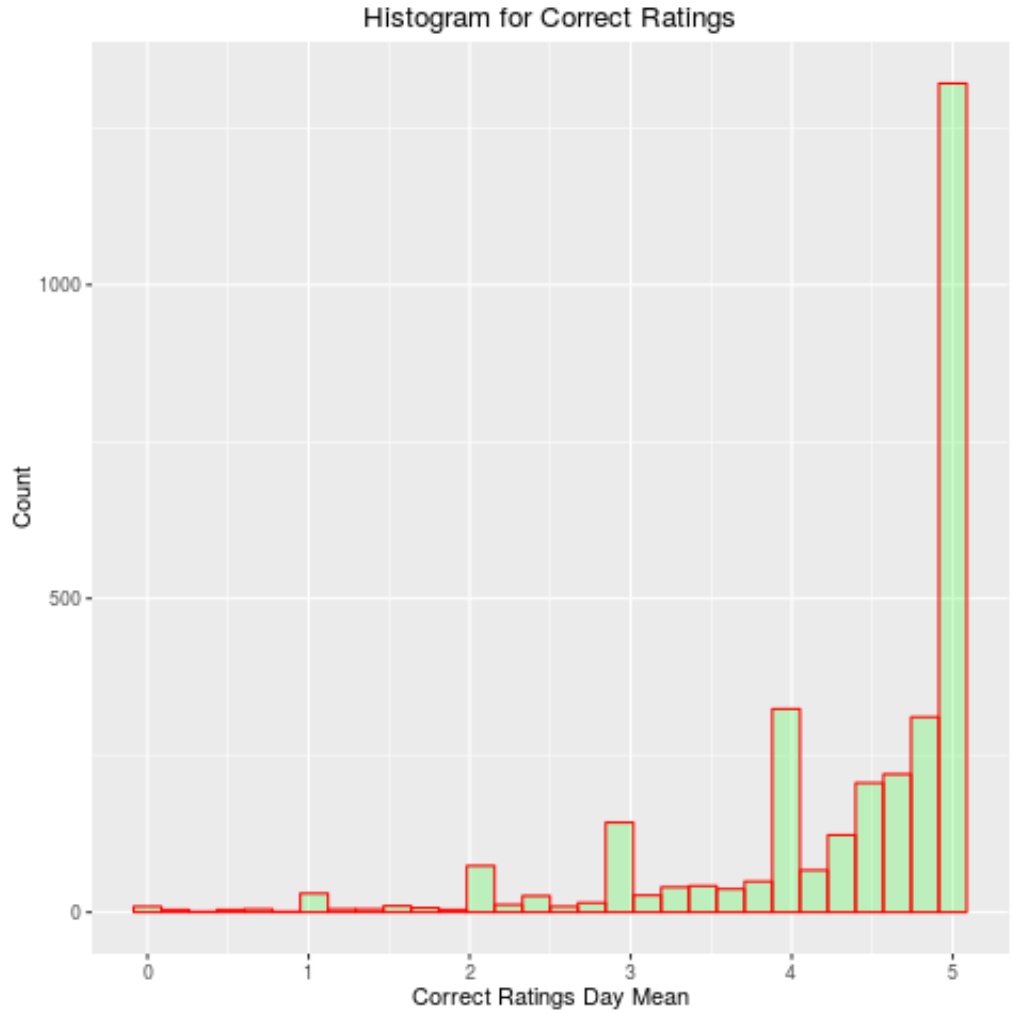


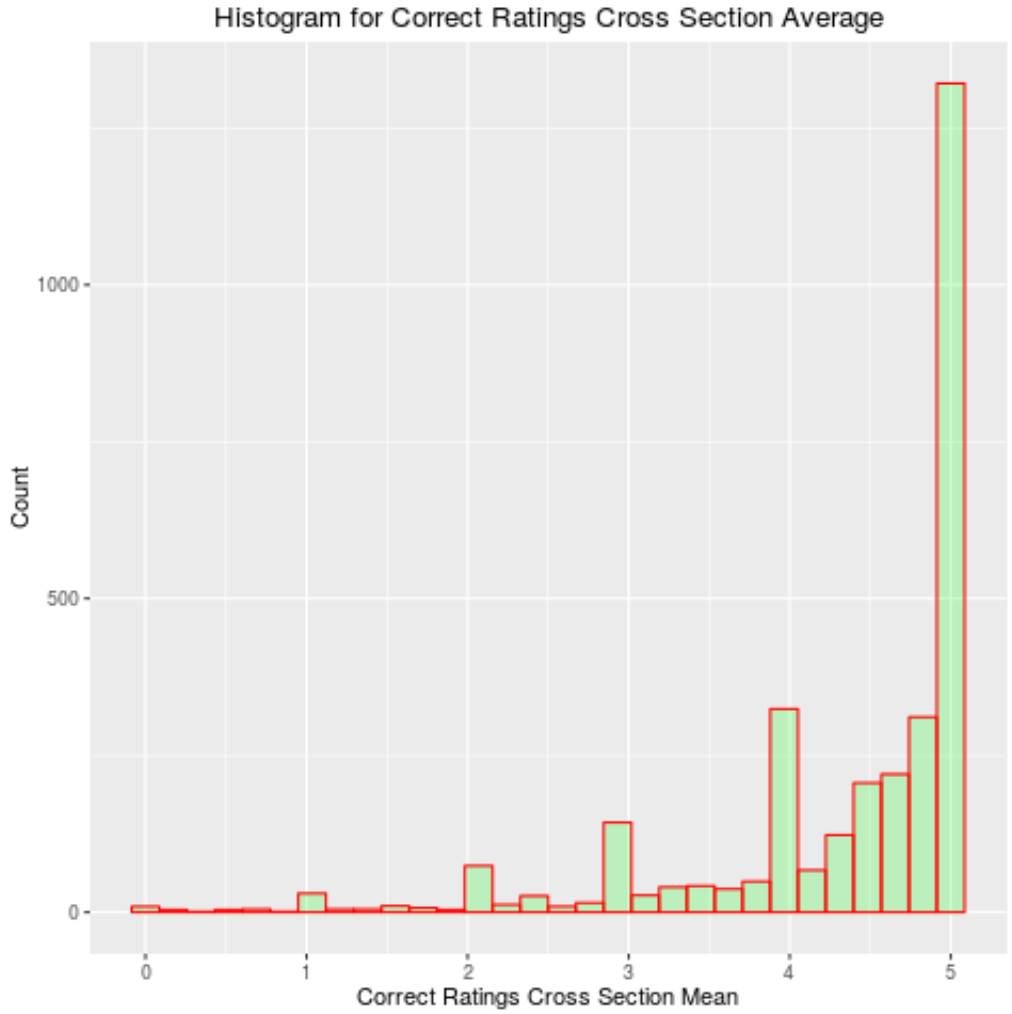
Figure 8: Histogram of Correct Appropriateness Ratings Averaged over Days



