Effort, Wage Premia, and Reservation Wages: A Field Test of the Shirking and Sorting Efficiency Wage Models

Nail Hassairi

Department of Economics

University of Washington

November 16, 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 a US freelance 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 provide an explanation to a variety of puzzling labor market phenomena – non-competitive wage premia, job queues, inter-industry wage differentials, and more – by arguing they came about as a result of actions of agents optimizing in the presence of informational asymmetries. The theory was first presented as a moral hazard problem, with a principal being unable to observe shirking. Employers offer wage premia to create a stake in the employment relationship. These wage premia are 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 ensures that workers do not shirk in equilibrium. Eventually, all firms offer wage premia to compete for workers, resulting in a job rationing and queues. Consequently, job queues take over the role of the disciplining device.

Carmichael et al. [1985] criticized this explanation, offering that workers could a) buy jobs if they were in a job queue (e. g. via re-location to a region with lower unemployment), or b) post performance bond that the employer would get to keep if they were found shirking. Shapiro and Stiglitz [1985] counter this "bonding critique", as it has come to be known, by pointing to another information asymmetry underlying efficiency wages – unobserved worker heterogeneity. The sorting model of the efficiency wage theory argue that lower wages attract lower quality labor force (Weiss [1980]). Both performance bonds or job buying would decrease the job value for workers, an effect equivalent to a lower wage. Consequently, to evaluate the relevance of the efficiency wage theory it is important to establish not only whether higher wages cause an increase in worker productivity, as demonstrated by Cappelli and Chauvin [1991], but also whether this increase is driven by incentives or heterogeneity.

My paper follows in the footsteps of Cappelli and Chauvin [1991], Esteves-Sorenson, Pohl, and Freitas [2016] and Guiteras and Jack [2014]. Cappelli and Chauvin [1991] convincingly demonstrated the causal effects of wage premia on productivity. A rather less convincing was their argument that this is primarily driven by incentives Esteves-Sorenson et al. [2016] studied screening in a context of Portuguese apprenticeship system and show that screening leads to selection for workers with better observable characteristics (education etc). Finally, Guiteras and Jack [2014] study selection and

heterogeneity using reservation and actual wages in a developing country context. My paper is the first to separate incentive and selection effects of wage premia in the context of a developed country labor market, and the first to find positive effects of both.

My paper utilizes data from a field experiment on Amazon Mechanical Turk online labor market for freelance labor. This labor market has characteristics similar to what the literature usually refers to as a secondary job market – routine, non-skilled work. The only skill required is common sense and very basic computer literacy. This makes it suitable to test the efficiency wage theory since job queues are most often found in the non-skilled sector of the economy (Bulow and Summers [1986]). The advantage of this platform for the purposes of a field experiment is that there is no face-to-face contact between employers and employees, ruling out experimenter bias. The lack of face-to-face contact also prevents workers from engaging in rent-seeking behavior as described in Prendergast [1999] (fraternizing with managers etc.) Additional advantage is that there are very few frictions in this market and there is a large and representative population of workers available. This contrasts with the institutional context in which Guiteras and Jack [2014]. Finally, workers on Mechanical Turk have very little opportunity to talk to each other and come from locales that are far away from each other preventing dissemination of information about the various conditions of the experiment. Consequently, if some participants find a way to game the experiment, this strategy is unlikely to spread across the pool of the experiment participants. Mechanical Turk also protects workers' anonymity so that there are no observable characteristics. This prevents gender or race discrimination, which in turn leads to more representative population of workers on this platform.

The experiment took 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. The job offered to workers consisted of marking images as appropriate or inappropriate for sensitive audiences (due to the possible presence of nudity, violence, explicit content etc.) The images were pre-selected by experimenters as belonging to either category and the workers' assessment was compared with the researchers assessment, providing a measure of effort. Our effort measure conformed to expectations, being positively correlated with the level of monitoring, experience,

wages and reservation wages. Workers were told that their work will be reviewed and payment will only be made to submissions satisfying a certain standard. They were also told that a certain proportion of workers on a given day were not paid due to the low quality nature of the work submitted. This information was experimentally varied to gauge an impact of monitoring on worker performance. In reality, all work was paid for. So this treatment was another intentional misinformation (on top of the task being real). Workers' wages were varied, allowing me to estimate incentive effect of wages while learning about workers' reservation wages and estimating the impact of their heterogeneity on performance. Workers were being experimented on without their knowledge and the job offered was fairly similar to other jobs in this labor market, ruling out Hawthorne effect (again in contrast to Guiteras and Jack [2014]). A large sample of workers (1,800) was recruited and the acceptance rate was around 40%. This is more than twice the size of the sample recruited in Guiteras and Jack [2014].

I construct a simple static model that in which effort increase the probability of being paid for the task and ability lowers the cost of effort. These two aspects merge the ideas from the shirking model (Shapiro and Stiglitz [1984]) and the adverse selection model (Weiss [1980]). By choosing an effort level worker reveals her cost of effort as well as her ability's effect on it. Since in the experiment the monitoring level was varied and the claim was made that low-quality work would not be paid for the utility function models beliefs of being paid as a function of effort, experience and stated level of monitoring. I derive the optimal level of effort as a function of reservation wage, current offered wage, job characteristics and worker's experience. The comparative statics show that the comparative importance of selection and incentives represents the comparative importance of ability in lowering the cost of effort and the importance of higher effort on increasing expected utility from income. 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." Additionally, the coefficient on reservation wage is twice as high as the one on actual wage demonstrating that at current levels of effort the cost of effort played larger role than the potential benefits from it.

My work is most closely related to Cappelli and Chauvin [1991], Esteves-Sorenson et al. [2016] and Guiteras and Jack [2014]. Cappelli and Chauvin [1991] finds that

selection has negative impact on productivity (not statistically significant), most likely due to their very noisy measure of reservation wage. They use wage premium from the year in which the median worker in their dataset was hired which means all workers except the median worker were offered a different wage at the time they were hired. In contrast, I observe the workers working at different wages in the course of six days and use the lowest accepted wage as a proxy for reservation wage. Additionally, I have data on how much time workers spent on the task and using imputed hourly reservation wage instead of just reservation piece rate yields equivalent results.

Esteves-Sorenson et al. [2016] only study selection on observable worker characteristics such as years of education. Unfortunately, studying observable worker characteristics does not help us understand how higher wages are used as a screening mechanism in the presence of unobservable worker productivity. Unlike them, I used a revealed preference argument to show that workers with higher reservation exhibit higher productivity. This actually corresponds to the adverse selection theory from Weiss [1980] in which it is the observable rather than unobservable characteristics that forced firms to use wages as a screening device. In the presence of observable differences between workers a firm simply pays the market price for a worker with those characteristics, there are no informational asymmetries, and no reason to expect wage premia. Esteves-Sorenson et al. [2016] also show that post-tenure shirking increases less in wage-premium-paying firms than firms paying competitive wages. They attribute this to the adverse selection story, however, this behavior is also consistent with the shirking model.

