Essays on Labor Supply

Nail Hassairi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Yoram Barzel, Chair

Claus Pörtner

Quan Wen

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Essays on Labor Supply

Nail Hassairi

Chair of the Supervisory Committee:
Professor Yoram Barzel
Economics

This dissertation consists of essays studying labor supply empirically through the use of an experimental platform set up on Amazon Mechanical Turk online labor market. The essays contribute to measuring labor supply elasticties; response of effort to pay; and the willingness to pay for job attributes. What these topics have in common is that a) it's extremely hard to estimate these parameters using observational data, b) the parameters estimated from these studies serve as input parameters for macroeconomic models of employment, welfare policy, and tax policy.

The notion that increased pay raises productivity, either through sorting or incentives, lies at the heart of the efficiency wage theory. The alternative competitive model also implies a correlation between wages and productivity, however, the causal effect implied is from productivity to pay, that is more productive workers receive higher pay. A conclusive test with the ability to nest both of these hypotheses and rule between them then has to provide a credible evidence of the direction of causality in the wage-productivity relationship. This challenge has so far been unmet in the literature. To provide an unequivocal answer, the first essay in this volume presents findings from a large-scale field experiment conducted on Amazon Mechanical Turk. The findings confirm the intuition behind the efficiency wage theory, demonstrating that increase in pay indeed has a causal effect on productivity; through sorting, and, to a lesser extent, incentives. These findings also suggest that this causal

relationship between pay and productivity is dynamic in nature, weakening over the course of the tenure of a worker. The findings also indicate that heterogeneity in tenure length among workers is correlated with their heterogeneity in response to higher pay. As the competitive model continues to be used as a microeconomic foundation of macroeconomic theories of unemployment, these findings give a vote of confidence to its main alternative – the efficiency wage theory.

An integral part of Adam Smith's compensating wage differentials theory is that workers trade off between job characteristics and wage. Other than risk of death, however, no job characteristics have consistently been found to affect wages, likely because of problems with self-selection and unobservable job characteristics. The second essay in this volume presents results from the experiments on Mechanical Turk, randomizing offered pay and job characteristics, thereby overcoming both problems. The findings indicate, as predicted by the theoretical model, that increasing job disamenities significantly reduces both likelihood of working and amount of work supplied. Correspondingly, the wage increases necessary to compensate workers for worse job disamenities are substantial.

The last essay in this volume estimates extensive and intensive margin labor supply elasticity. Contrary to prior analyses using micro data, the findings indicate that the intensive margin elasticities are more than twice the size of extensive margin elasticities and that both are substantial, even if conditioning on working. Furthermore, using data on all workers in the experiments whether they decide to work on the experiments or not, the elasticities range from 1.2 to 2.9, depending on experiment and specification. These results are consistent with the idea that off-line labor markets are characterized by frictions that lower elasticities and may reverse the ordering of extensive margin and intensive margin elasticities.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

HIT: Human Interaction Task, a unit of work on Mechanical Turk, a single submission of work

MT: Mechanical Turk

REQUESTER: an employer, a party that posts a job on the Mechanical Turk and requests the help of the workers to complete this job

ACKNOWLEDGMENTS

I would like to express my gratitude to Yoram Barzel, who with a keen eye was able to spot any loose end in the dissertation project, challenging me to re-think my argument and making sure I understand what I am trying to say and help me think through how I want to lay down the argument.

I would like to thank Claus Pörtner for the trust, support, and friendship. In the time I spent with Claus, I have had the opportunity to observe an efficient approach to conducting empirical research, and the daily grind of research work in general.

I would like to thank Rachel Heath for excellent comments on my earlier drafts.

I would like to thank my classmates Jocelyn Yunling Wang, Austin Gross, Javier Pereira, Bijetri Bose, Guillermo Gallacher, Luke Chen, Mahama Bandaogo for being great friends and sources of inspiration.

I would like to thank Gail Joseph and Janet Soderberg from CQEL UW for a wonderful job experience working on early childhood education research.

DEDICATION

to my mom

Chapter 1

EFFORT, WAGE PREMIA, AND RESERVATION WAGES: A FIELD TEST OF THE EFFICIENCY WAGE THEORY

1.1 Introduction

Efficiency wage theories were designed to explain why labor markets do not clear, a central sore spot of the competitive model. The competitive model assumed perfect information on the part of all agents. The efficiency wage model relaxed this assumption, positing information asymmetry between the firm and the worker and mechanisms that sprung up to remedy this asymmetry, pushing the wage rate above the equilibrium level as a result.