5.2 Methodology – Data Granularity

Table 1 and Table 3 provide results from a several econometric specifications. Following the discussion in Section 3.3 I have run several econometric specifications with different data granularity and emphasizing different sources of variation. My preferred specification is in Column (3) of Table 1. This is a random effects model using HIT level granularity of the data. One might argue that this granularity overrepresents workers who work more, however, to me this actually makes sense since these workers have contributed more work and are responsible for a bigger part of the total product. Furthermore, on this level of granularity I am able to control for dynamic effects such as learning and dynamics of incentive effects such as the incentive effect of wages de-

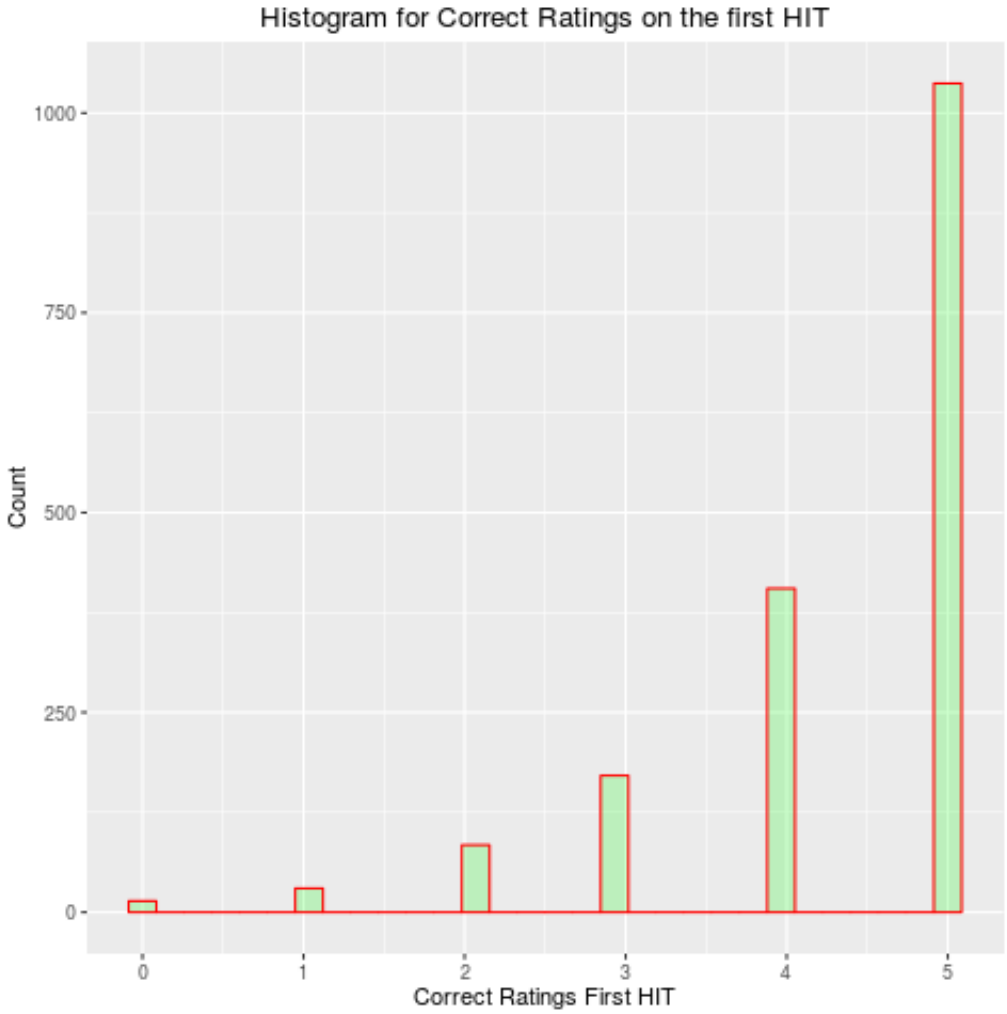
Figure 9: Histogram of Correct Appropriateness Ratings Averaged the Whole Experiment



creasing toward the end of worker's tenure. Table 1 contains results from a random effects model and from a pooled OLS model. Random effects model would control for a worker-level effects if present, however, the two models provide fairly comparable results.

For reference, I have also run some additional regressions in Table 3 but they exhibit some inexplicable results that I believe are due to dynamic effects which I am unable to control for at this level of granularity. This specification could also overrepresent the impact of workers who do not submit a lot of work but have perverse response to incentives (wages).

Figure 10: Histogram of Correct Appropriateness Ratings on the first HIT



5.3 Testable Implication of the Shirking Model: Nexus Between Productivity/Effort and Wages

Now I connect my theory with the experimental data. This subsection discusses the shirking model and the next will discuss the sorting model.

Is there a positive causal relationship between paid wages and productivity on the labor supply side, as implied by the shirking model? Yes. Column (3) shows results from my preferred specification, a worker-HIT panel data using a random effects estimator. The estimate of $log(wage)$ on effort (see Section 3.4 for details on this measure of effort) in this model is $\beta_1 = .037$ and highly statistically significant (the

tables always indicate statistical significant at the .1, .05 and .01 levels). Column (2) in Table 1 shows equivalent results.

In my preferred specification in Column (3) of Table 1 the estimate for the $log(wage)$ coefficient, $\beta_1 = \frac{1}{\psi-\omega}$ in my theoretical model, is equal to .037. This result shows that $\frac{1}{\psi-\omega} > 0$ which confirms (does not refute) the implication of the shirking model (Shapiro and Stiglitz [1984], Bulow and Summers [1986]) as expressed in the Testable Implication 17 from Section 4. Using a pure moral hazard model I would be able to quantify the relevant parameter but using the model nesting both shirking and sorting models allows me to provide only qualitative insights on these parameters. These qualitative insights are enough to confirm (or not refute when the data provided an opportunity for refutation) the conclusion of the shirking model. What I cannot say is whether this result comes from effort having low cost or high impact on the perceived probability of being paid for a HIT.