Both Cappelli and Chauvin [1991] and Esteves-Sorenson et al. [2016] could be potentially dealing with a situation in which workers receive what appear to be premium because the worker quality is observed to the firm (from the recruiting process) but unobserved to the econometrician. My study does not suffer from this problem. All workers willing to work are invited to submit their work so the sorting rather than being double-sided as in Cappelli and Chauvin [1991] or Esteves-Sorenson et al. [2016] is driven exclusively by workers selecting themselves into jobs.

Guiteras and Jack [2014] devise an ingenious way how to have workers reveal truthfully their reservation wages and use this information alongside actual wages paid to workers which are randomized. Curiously enough, they find that individuals with

higher reservation wages exhibit lower productivity. A number of possible explanations are possible. First, Guiteras and Jack [2014] offer that the rural labor market in Malawi, where the study took place, exhibits frictions. They argue that the men had higher reservation wages because of outside opportunities using different skillset than was required in the experimental task. As a consequence, mean with higher reservation wages in their experiment deliver output of lower quality. Why women exhibited no positive impact from higher reservation wages is not clear. I suspect some of the following possibly causes may be relevant. Firstly, workers were aware of being part of an experiment which introduces the possibility of a Hawthorne effect. Secondly, there was a possibility of an experimenter effect since there was personal interaction involved in running the experiment. Third, workers were not isolated from each other so they could have communicated with each other about the experiment treatment. Finally, I believe that Guiteras and Jack [2014] should have offered more than one task during the experiment for the sake of external validity since it appears that the task was poorly chosen to match the skillset of the recruited workers. In my experiment, the task offered was fairly standard work and the market large enough so that workers could sort themselves into task that best suited their preferences and abilities. My paper likewise studies the comparative effect of selection and incentives using reservation wages alongside actual wages in the context of a developed country online labor market while avoiding the issues mentioned above.

My work is part of a broader literature on testing efficiency wage theory. Early anecdote on the existence of efficiency wages comes from Smith [1776] and Raff [1988]. Krueger and Summers [1988] find inter-industry wage differentials that are not explained by regional, demographic and human capital variables. 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) to show that the link between wage premia and productivity is causal and that the direction is from wages to productivity. 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. Levine [1992]

finds that it is in fact profitable for the firm to pay efficiency wages using the PIMS line-of-business data. Esteves-Sorenson et al. [2016] confirm that savings from reducing absenteeism justify paying efficiency wages using data from the Portuguese tourism industry. All this evidence is broadly supportive of the efficiency wage theory without distinguishing between selection or heterogeneity. Rebitzer and Taylor [1995] find that law firms using tournament like incentive scheme for their associate employees still pay efficiency wages, evidence they find to be in contrast with the shirking model (incentive story) of the efficiency wage theory. Akerlof and Katz [1986] makes the case, however, that tournament schemes are not sufficient to resolve the moral hazard problem.

Rebitzer [1995] find a negative link between supervision (monitoring) and wages which they see as a confirmation of the shirking model of the efficiency wage theory since the production isoquant would imply a trade-off between the two. Groshen and Krueger [1990] reach the same conclusion using hospital data on supervision and wages. Esteves-Sorenson et al. [2016] finds that once a worker receives tenure, making her more difficult to fire, her absenteeism tends to increase using data on Portuguese workers in the tourism industry. However, as Prendergast [1999] points out a firm could move not only along the isoquant between monitoring and wages but also from one isoquant to another. Prendergast [1999] also points out that the interaction between wages and monitoring depends on the nature of the production technology of the firm and the nature of the monitoring technology. Monitoring and wages could be both substitutes or complements to each other. Consequently, Prendergast [1999] questions the feasibility of testing the shirking model by testing the link between monitoring and wages.

My work is also somewhat similar to Lazear [2000] in a sense that both study impact of selection and incentives on productivity. However, Lazear [2000] studies the impact of contract type, hourly rate vs piece rate, on productivity via both selection and incentives. However, a more efficient contract type rather than causing wage premia is a substitute for them as far as increasing productivity is concerned. So while heterogeneity and selection are discussed in the empirical literature on contract types, this literature is not of central importance to my paper.

Section 2 describes the experimental design, the Mechanical Turk labor market, and the structure of the data. Section 3 describes my theoretical model and identifi-

cation strategy. Section 4 provides results from the analysis of the experimental data. Section 6 concludes.

2 Experimental Design

2.1 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).[1] 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.

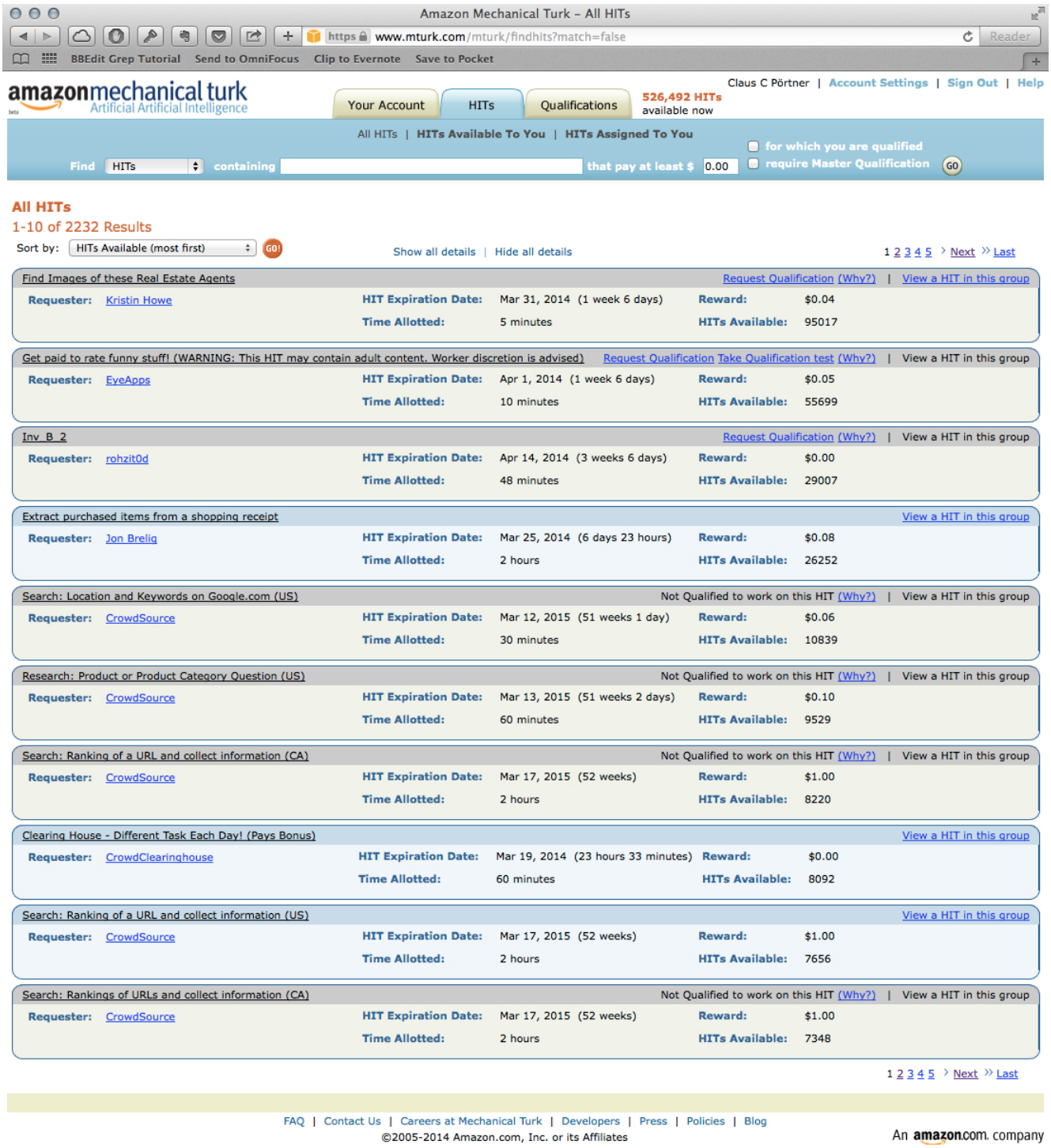
Work is paid per task, and although the corresponding hourly wage may not be typical of the overall US labor market, it will be close for workers on the current U.S. minimum wage.[2] [3] There are generally between 5,000 to 30,000 tasks completed each day [Ipeirotis, 2010]. Workers communicate on 3rd-party 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].