The proponents of the competitive model, following the introduction of the efficiency wage theory, have introduced an updated competitive model which relaxes the restrictive perfect information assumption, however, still maintained that all agents possessed the same limited amount of information ([189]). The emphasis was on the frictions resulting from incomplete information rather than information asymmetry.

The search model, facilitated by the increased availability of panel data on employment-unemployment flows, has grown into a general framework (see [215] for a recent survey) that is able to accommodate any microeconomic theory of unemployment – the implicit contract model ([165]), insider-outsider theory, union bargaining, decentralized Nash bargaining/competitive model ([106]), and the efficiency wage theory (see [231] for the seminal contribution and [177] for its reinterpretation within the search model)[1].

The earlier rivalry between these models in their bid to explain unemployment is now taking place within the confines of various versions of the search model where they are employed to provide an explanation of the wage setting process. Union participation is quite

[1]See [163] for a comprehensive survey of the unemployment literature.

low in the US, so I will omit it from further discussion. Among the wage setting models, the efficiency wage theory is the only theory that implies there is a causal impact of pay on productivity. The main alternative to the efficiency wage theory, the competitive model, re-branded as Nash bargaining model of wage setting, implies the causality runs the other way – more productive worker will receive a higher pay. These explanations are not mutually exclusive. The efficiency wage theory does not contest the idea that more productive workers receive higher pay. Rather, it complements this competitive story by adding that workers receive wage above their competitive levels to motivate them to exert effort. Finding that there is a causal effect from human capital to wages would not disprove neither theory. However, finding that wages causally effect productivity would disprove the competitive theory. This is the bigger part of the agenda of this paper. In this paper, an empirical test will be performed on the question of whether pay has a causal effect on productivity. The null hypothesis of no effect will be in line with the competitive/Nash bargaining model, while rejection of the null hypothesis would be an evidence in favor of the efficiency wage theory.

The efficiency wage theory was first presented as a moral hazard problem, with a principal (employer) being unable to observe shirking ([231], [35]). The employer would then offer wage premia to eliminate the marginal worker's indifference between her job and alternative opportunities. These wage premia put their wage above their reservation wage (outside options, competing job offers) and motivates them to perform adequately to avoid losing the job. Partial monitoring complements the wage premia to ensure that workers do not shirk in equilibrium. Eventually, all firms offer wage premia to compete for workers, resulting in a job rationing and queues. The resulting job queues provide an additional incentive for workers to perform adequantely to keep their job. The carrot (wage premia) and stick (threat of dismissal into an unemployment pool that does not clear quickly) circle is then complete.

The main aim of this paper is to test the shirking model. In this respect, extant empirical studies ([257], [42], [158], [75]) have failed to provide a clean identification strategy that would rule out alternative explanations for their findings. For instance, [42] uses regional variation in wages and treats it as if it were random, rather than stemming from regional differences

in productivity ([188]), re-introducing the issue of spurious correlation. [75] finds evidence consistent with both efficiency wage theory as well as moral hazard problem without rents. Another finding in [75] (workers increase absenteeism after begin granted a contract with more job security) is also consistent with increased absenteeism being seen as a "perk" (job attribute) of a job with tenure and a tournament model for job contracts with more security (for example partners in law firms). The identification strategy in [75] also relies on hand-picking the premium-paying firms and non-premium paying controls, allowing for spurious correlation between unobserved worker/firm productivity and wages – hypothesis in line with the researchers' own admission that the premium-paying firms attract better educated workforce.

The field experiment conducted by [104] could possibly suffer from a Hawthorne effect – change in workers' behavior due to their being aware of being subjects in the experiment and trying to conform to some social norms – and for testing samples not very well-known (or economically important) population of workers (Malawi). 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 the fact that there was only one type of task in the experimental design makes it hard to assess the external validity of the experiment, especially given the suspected mismatch of workers' skill and the task chosen, and given the odd results regarding the effect of selection on productivity. For a more comprehensive overview of these issues, see [204].

Immediately after the shirking model was introduced, [43] criticized it, suggesting there are alternative mechanisms through which shirking could be eliminated – 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. [232] 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 ([258]). 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 whether the ideas of Shapiro, Stiglitz, and Weiss – henceforth SSW – hold together within a joint model. This is the second aim of this paper.

[42], [104], and [75] attempt to test the validity of this joint model. In addition to the shirking model identification issues mentioned above, these papers leave much to be desired in their treatment of testing the selection model. [104] use an imperfect labor market that does not allow workers to self-sort based on their skills, resulting in workers skilled at task A accepting employment in task B they lack skill for, while still requiring higher reservation wages. This results in their finding of negative relationship between reservation wages and productivity. [42] find that in their study 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. [75] studied screening in a context of Portuguese apprenticeship system and show that screening leads to selection for workers with better observable characteristics (education etc). 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.