5.4 Testing the Heterogeneous Agent Model

The coefficient on the regressor `log(min_accepted_wage)` in Table 1 provides the result of the test of the heterogeneous agent model. This coefficient is fairly comparable accross the four specifications in the table. In my preferred specification the coefficient for `min_accepted_wage`, $\beta_3 = \frac{\theta-\zeta}{\psi-\omega}$, turns out to be .155 and highly significant. This is much stronger than the incentive effect of wages (coefficient on `log(wage)`) suggesting that heterogeneity is bigger issue than incentives in our data and on Mechanical Turk in general. This makes sense given that there is lack of screening procedures on Mechanical Turk and given that the minimum accepted wage turned out to be on average very close to the average wage offered in the experiment ($.248 mean minimum accepted wage against the mean $.29 of the wage offered to workers in the experiment). In the end, our wage range was quite reasonable given the wages offered on Mechanical Turk and we did not wildly overpay or underpay the workers.

As detailed in Section 4 the sign of the expression $\frac{\theta-\zeta}{\psi-\omega}$ is of significance when drawing conclusion about incentives on one hand, and selection on the other. This reduced form parameter meshes influences of four structural parameters – providing comparative insight into how ability, beliefs about ability, and effort affect each other

and equilibrium behavior in the process.

I noted in Section 4 that $\frac{\theta-\zeta}{\psi-\omega} > 0$ would imply that effort and ability are treated as complements by the workers and that θ has a smaller effect than ζ, while $\frac{\theta-\zeta}{\psi-\omega} < 0$ would imply that effort and ability are treated as substitutes and that θ has larger effect than ζ. In the end, it turns out that neither the former is the case and that $\frac{\theta-\zeta}{\psi-\omega} > 0$.

θ is worker's belief about the contribution of their ability to their chances of having their work accepted and paid for. ζ is revealed preference estimate of the contribution of their ability towards lowering their disutility of effort. What can we learn from the magnitude of these parameters? Since θ has a weaker impact than ζ it shows that workers act mainly on the effect of their ability on the experienced cost of effort rather than on their hypothesized belief in their own ability.

Furthermore, $\frac{\theta-\zeta}{\psi-\omega} < 1$ suggests that the magnitude of ζ is limited from above by $\psi - \omega$.

Table 1: Censored Cross-Section and Worker-HIT Random Effects Models

	Dependent variable:			
	correct_ratings1			
	Pooling	Pooling	Worker-HIT RE	Worker-HIT RE
	(1)	**(2)**	**(3)**	**(4)**
log(wage)	−0.056	0.089***	0.037***	−0.022
	(0.036)	(0.026)	(0.012)	(0.017)
HITsLeft	0.004***			−0.0002
	(0.001)			(0.0003)
HITsDone	0.003***	0.003***		
	(0.0003)	(0.0003)		
log(min_accepted_wage)	0.210***	0.189***	0.155***	0.079***
	(0.020)	(0.020)	(0.011)	(0.011)
disagreeable	−0.039***	−0.039***	−0.035***	−0.035***
	(0.0005)	(0.0005)	(0.0002)	(0.0002)
success_rate	−0.003***	−0.003***	0.0003	0.0003
	(0.001)	(0.001)	(0.0003)	(0.0003)
as.factor(day)2	0.005	−0.043	0.144***	0.101**
	(0.079)	(0.079)	(0.040)	(0.042)
as.factor(day)3	0.225***	0.173**	0.295***	0.150***
	(0.080)	(0.080)	(0.041)	(0.041)
as.factor(day)4	0.331***	0.262***	0.375***	0.236***
	(0.079)	(0.078)	(0.039)	(0.041)
as.factor(day)5	0.373***	0.289***	0.402***	0.194***
	(0.079)	(0.077)	(0.039)	(0.042)
as.factor(day)6	0.297***	0.194**	0.520***	0.252***
	(0.082)	(0.077)	(0.039)	(0.044)
log(wage):HITsLeft	0.003***			0.002***
	(0.001)			(0.0002)
Constant	7.827***	8.076***	7.000***	7.094***
	(0.109)	(0.102)	(0.049)	(0.054)
Observations	41,963	41,963	41,963	41,963

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Worker-HIT panel summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
time_period	41,967	59.614	55.659	1	341
day	41,967	4.317	1.504	1	6
disagreeable	41,967	42.017	33.179	0	100
training	41,967	0.445	0.497	0	1
wage	41,967	0.332	0.127	0.050	0.500
success_rate	41,967	78.199	17.671	49	95
time_on_hit	41,967	253.085	303.457	30	3,617
HITsDone	41,967	53.920	50.858	1	293
correct_ratings_raw	41,963	4.490	0.942	0	5
hourly_wage	41,967	7.902	5.591	0.056	48.649
min_hourly_wage	41,967	7.902	5.591	0.056	48.649
totalHITs	41,967	105.873	69.972	1	293
HITsLeft	41,967	51.953	50.516	0	292
min_accepted_wage	41,967	0.201	0.116	0.050	0.500

5.5 Dynamics of Wage Incentives

Column (4) and Column (1) in Table 1 break down the incentive effect of wages by tenure. $log(wage)$ now turns statistically insignificant (for `HITsLeft` = 0) and the interaction term `log(wage):HITsLeft` is now .002 and highly statistically significant. This implies that on their last HIT workers are not incentivized by wages but the more HITs they have left the more they are bound by the incentive effet of wages. This make sense, since workers take into account loss of possible future stream of premia rather than the wage premium being offered in the current period. This effect is somewhat visible in Table 3 in Column (2) and Column (1). In the former column we see a

negative coefficient for `log(wage)` of $-.096$ and highly significant. Once `dayTotalHITs` and `dayHITsRemaining` are added to the regression specification, however, this effect loses its significance (while the newly added regressors turn out to be highly significant). This points to the presence of dynamic effects that are hard to control for on the worker-day level. Hence I prefer to base my conclusions on Column (3) and Column (2) of Table 1.