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

[2] 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.

[3] 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].[4] 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.

2.2 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.[5] 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

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

[5] 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.[6]

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.

[6] 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"

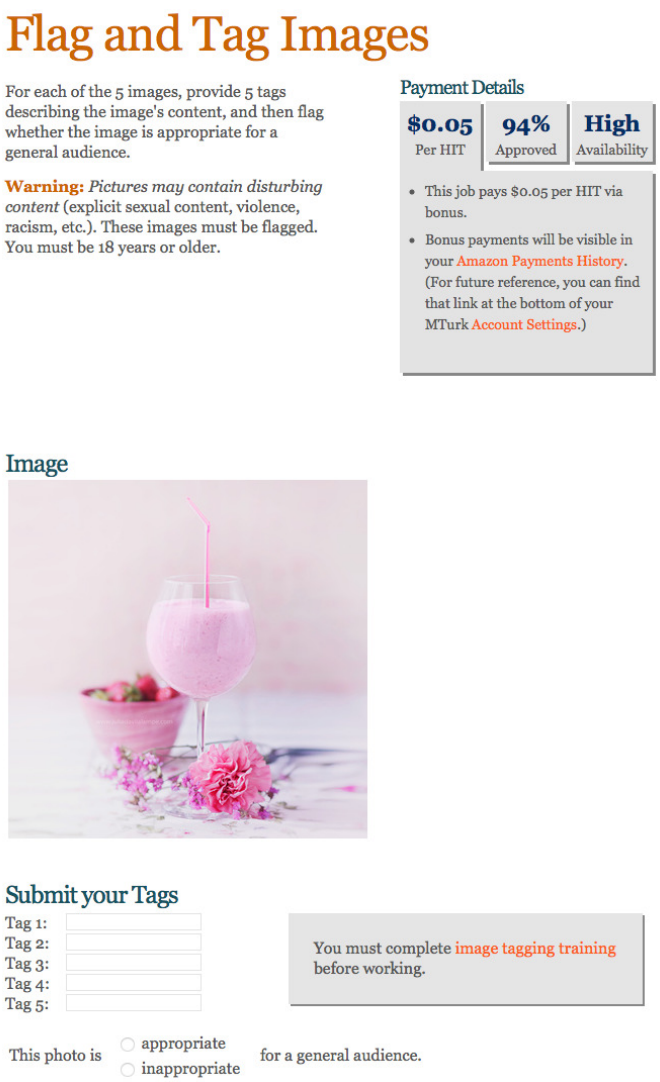
Mechanical Turk allows requestors to require skills and "certifications" of workers. Our only requirement is that the computer accessing the HITs must be in the US. This allows us to estimate consistent wage responses while achieving a sufficient sample size. It is possible to circumvent the location 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*[7]. 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

[7]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 2.4.

Figure 3: Image Tagging Experiment Page View

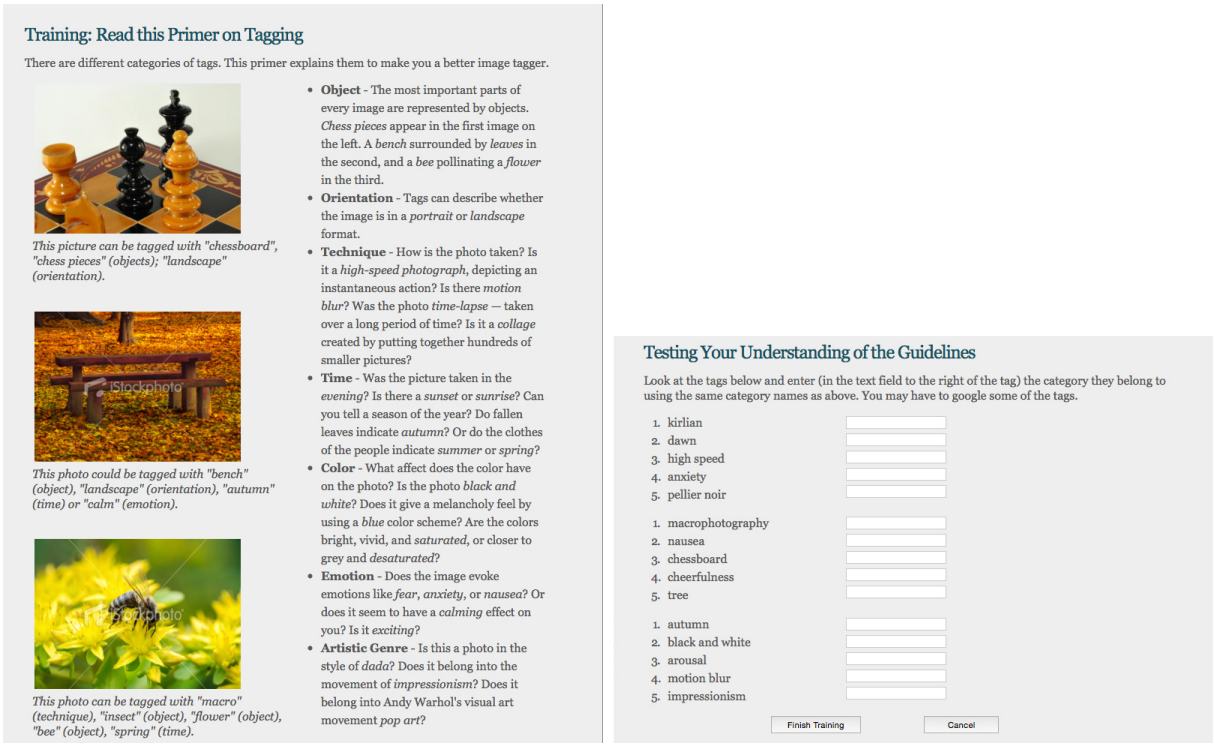


expressed as a ratio between 0 and 1. The number of disagreeable pictures do not change between HITs, but the ordering is randomly allocated, so that a worker with, say, one disagreeable image per HIT (20%) may see that as, for example, the first image on one page and as the third on the next. The agreeable 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.[8] This process

[8] 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.