5.6 Heterogeneity in Response to Incentives

Results from Table 3 are a bit puzzling as far as the coefficient on `log(wage)` is concerned. Column (3) shows a negative effect of wages on performance. This could mean that there are some workers who respond negatively to higher wages and that they do not submit a lot of HITs (which is why they would be overshadowed in the worker-HIT and worker-day specifications). Conceivably, it could be the case that wage offered by employer is a signal of savviness of this employer and if the wage is too high then employer would be judged as being of low quality and unable to perform monitoring very well, potentially attracting workers attempting to scam this employer.

Table 3: Censored Models Aggregating the Data on the Day and Worker Level

	Dependent variable:			
	correct_ratings1			
	Worker-Day RE	Worker-Day RE	Worker-HIT BE	First HIT
	(1)	**(2)**	**(3)**	**(4)**
log(wage)	−0.055	−0.096**	−0.234***	−0.126
	(0.044)	(0.042)	(0.078)	(0.130)
log(min_accepted_wage)	0.098**	0.117***	0.234***	0.127
	(0.041)	(0.039)	(0.068)	(0.124)
disagreeable	−0.025***	−0.025***	−0.024***	−0.038***
	(0.001)	(0.001)	(0.001)	(0.002)
success_rate	−0.001	−0.001	−0.002	−0.002
	(0.001)	(0.001)	(0.001)	(0.003)
dayTotalHITs	−0.005***			
	(0.002)			
dayHITsRemaining	−0.002*			
	(0.001)			
training			−0.069	−0.135
			(0.054)	(0.099)
totalHITs			−0.001	
			(0.001)	
logSigma	0.062***	0.059***	−0.028*	0.518***
	(0.014)	(0.014)	(0.017)	(0.034)
Constant	6.131***	6.062***	5.958***	7.490***
	(0.114)	(0.113)	(0.143)	(0.276)
Observations	3,131	3,131	1,776	1,741

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Worker-Day Panel Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
day	3,135	3.924	1.583	1	6
wage	3,135	0.295	0.139	0.050	0.500
success_rate	3,135	77.264	17.759	49.000	95.000
min_hourly_wage	3,135	5.061	4.150	0.109	32.429
min_accepted_wage	3,135	0.216	0.130	0.050	0.500
correct_ratings1	3,131	4.360	0.930	0.000	5.000
disagreeable	3,135	46.494	33.933	0	100
time_on_hit	3,135	383.734	341.546	58.320	3,307.000
hourly_wage	3,135	5.061	4.150	0.109	32.429

Table 5: First HIT Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
time_period	1,741	1.431	1.080	1	13
day	1,741	3.354	1.595	1	6
disagreeable	1,741	47.099	33.654	0	100
training	1,741	0.396	0.489	0	1
wage	1,741	0.290	0.140	0.050	0.500
success_rate	1,741	77.836	17.594	49	95
time_on_hit	1,741	583.195	461.033	98	3,617
HITsDone	1,741	1.000	0.000	1	1
correct_ratings1	1,741	4.317	1.035	0	5
hourly_wage	1,741	2.737	2.291	0.056	15.652
min_hourly_wage	1,741	2.737	2.291	0.056	15.652
totalHITs	1,741	23.469	43.968	1	293
HITsLeft	1,741	22.469	43.968	0	292
min_accepted_wage	1,741	0.248	0.138	0.050	0.500

Table 6: Cross-Section Data Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
wage	1,780	0.299	0.126	0.050	0.500
success_rate	1,780	77.883	15.925	49.000	95.000
min_accepted_wage	1,780	0.248	0.138	0.050	0.500
correct_ratings1	1,776	4.359	0.888	0.000	5.000
disagreeable	1,780	45.499	30.102	0.000	100.000
time_on_hit	1,780	406.817	346.490	66.440	3,307.000

5.7 Wage per HIT, Hourly Wage, Reservation Wage and Learning

My choice of the acceptance wage measure is based on the assumption that this acceptance wage does not change over the course of our six day experiment. This would imply that workers are not adding significantly to their human capital or increase their value on the Mechanical Turk market by working on our HITs or over the course of a week. This also implies that workers know their value (and this value has converged to a stable equilibrium) on the job market and are not significantly updating their belief about their ability/opportunities. Given the short period of time, this would seem to be a reasonable assumption.

Furthermore, I have replicated the analysis in this paper with minimum accepted hourly wage by dividing wages by time spent on a given HIT and selecting the minimum of such a time series for all workers; the results did not materially change. The implied per hour wage would have incorporated learning. Workers who learn over time how to submit my HITs faster are effectively getting paid more, ceteris paribus. This means that these workers might be more willing to accept a lower per HIT wage as this is offset by their increased ability to finish HIT quickly keeping their de facto hourly wage constant. By looking at imputed hourly wages I made my analysis robust to this issue, incorporating the possible dynamics between a reservation wage and learning.

Given that we have manipulated not only wages but also job attributes I need to take account of that when I construct my measure of opportunity cost. This is another advantage of working with the implied hourly wages because I find that job disamenities increase the amount of time workers take to submit a HIT (see Appendix **??**) so my hourly wage measure already incorporates the possible impact of job attributes on accepted wages.

Appendix C explores how accepted wages and imputed accepted hourly wages are changing over the course of the experiment. It turns out that the implied hourly wages for accepted HITs are increasing every day within the week I ran my experiment.

5.8 Job Attributes and Screening on Wages

It has been suggested recently that in the context where job attributes play a major role higher wage helps attract workers with inferior match-specific utility rather than high productive workers (Delfgaauw and Dur [2007]). The implication of this claim is rather opposite to the claim advanced by Weiss [1980] and if empirically relevant would in fact render screening by wages an infeasible strategy to recruit high ability workers.

As discussed in Section 3.2 we have manipulated the job attribute of our HITs by asking workers to provide keywords (tags) for agreeable (positive job attribute) or disagreeable (negative job attribute) images[15]. Pörtner et al. [2015] clearly shows that this had the desired effect as the workers required a compensating wage differential to work on these HITs.

My paper incorporates job attributes in the joint sorting and shirking model so I can test this hypothesis as well. Table 1 shows the result in Column (3). The coefficient on the `disagreeable` regressor is $-.035$ and highly significant. For comparison, the coefficient on the `log(wage)` in the same column is .037. This means that the job attribute has an equivalent effect to wage paid and a firm could motivate workers through a job attribute as well as it could through wages paid[16]. The coefficient on `log(min_accepted_wage)`, however, shows that the reservation wage has much stronger effect than job attributes – .155. Column (2) in Table 1 confirms this conclusion. Other models in Table 3 confirm the importance of job attributes on effort, as well as relatively larger importance of ability (as proxied by the reservation wage).