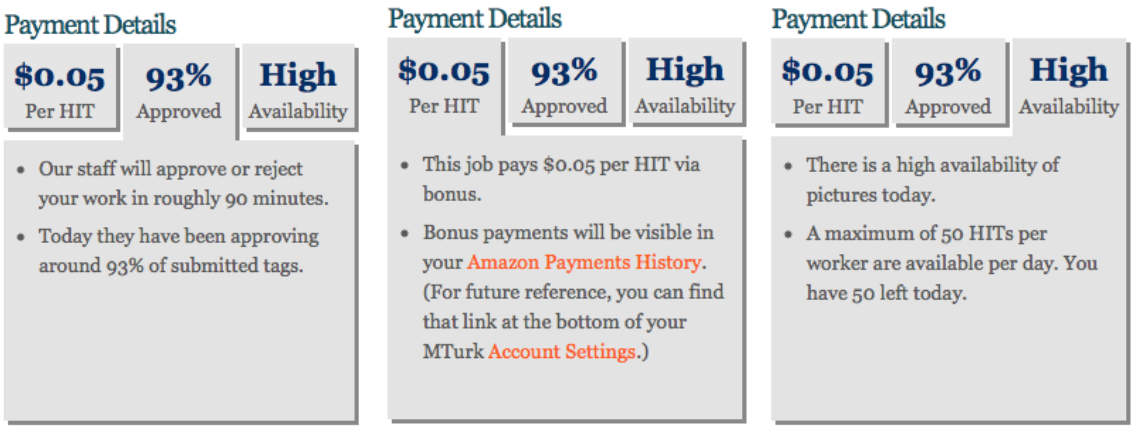
Figure 4: Image Tagging Experiment—Training and Test



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[9]. 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.

[9]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 for, as well as to see what is the incentive effect of increase the wage above this minimum.

2.3 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 2.1 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:

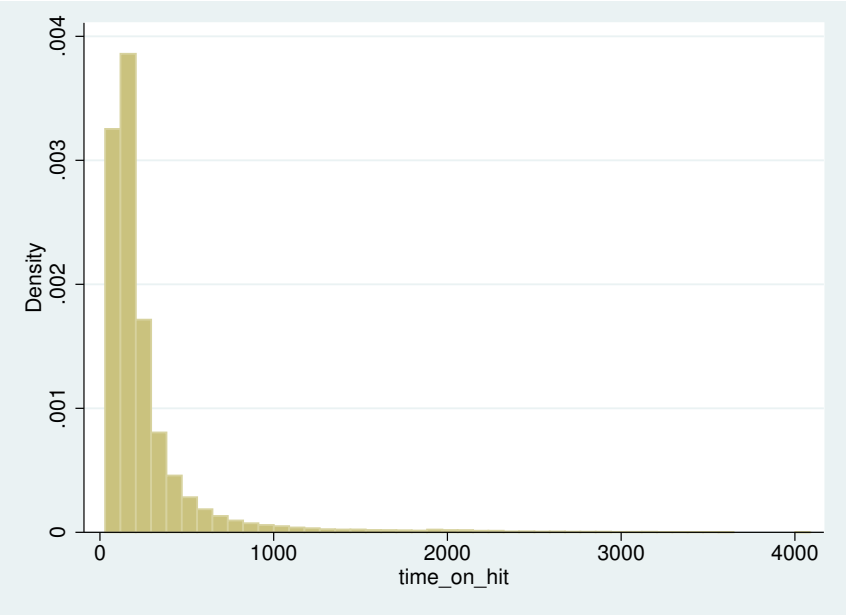
1. HIT count as a measure of experience (or job fatigue),
2. HITs left as a measure of how much longer a worker desires to work on my HITs (assuming rational expectations that are on average correct) – this will reveal whether incentive effects vary with tenure,
3. HIT quality,
4. time spent on HIT.

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 4 provides more details on the respective results obtained from the worker-HIT and worker-day models.

2.4 Output/Productivity Data

As described in Subsection 2.2 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".

2.5 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.

2.6 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 2, 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.

3 Theory

I construct a simple model in which principal's behavior is taken as exogenous (since it is experimentally varied) and the agent (worker) chooses her optimal effort level given the wage, her reservation wage, her experience, her preference for the given job attributes and the stated level of monitoring. Lower effort level will not lead to an increase in the probability of being fired (as is the case in Shapiro and Stiglitz [1984]) but rather to higher probability of not being paid for a given job (as is the setup in our experiment). The worker's ability will enter the cost of effort function since ability is assumed to lower the cost of effort. Ability will be proxied by worker's reservation wage (as proposed in Weiss [1980]). The model then combines the aspects of moral hazard from Shapiro and Stiglitz [1984] as well as the notion that higher reservation wage correlates with higher unobserved productivity proposed in Weiss [1980]. A participation constraint relating wage offered to reservation wage is not included as the participation decision is not a salient feature of my paper and would only serve to destract the exposition from the effort decision. I am assuming that reservation wage of a worker is constant throughout the experiment. Given that the experiment took place over the course of six days this assumption should not raise too many eyebrows (see Section 5.3 for more detailed discussion of this assumption). Negative job attribute enters the worker's cost of effort function. The worker's payoff function has the following form:

$$u(e) = [1 - P(e, p, n)]U(w) - C(e, \bar{w}, J) \tag{1}$$

where $P(e, p, n)$ is the probability of not receiving payment (work being judged as subpar) and $C(e, \bar{w}, J)$ is the cost of effort function. e stands for the effort level, p stands for the advertised probability of success (quality standards/monitoring), n is a number of HITs done by the work up until now (experience), w is the current wage, J is a job attribute and $\bar{w}$ is worker's reservation wage or opportunity cost.