5.9 External Validity

How does our unique experimental environment bear on the external validity of my results? Our environment is very similar to the model described in Weiss [1980], much more so than any other labor market – there are no interviews, no resumes, identity of the workers is unknown to employers, screening on wages is the only screening available

[15] There were five images in every HIT. The proportion of the disagreeable images varied experimentally from 0 to 100% with a step of 20%.

[16] One would have to calculate the cost-benefit to the firm from optimizing job attributes for workers to make a policy conclusion in this context.

to the employers. Our environment is uniquely suited to testing the sorting theory and to sort it out from the shirking model.

As far as the shirking model is concerned, Mechanical Turk is based on a piece rate contract. Where applicable, piece rate contract is one of the best ways to maximize efficiency (Lazear [2000]). On Mechanical Turk one has to submit a text field with content and it is easy to implement checks to make sure that the field is not empty. The combination of these features makes sure that workers submit as much work as possible while making sure they are in fact submitting it. In this sense there is less room for efficiency wage theory in this environment than others. Positive finding on efficiency wage theory at work within this environment provides an effective lower bound for the role that the efficiency wage theories would have elsewhere. My findings indicate that even in a simple piece rate environment it is impossible to completely specify the nature of desired output resulting in variation in its quality.

6 Conclusion

The literature has been treating efficiency wage theory either as a single theory exhibiting a positive relationship between wages and productivity or singling out the shirking model as the efficiency wage theory. As a result, the empirical tests either provide only evidence on the shirking model or evidence that is broadly consistent with all three/four models of efficiency wages. This paper uses a field experiment that provides a measure of reservation wages of workers along with their actual wages in an environment where workers are anonymous and all their characteristics are unobservable. I find that the labor supply behavior of these workers is consistent with both shirking and heterogeneous agent models, a conclusion providing validity to Stiglitz's claim that having workers post bonds when they enter employment is not a feasible strategy to eliminate wage premia. Additionally, I find evidence that workers' propensity to be motivated by wages falls with the nearing end of their tenure, a finding hard to reconcile with the gift-exchange model of the efficiency wage theory. These dynamic effects also suggests a role of labor turnover in generating wage premia and job queues, a role inter-connected with the roles of heterogeneity and incentives.

Appendices

A Summary Statistics of the Data

Figure 11: Histogram of Wages for Abandoned HITs only

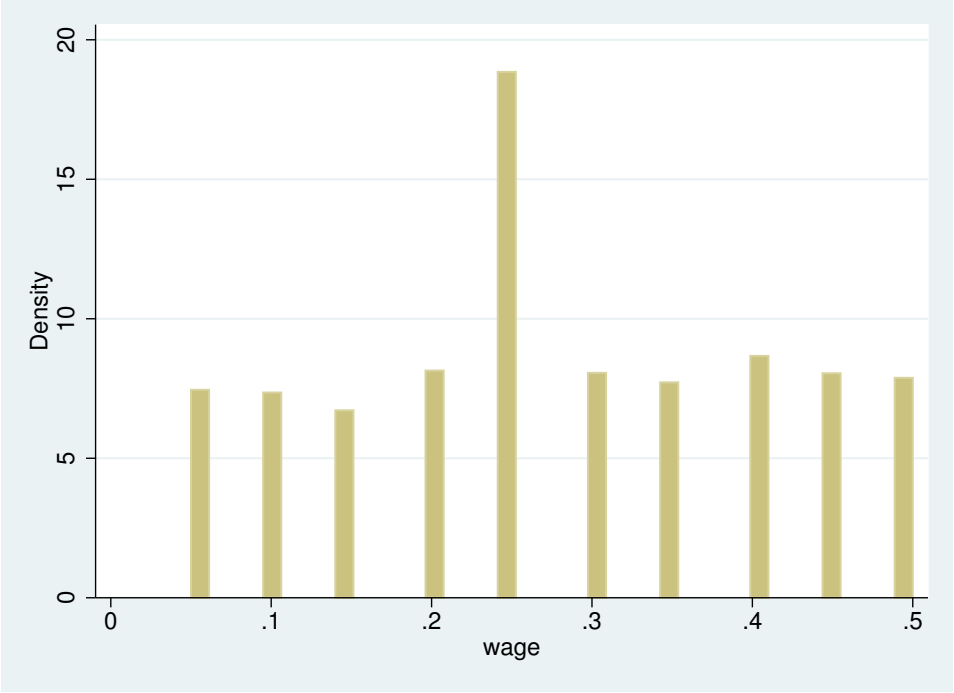


Figure 12: Histogram of Wages for all HITs

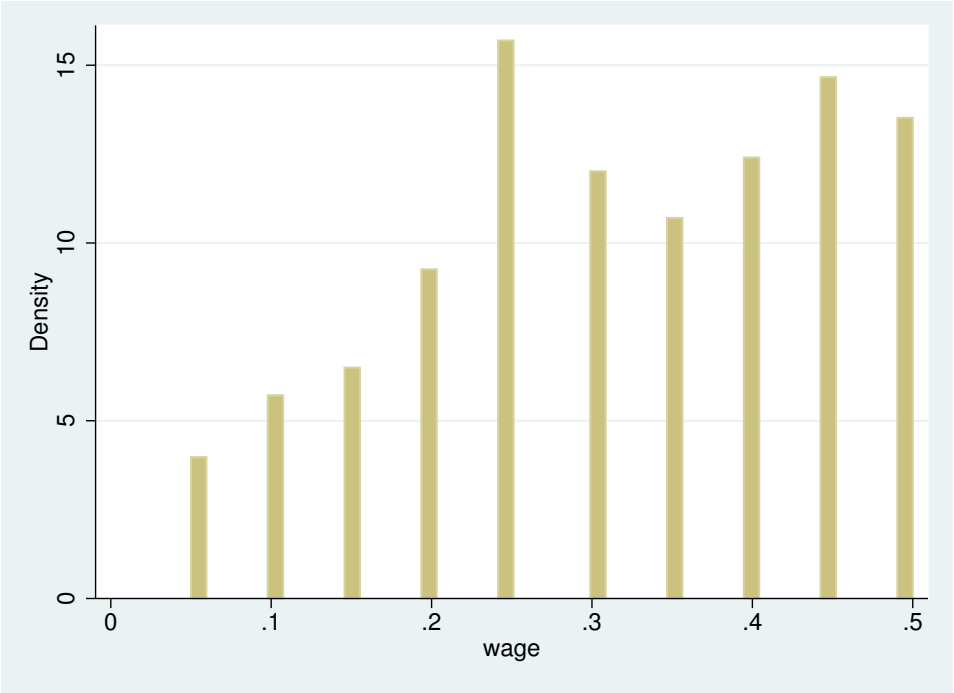


Figure 13: Histogram of Hourly Wages for submitted HITs

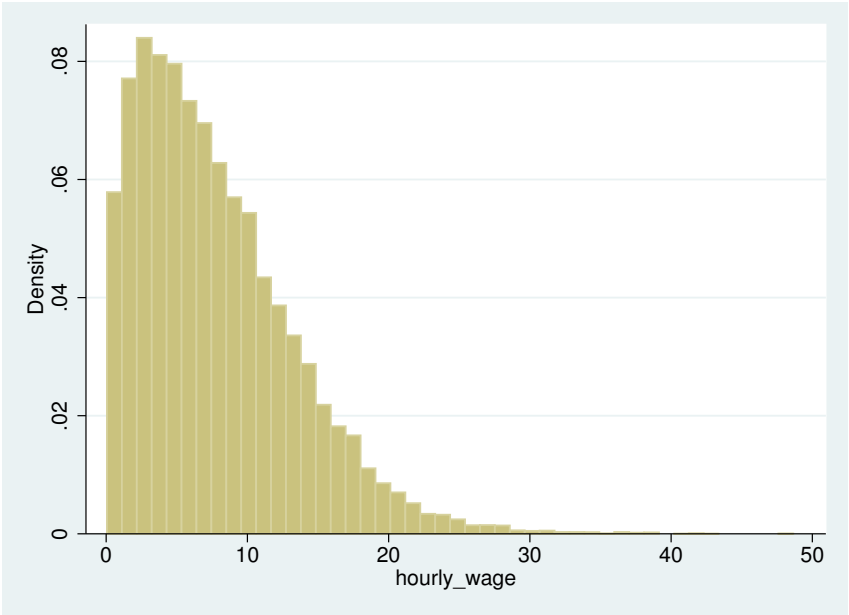
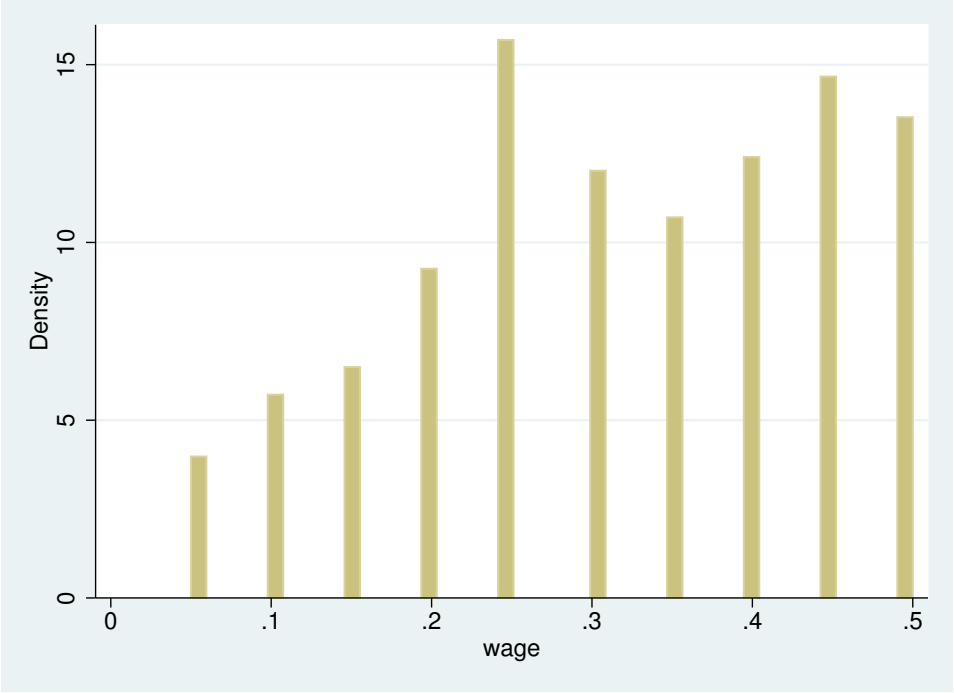


Figure 14: Histogram of Min Accepted Wages for all HITs



B Censored Regression Models vs non-Censored ones

As the Figures in Section 5 show the quality measure of output seems to be restricted from the top in a sense that workers would want to provide more quality but were not able to because of the experimental design (design of the HIT). The Tables in this section further show that the results from the censored regression model make more sense and are very different from the results that does not model the censoring properly.

Table 7: First HIT Analysis Cross Section

	Dependent variable:	
	correct_ratings1	correct_ratings
	censored regression	*OLS*
	(1)	(2)
log(wage)	−0.065	0.010
	(0.119)	(0.052)
min_accepted_wage	0.305	0.007
	(0.574)	(0.255)
disagreeable	−0.038***	−0.017***
	(0.002)	(0.001)
success_rate	−0.003	−0.001
	(0.003)	(0.001)
as.factor(day)2	0.354*	0.149*
	(0.193)	(0.086)
as.factor(day)3	0.510**	0.178*
	(0.218)	(0.098)
as.factor(day)4	0.424**	0.172*
	(0.212)	(0.096)
as.factor(day)5	0.147	0.006
	(0.212)	(0.096)
as.factor(day)6	0.234	0.102
	(0.217)	(0.100)
logSigma	0.515***	
	(0.034)	
Constant	6.956***	0.578***
	(0.399)	(0.178)
Observations	1,741	1,741
R^2		0.268
Adjusted R^2		0.264
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8: Between Effects Estimator Demonstrating Selection Effects