Agent choose effort to maximize the payoff function in Equation 1. The first order condition for this problem is:

$$-P'(e, p, n)U(w) - \frac{\partial C}{\partial e}(e, \bar{w}, J) = 0 \tag{2}$$

While this cannot yield explicit solution for optimal effort without choosing func-

tional forms for probability and cost functions we can use the implicit function theorem to conduct some comparative statics analysis. This yields the following results for incentives:

$$\frac{\partial e^*}{\partial w} = \frac{-P'(e^*)U'(w)}{P_{ee}(e^*)U(w) + C_{ee}} \tag{3}$$

and the following for heterogeneity:

$$\frac{\partial e^*}{\partial \bar{w}} = \frac{-C_{e\bar{w}}}{P_{ee}(e^*)U(w) + C_{ee}} \tag{4}$$

These effects cannot be signed since the denominator's sign is ambiguous. It is in accordance with the law of diminishing marginal productivity to assume that $P_{ee}(e^*) < 0$, while it also commonplace in the literature to assume that $C_{ee} > 0$. The sign of the denominator would depend on the current level of wages and effort. I will be able to sign these using our data later on, however, I won't be able to make a judgement on whether this does or does not conform to the theory. The only thing I can say about the sign of these effects is that they should be the same since the uncertainty about the sign comes from the denominator and the denominator is the same for both of these effects. Hence, my focus will be on comparing the effects of heterogeneity and incentives.

Taking the ratio of the two comparative statics effects gives:

$$\frac{\frac{\partial e^*}{\partial \bar{w}}}{\frac{\partial e^*}{\partial w}} = \frac{C_{e\bar{w}}}{P'(e^*)U'(w)} = \frac{.189}{.089} = 2.1 \tag{5}$$

Equation 5 can be unambiguously signed if one is to believe that:

- $C_{e\bar{w}} \geq 0$ (ability has non-negative impact on the slope of the disutility of effort function with respect to effort)
- $P'(e^*) > 0$ – agent believes that effort leads to higher probability of success
- $U'(w) > 0$ – the marginal utility of income is positive

If all of the above conditions hold the fraction in Equation 5 has a positive sign. Of more interest is whether this fraction is bigger or smaller than 1. Theory has no prediction to this effect and I will learn this from our data.

The intuition behind Equation 5 is that heterogeneity is only important if $C_{e\bar{w}}$ is different from zero. This means that for heterogeneity to really matter ability should

not only levels of disutility of effort but also the slop of the disutility function as the effort increases. For an individual with less steap effort-disutility relationship incentives are more effective. Equation 5 also shows that the relationship between heterogeneity and incentives is one between effort disutility and gains from increased effort. It is important here that agents believe that probability of success increases with their effort and that they derive sufficient utility from the additional income given their current income levels.

4 Identification Strategy

In Section 3 I derived a general model of the agent's choice of effort and derived some comparative statics for the optimal effort choice would vary with wage and reservation wage (worker ability). In this section I will make use of the experimental data to estimate the following reduced form equation:

$$correctRatings^* = \beta_0 + \tag{6}$$

$$\beta_1 \log(wage) + \tag{7}$$

$$\beta_2 SuccessRate + \tag{8}$$

$$\beta_3 log(MinAcceptedWage) + \tag{9}$$

$$\beta_4 HITsDone + \tag{10}$$

$$\beta_5 Disagreeable + \epsilon \tag{11}$$

This reduced from equation will map to my comparative statics results, from Equation 3, Equation 4, and Equation 5, in the following way:

$$\frac{\partial e^*}{\partial w} = \beta_1 \tag{12}$$

$$\frac{\partial e^*}{\partial \bar{w}} = \beta_3 \tag{13}$$

$$\frac{\frac{\partial e^*}{\partial \bar{w}}}{\frac{\partial e^*}{\partial w}} = \frac{C_{e\bar{w}}}{P'(e^*)U'(w)} = \frac{\beta_3}{\beta_1} \tag{14}$$

4.1 Data Granularity

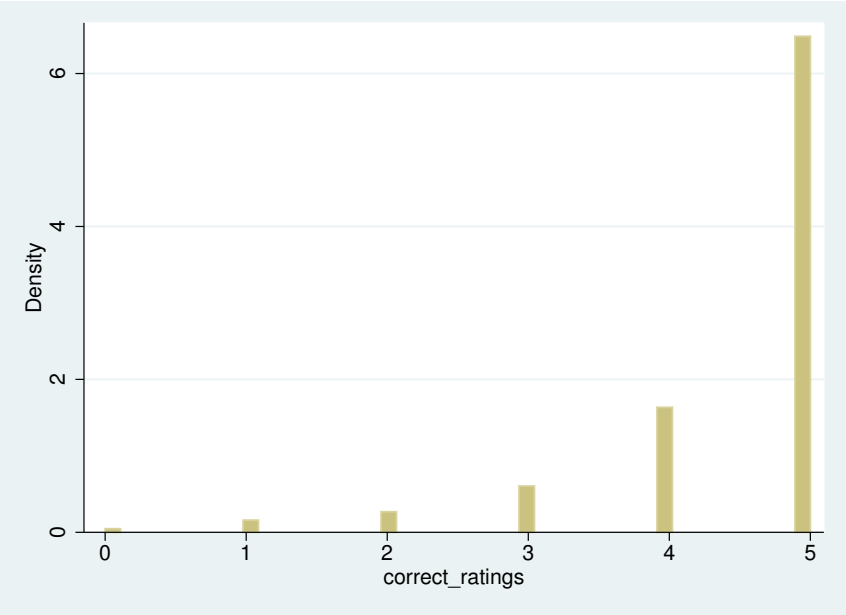
Table 1 and Table 3 provide results from a several econometric specifications. Following the discussion in Section 2.3 I have run several econometric specifications with different data granularity and emphasizing different sources of variation. My preferred specification is in Column (3) of Table 1. 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 decreasing 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).

4.2 Censored Dependent Variable

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 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