	Dependent variable:			
	correct_ratings			correct_ratings1
		panel		*censored*
		linear		*regression*
	Worker-HIT BE	Worker-HIT BE	Worker-HIT BE	Worker-HIT BE
	(1)	(2)	(3)	(4)
log(wage)	0.061	0.043		-0.267^{***}
	(0.050)	(0.032)		(0.069)
min_accepted_wage	-0.099		0.104	1.313^{***}
	(0.215)		(0.137)	(0.298)
disagreeable	-0.017^{***}	-0.017^{***}	-0.017^{***}	-0.023^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
success_rate	-0.001	-0.001	-0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)
as.factor(day)2	0.294^{***}	0.294^{***}	0.292^{***}	
	(0.103)	(0.103)	(0.103)	
as.factor(day)3	0.349^{***}	0.352^{***}	0.352^{***}	
	(0.111)	(0.111)	(0.111)	
as.factor(day)4	0.257^{**}	0.261^{**}	0.262^{**}	
	(0.110)	(0.110)	(0.110)	
as.factor(day)5	0.126	0.130	0.133	
	(0.107)	(0.107)	(0.107)	
as.factor(day)6	0.254^{**}	0.256^{**}	0.258^{**}	
	(0.104)	(0.104)	(0.104)	
logSigma				-0.028^{*}
				(0.017)
Constant	0.584^{***}	0.533^{***}	0.447^{***}	5.149^{***}
	(0.182)	(0.145)	(0.143)	(0.211)
Observations	1,780	1,780	1,780	1,776
R^2	0.295	0.295	0.294	
Adjusted R^2	0.293	0.293	0.293	
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: OLS Regressions of Effort with Errors Clustered on the Day Level

	Dependent variable:		
	correct_ratings		correct_ratings1
	OLS		*censored regression*
	(1)	(2)	(3)
log(wage)	0.011	0.048**	−0.049
	(0.044)	(0.022)	(0.037)
HITsLeft	0.001**	0.0005	0.004***
	(0.001)	(0.001)	(0.001)
disagreeable	−0.014***	−0.014***	−0.038***
	(0.001)	(0.001)	(0.0005)
success_rate	−0.001*	−0.001*	−0.003***
	(0.001)	(0.001)	(0.001)
HITsDone	0.001**	0.001**	0.003***
	(0.0004)	(0.0004)	(0.0003)
min_accepted_wage	0.375**	0.350**	1.052***
	(0.152)	(0.139)	(0.119)
as.factor(day)2	−0.006	−0.012	0.009
	(0.017)	(0.018)	(0.079)
as.factor(day)3	0.125***	0.119***	0.222***
	(0.027)	(0.028)	(0.080)
as.factor(day)4	0.151***	0.143***	0.331***
	(0.025)	(0.027)	(0.078)
as.factor(day)5	0.173***	0.165***	0.379***
	(0.046)	(0.048)	(0.079)
as.factor(day)6	0.141**	0.135**	0.298***
	(0.065)	(0.067)	(0.081)
log(wage):HITsLeft	0.001**		0.003***
	(0.0004)		(0.001)
logSigma			0.654***
			(0.009)
Constant	0.434**	0.490***	7.242***
	(0.170)	(0.161)	(0.117)
Observations	41,963	41,963	41,963
R^2	0.218	0.217	
Adjusted R^2	0.217	0.217	
Residual Std. Error	0.885 (df = 41950)	0.885 (df = 41951)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10: Effort Regressions Distinguishing Selection from Incentives

	Dependent variable:			
	correct_ratings			correct_ratings1
	panel			*censored*
	linear			*regression*
	Worker-Day FE	Worker-Day Pooling	Worker-Day RE	Worker-Day RE
	(1)	(2)	(3)	(4)
log(wage)	0.005	0.025	0.015	−0.083**
	(0.026)	(0.029)	(0.024)	(0.041)
min_accepted_wage		0.101	0.053	0.537***
		(0.144)	(0.156)	(0.206)
disagreeable	−0.015***	−0.016***	−0.015***	−0.025***
	(0.0005)	(0.0004)	(0.0004)	(0.001)
success_rate	−0.0002	−0.001	−0.0005	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)
logSigma				0.060***
				(0.014)
Constant		0.663***	0.613***	5.761***
		(0.094)	(0.088)	(0.140)
Observations	3,131	3,131	3,131	3,131
R^2	0.415	0.290	0.337	
Adjusted R^2	0.179	0.289	0.336	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Worker-HIT Analysis

	Dependent variable:			
	correct_ratings			correct_ratings1
		panel		*censored*
		linear		*regression*
	Worker-HIT FE	Worker-HIT RE	Worker-HIT Pooling	Worker-HIT RE
	(1)	(2)	(3)	(4)
log(wage)	0.028***	0.028***	0.065***	0.044***
	(0.009)	(0.009)	(0.009)	(0.012)
min_accepted_wage		0.042	0.234***	0.608***
		(0.124)	(0.041)	(0.062)
disagreeable	−0.013***	−0.014***	−0.014***	−0.034***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)
success_rate	−0.001*	−0.001**	−0.001***	0.00001
	(0.0003)	(0.0003)	(0.0002)	(0.0003)
as.factor(day)2	0.150***	0.163***	−0.018	0.147***
	(0.026)	(0.026)	(0.027)	(0.042)
as.factor(day)3	0.239***	0.246***	0.123***	0.297***
	(0.026)	(0.025)	(0.027)	(0.042)
as.factor(day)4	0.282***	0.285***	0.153***	0.381***
	(0.026)	(0.025)	(0.026)	(0.041)
as.factor(day)5	0.302***	0.299***	0.175***	0.420***
	(0.025)	(0.025)	(0.026)	(0.040)
as.factor(day)6	0.285***	0.287***	0.146***	0.537***
	(0.025)	(0.025)	(0.026)	(0.040)
logSigmaMu				−0.010*
				(0.005)
logSigmaNu				0.326***
				(0.003)
Constant		0.316***	0.595***	6.611***
		(0.049)	(0.035)	(0.052)
Observations	41,963	41,963	41,963	41,963
R^2	0.187	0.194	0.215	
Adjusted R^2	0.179	0.194	0.215	

Note: *p<0.1; **p<0.05; ***p<0.01

C Opportunity Costs over Time

Given my experimental design the average wages should be constant over the duration of the experiment (6 days). In one of my specification above I have used the lowest accepted wage over the experiment duration as a proxy for opportunity cost. This implicitly assumes that opportunity costs of workers are not changing over time. To make this assumption a bit more explicit I will investigate whether wage of accepted HITs change over the course of the days of the experiment by using the following specification:

$$wage = \alpha + \beta_1 * I(day == 1) + \beta_2 * I(day == 2) + \cdots + \epsilon \tag{30}$$

Results from this regression are presented in Table 12. We see some changes on day2 but the size of this effect is less than 1 percent of the baseline. Furthermore, I have run the original regression with opportunity cost proxy from the body of my paper using interaction with days and the positive relationship between productivity and opportunity costs of workers are robust across all days in my experiment (see Table **??**).