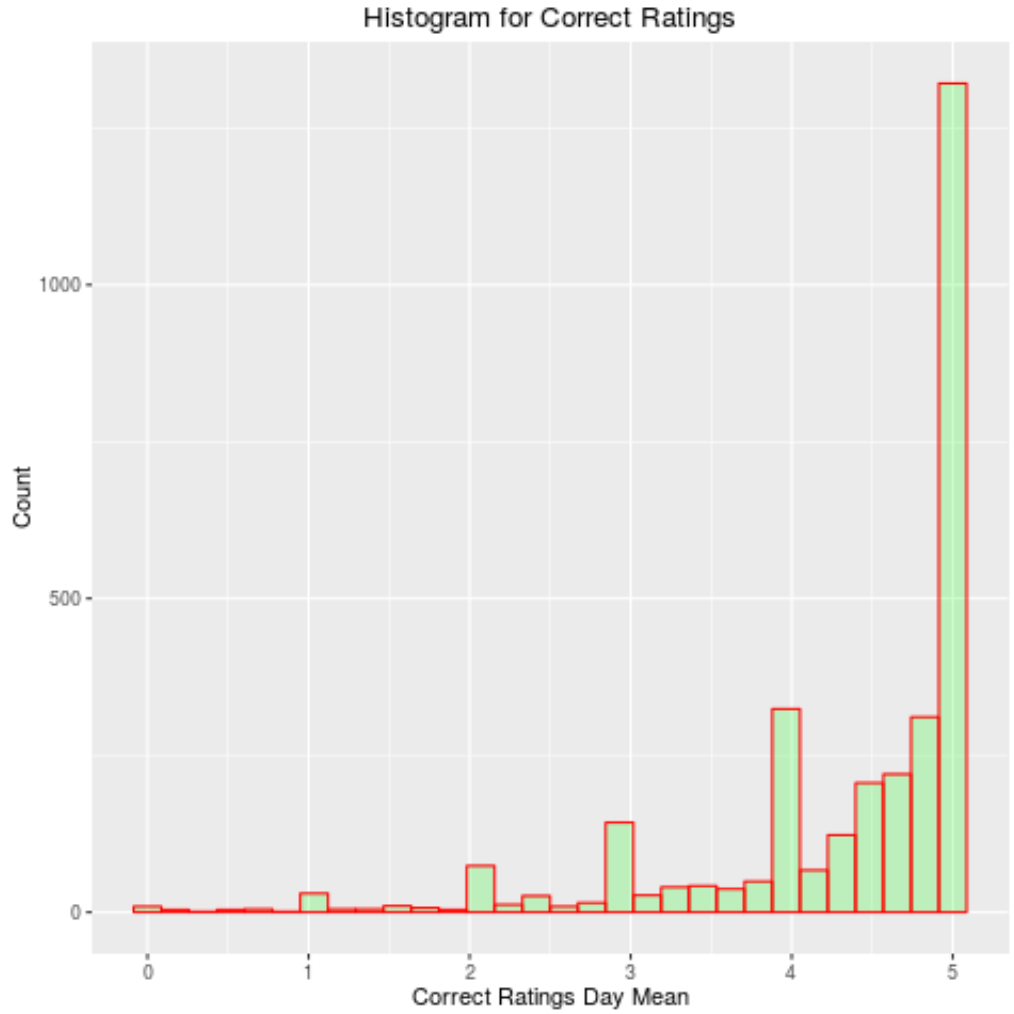


4.3 External Validity

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 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

Figure 8: Histogram of Correct Appropriateness Ratings Averaged over Days



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.

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

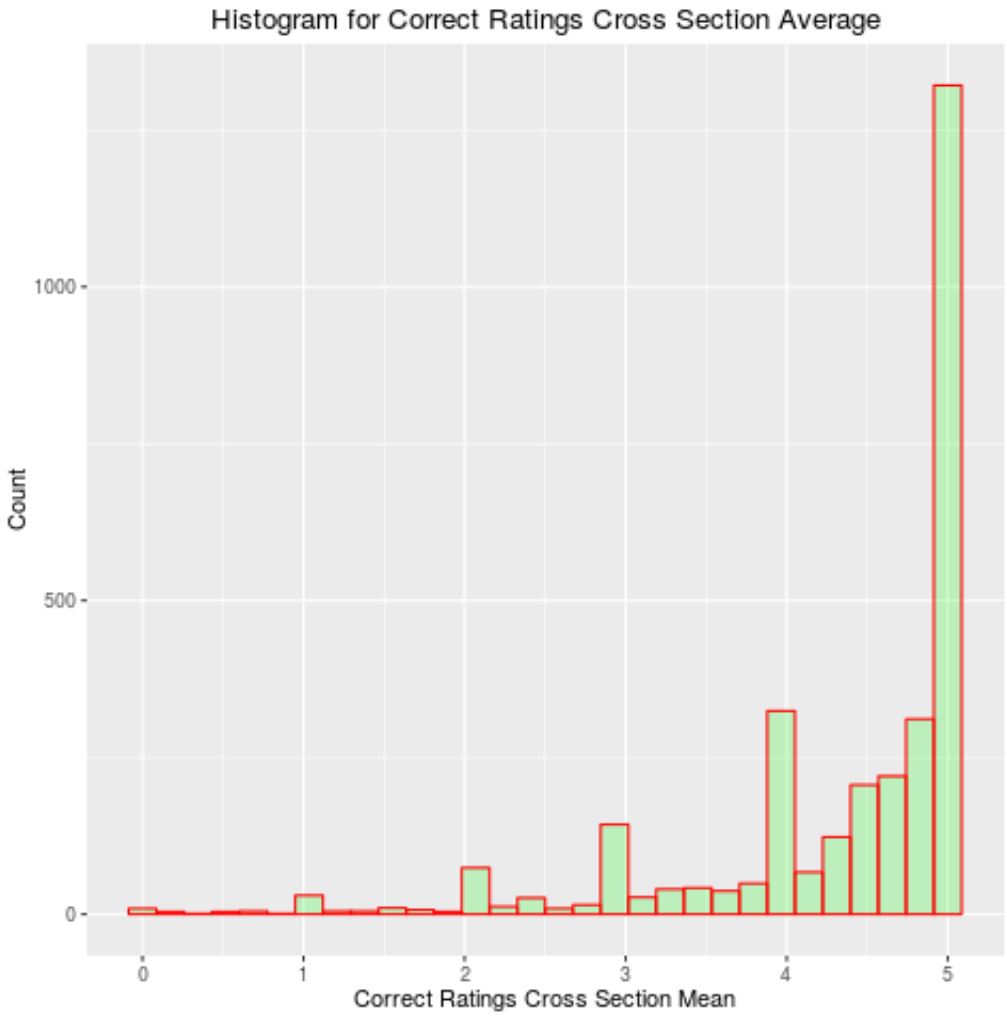
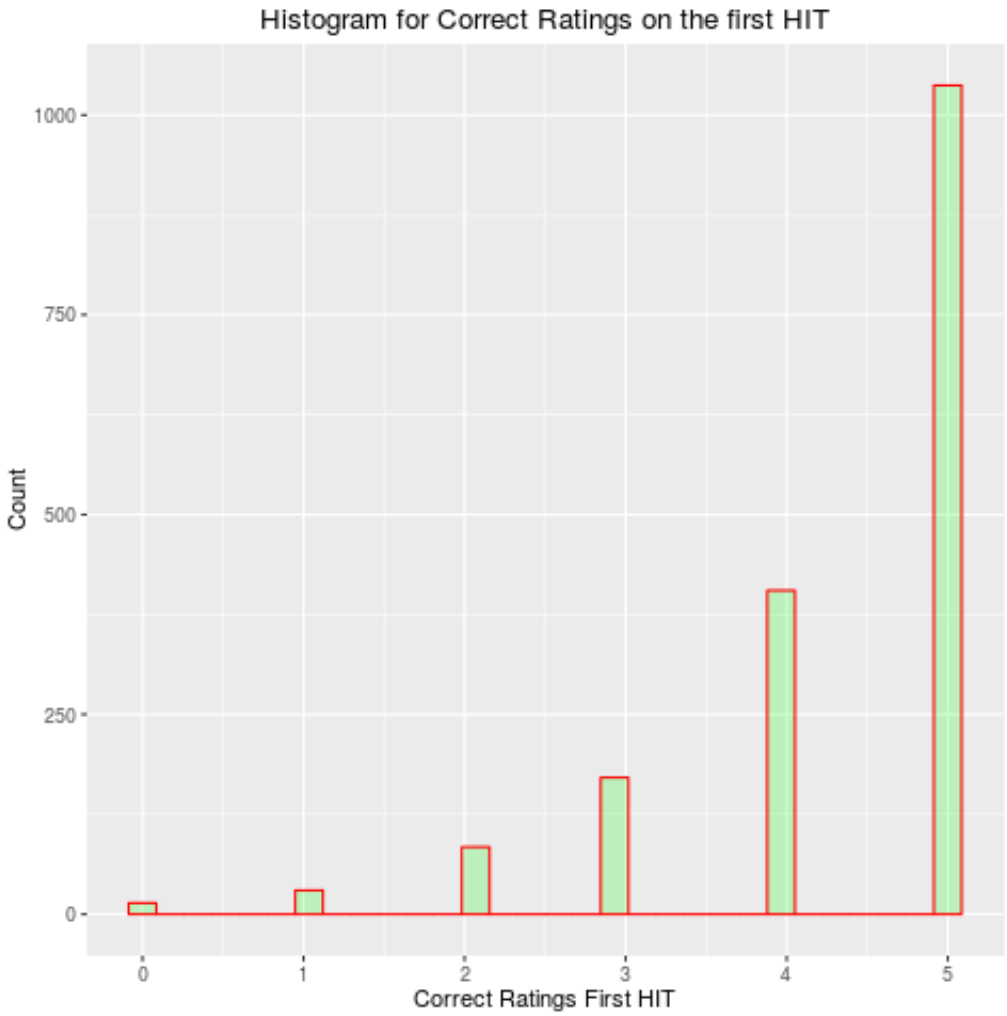