Table 12: Accepted Wages over the Duration of the Experiment

	Dependent variable:
	wage
day1 (baseline)	0.320^{***}
	(0.003)
day2	-0.012^{***}
	(0.003)
day3	0.004
	(0.003)
day4	-0.003
	(0.003)
day5	-0.005^{*}
	(0.003)
day6	0.002
	(0.003)
Observations	55,606
R^2	0.002
Adjusted R^2	0.002
Residual Std. Error	0.129 (df = 55600)
F Statistic	18.993^{***} (df = 5; 55600)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 13: Worker HIT Panel Linear Random Effects Accepted Hourly Wage and Quality over the Course of the Experiment

	(1)	(2)
	hourly_wage	correct_ratings
day=1	0	0
	(.)	(.)
day=2	0.174	0.119***
	(1.20)	(4.16)
day=3	0.984***	0.271***
	(6.82)	(9.51)
day=4	1.125***	0.163***
	(7.88)	(5.79)
day=5	1.359***	0.272***
	(9.67)	(9.79)
day=6	2.206***	0.229***
	(15.64)	(8.20)
Constant	4.355***	−0.342***
	(28.59)	(−10.20)
Observations	41967	41963

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

References

George A Akerlof. Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, pages 543–569, 1982.

George A Akerlof and Lawrence F Katz. Do deferred wages dominate involuntary unemployment as a worker discipline device?, 1986.

Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar. Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469, 2009.

Michael Buhrmester, Tracy Kwang, and Samuel Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6:3–5, 2011.

Jeremy I. Bulow and Lawrence H. Summers. A theory of dual labor markets with application to industrial policy, discrimination, and keynesian unemployment. *Journal of Labor Economics*, 4(3):376–414, 1986.

Peter Cappelli and Keith Chauvin. An interplant test of the efficiency wage hypothesis. *The Quarterly Journal of Economics*, pages 769–787, 1991.

Lorne Carmichael et al. Can unemployment be involuntary? comment [equilibrium unemployment as a worker discipline device]. *American Economic Review*, 75(5): 1213–14, 1985.

Linda M Collins, John J Dziak, Kari C Kugler, and Jessica B Trail. Factorial experiments: efficient tools for evaluation of intervention components. *American journal of preventive medicine*, 47(4):498–504, Oct 2014. doi: 10.1016/j.amepre.2014.06.021.

Josse Delfgaauw and Robert Dur. Signaling and screening of workers' motivation. *Journal of Economic Behavior & Organization*, 62(4):605–624, 2007.

Costanca Esteves-Sorenson, R. Vincent Pohl, and Ernesto Freitas. Wage premiums, shirking deterrence, gift exchange and employee quality: Firm evidence. Technical report, 2016.

Ronald Fisher. *The Design of Experiments*. Macmillan, 1935.

Uri Gneezy and Aldo Rustichini. Pay enough or don't pay at all. *Quarterly journal of economics*, pages 791–810, 2000.

Erica L Groshen and Alan B Krueger. The structure of supervision and pay in hospitals. *Industrial & Labor Relations Review*, 43(3):134S–146S, 1990. TO READ: test on whether supervision is a subsitute for rents in worker employment.

J. Horton. The condition of the turking class: are online employers fair and honest? *Economic Letters*, 2011.

P. Ipeirotis. Mechanical Turk: The Demographics. `http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html`, 2008. URL `http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html`. Accessed: 9/18/2009.

P. Ipeirotis. Demographics of mechanical turk. *New York University Working Paper*, 2010.

Alan B Krueger and Lawrence H Summers. Efficiency wages and the inter-industry wage structure. *Econometrica: Journal of the Econometric Society*, pages 259–293, 1988.

Edward P. Lazear. Performance pay and productivity. *American Economic Review*, 90(5):1346–1361, 2000.

David I Levine. Can wage increases pay for themselves? tests with a productive function. *The Economic Journal*, 102(414):1102–1115, 1992.

MBO Partners. The state of independence in america - third annual independent workforce report. Technical report, Sep 2013.

Canice Prendergast. The provision of incentives in firms. *Journal of economic literature*, 37(1):7–63, 1999. skimmed: very good overview of incentives, worker supply of quality labor and the form of contracts firms use.

Claus C. Pörtner, Nail Hassairi, and Michael Toomim. Only if you pay me more: Field experiments support compensating wage differentials theory. Technical report, SSRN, 2015.

Daniel MG Raff. Wage determination theory and the five-dollar day at ford. *The Journal of Economic History*, 48(02):387–399, 1988. TO READ: anecdote on the use of efficiency wages.

James B Rebitzer. Is there a trade-off between supervision and wages? an empirical test of efficiency wage theory. *Journal of Economic Behavior & Organization*, 28(1): 107–129, 1995.

James B Rebitzer and Lowell J Taylor. Efficiency wages and employment rents: The employer-size wage effect in the job market for lawyers. *Journal of Labor Economics*, pages 678–708, 1995.

Steven C Salop. A model of the natural rate of unemployment. *The American Economic Review*, 69(1):117–125, 1979.

Carl Shapiro and Joseph E. Stiglitz. Equilibrium unemployment as a worker discipline device. *The American Economic Review*, 74(3):pp. 433–444, 1984.

Carl Shapiro and Joseph E. Stiglitz. Can unemployment be involuntary? reply. *The American Economic Review*, 75(5):1215–1217, 1985.

Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Straman and T. Cadell, 1776.

Sushil B Wadhwani and Martin Wall. A direct test of the efficiency wage model using uk micro-data. *Oxford Economic Papers*, 43(4):529–548, 1991.

Andrew Weiss. Job queues and layoffs in labor markets with flexible wages. *The journal of political economy*, pages 526–538, 1980.

C.F.J. Wu and M.S. Hamada. *Experiments: Planning, Analysis, and Optimization*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118211533. URL `https://books.google.com/books?id=SBgehORJ7hkC`.