Figure 10: Histogram of Correct Appropriateness Ratings on the first HIT



5 Results

As I showed in Section 3 the theory that includes both effects of incentives and worker heterogeneity does not allow me to make conclusive statements about the signs of β_1 and β_3 from Equation 6. However, the theory does predict that these signs should be the same since the uncertainty about the sign comes from the denominator of Equation 3 and Equation 4 and these denominators are identical. The interesting conclusion from our theoretical model is that if the wage is too high (so that marginal utility from the wage is sufficiently low) and the costs of effort too high the effect of an increase in the wage might lead to lower optimal effort. So what does our data show? Are the wages too high? No. Column (3) in Table 1 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 2.4 for details on this measure of effort) in this model is $\beta_1 = .037$ and highly statistically significant. Column (2) in Table 1 shows equivalent results.

The coefficient on the regressor `log(min_accepted_wage)` in Table 1 provides the estimate of the parameter β_3 from Equation 6. As mentioned above this coefficient is ambiguously signed but it should have the same sign as the coefficient β_1. This is indeed the case in the results in Table 1, however, not so in Table 3. This further confirms my suspicion that those results reflect dynamic effects that are not the direct focus of this paper.

Section 3 discussed the ratio of β_3 and β_1. Using the results from Table 1 this ratio can now be estimated:

$$\frac{\frac{\partial e^*}{\partial \bar{w}}}{\frac{\partial e^*}{\partial w}} = \frac{C_{e\bar{w}}}{P'(e^*)U'(w)} = \frac{.189}{.089} = 2.1 \tag{15}$$

This suggests that heterogeneity is bigger issue than incentives in our data and that one can increase productivity by finding workers with lower cost of effort rather than incentivizing the average worker by offering them higher wages.

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
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

5.1 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.

5.2 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

5.3 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.

5.4 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 2.2 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[10]. 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[11]. 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).

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

[11] 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.

6 Conclusion

The literature has provided several demonstration of how more efficient choice of contracts improves productivity via both selection and incentive effects. In the case of wage premia there has been a lack of a convincing demonstration on incentive and selection effects. My paper has filled this gap using a field experiment in a US online labor market. The experiment was conducted without the workers knowledge of the real purpose which together with the experimental design and advantages of the Mechanical Turk marketplace allowed me to get a clean estimate of the incentive and selection effects of wage premia. This is important as the theoretical literature has been criticizing the shirking model of the efficiency wage theory as being the sole source of wage premia. Our results show that both incentives and selection are important considerations. Consequently, an alternative incentive scheme would not be sufficient to eliminate wage premia. Furthermore, I found signs that the incentive effect is dynamic, varies with tenure and this suggests that labor turnover interacts with incentives and selection to maintain wage premia. All this evidence is supportive of the relevance of the efficiency wage theory as an explanation for the existence of wage premia and job queues.

Appendices

A Summary Statistics of the Data

Figure 11: Histogram of Wages for Abandoned HITs only

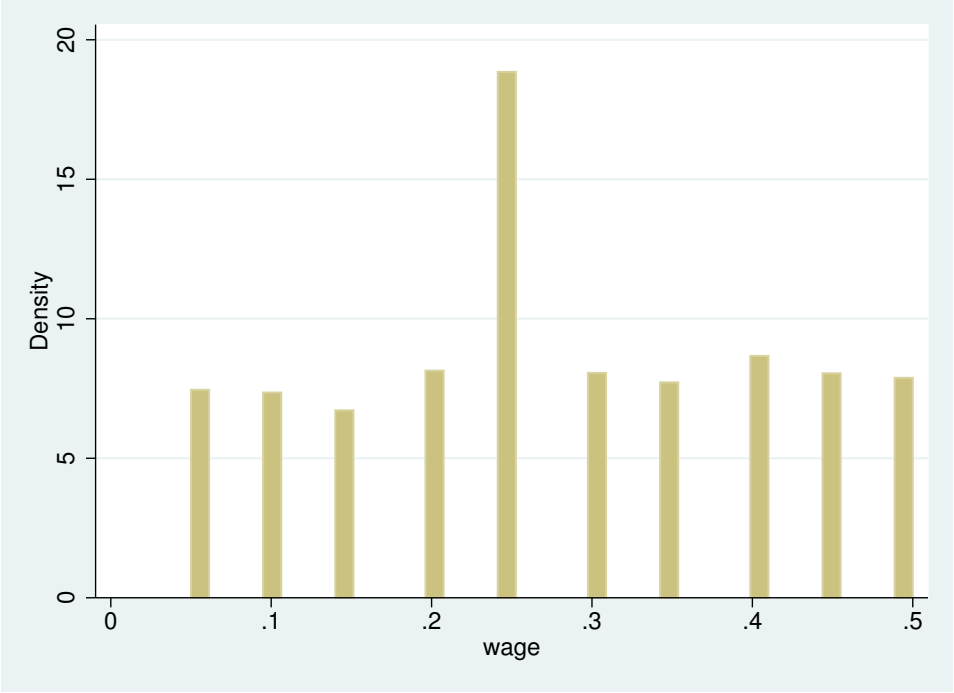


Figure 12: Histogram of Wages for all HITs

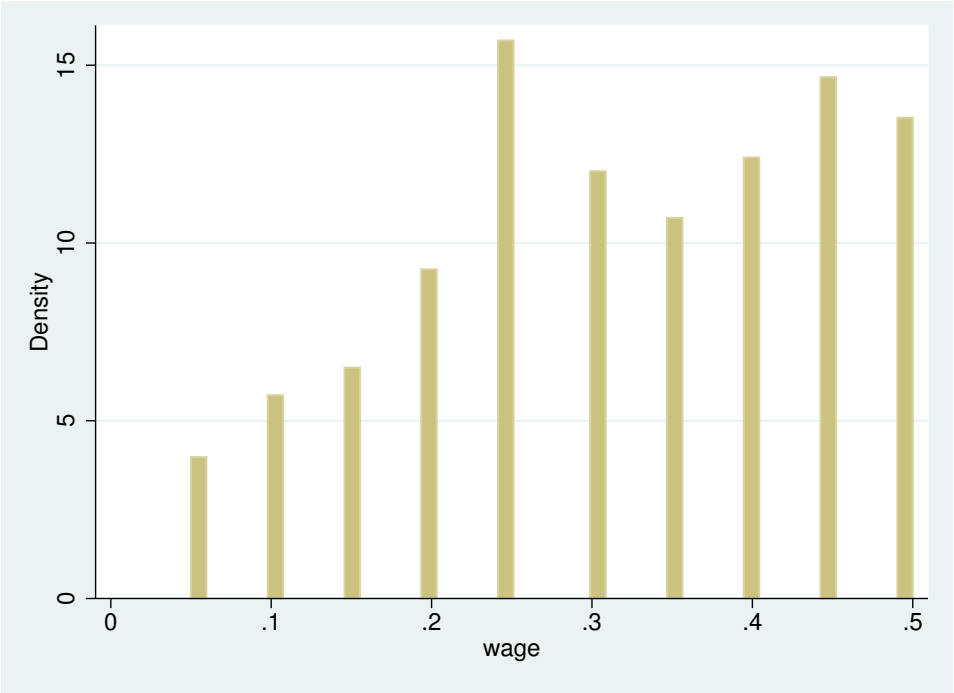


Figure 13: Histogram of Hourly Wages for submitted HITs

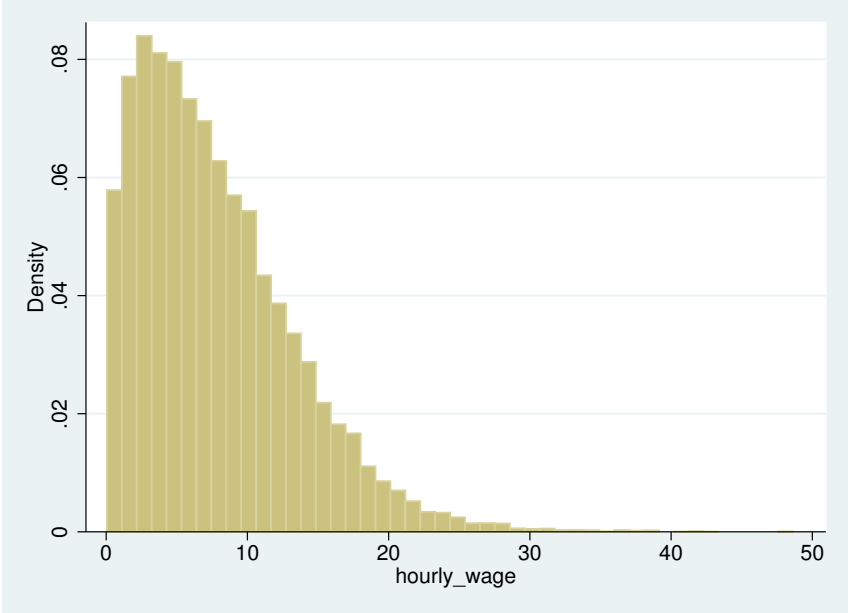
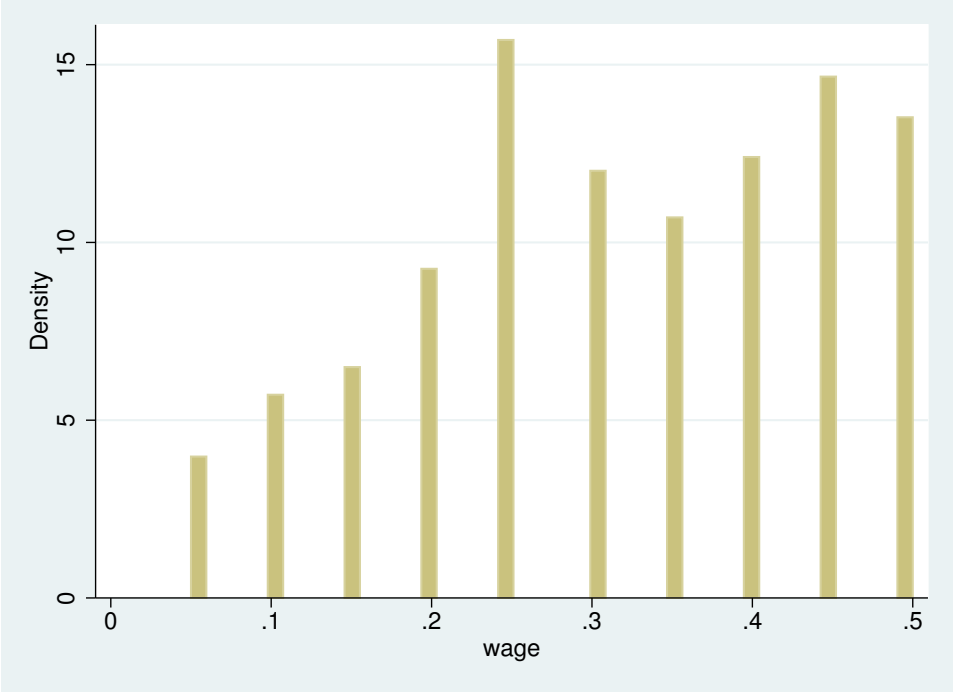


Figure 14: Histogram of Min Accepted Wages for all HITs



B Censored Regression Models vs non-Censored ones

As the Figures in Section 4 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:

$$wage = \alpha + \beta_1 * I(day == 1) + \beta_2 * I(day == 2) + \cdots + \epsilon \tag{16}$$

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 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.

Raymond P Guiteras and B Kelsey Jack. Incentives, selection and productivity in labor markets: Evidence from rural malawi. Technical report, National Bureau of Economic Research, 2014.

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.

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`.