The contribution of this paper to the literature above is an improved methodology – a cleaner identification of the effects of interest. Variation in wages is provided experimentally, eliminating concerns of spurious correlation. Experiment takes place in the field and workers are not aware of being experimented on, bolstering the external validity of the findings. This paper is the first to find that higher wages contribute to higher productivity through both sorting and incentives, providing support for [232] in their rebuttal of [43]; showing that sorting in fact contributes more strongly to the positive causal effect of wages on productivity than incentives.

This paper utilizes data from a field experiment on Amazon Mechanical Turk online labor market for freelance labor. Amazon Mechanical Turk is an auspicious environment to conduct this study for the following reasons. It 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. Since job queues are most often found in the non-skilled sector of the economy ([35], [163]), Mechanical Turk is a suitable environment to test the efficiency wage theory[2]. Additionally, 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 [204] (fraternizing with managers etc.) Moreover, there are very few frictions in this market and there is a large and diverse population of workers available. This contrasts with the institutional context in which [104]. 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, even if some participants find a way to game the experiment, this strategy is unlikely to spread across the pool of the experiment participants. Finally, Mechanical Turk also protects workers' anonymity – the employers have no information about the worker whatsoever other than a worker ID. This prevents gender, age, or race discrimination, which in turn leads to more representative population of workers on this platform. Unfortunately, this also makes Mechanical Turk less than ideal platform to test discrimination in the labor market.

The experiment took place over the course of six days. 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

[2]Unemployment rate for semi- and unskilled workers is four to five times higher than that for professional and managerial workers. More than three-quarters of unemployed men are manual workers ([163])

to expectations, being positively correlated with the level of monitoring (experimental treatment), experience with the experimental task, 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 [104]). 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 [104].

I construct a simple static model in which effort increase the probability of being paid for the task and ability lowers the cost of effort (the SSW model). A worker chooses an effort level and, in so doing, reveal her cost of effort as well as ability (which lowers the cost of effort). 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 shows the relative importance of selection and incentives. Selection represents the relative importance of ability in lowering the cost of effort. Incentives operate through the effect of higher effort on the 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.

In contrast to [42], 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. Unlike [42] and [75], I used a revealed preference method to show that workers with higher reservation wages exhibit higher productivity. This actually corresponds to the adverse selection theory from SSW 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 certain characteristic; there are no informational asymmetries and no reason to expect wage premia. [75] 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.

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 can be found already in [236] and later on in [209]. [158] find inter-industry wage differentials that are not explained by regional, demographic and human capital variables. [166] finds that it is in fact profitable for the firm to pay efficiency wages using the PIMS line-of-business data. [212] 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. [7] makes the case, however, that tournament schemes are not sufficient to resolve the moral hazard problem. [210] finds 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. [102] reach the same conclusion using hospital data on supervision and wages. However, as [204] points out, a firm could move not only along the isoquant between monitoring and wages but also from one isoquant to another. [204] 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, [204] questions the feasibility of testing the shirking model by testing the link between monitoring and wages.

The rest of the paper proceeds as follows. Section 1.2 describes the experimental design, the Mechanical Turk labor market, and the structure of the data. Section 1.3 describes my theoretical model, identification strategy, and provides results from the analysis of the experimental data. Section 1.4 concludes.

1.2 Experimental Design

The purpose of the experiment is to show that workers are more productive if they get paid more. Much effort goes into a) separating only one direction of causation, since the opposite claim is also valid (and not studied in this paper); b) making sure workers do not know they are part of an experiment (to avoid Hawthorn effect – see above). This section will elaborate on how the features of the experimental environment and design ensure that this goal is achieved.

1.2.1 The Mechanical Turk Labor Market

Employers can post almost any task 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 responding to surveys. Amazon's Mechanical Turk is the largest and most flexible of the emerging micro-task 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).[3] 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 [34].

Figure 1.1 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 at any time without penalty. 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 (see Figure 1.2 for histogram of wages paid out in this experiment based on recorded data

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

for time spent on HIT).[4] There are generally between 5,000 to 30,000 tasks completed each day [140]. 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 any requester can exclude workers with high rejection rates [127]. Workers are not guaranteed to be paid. If workers feel mistreated by a requester on the platform (non-payment, unfair rejection of a HIT), they can share this experience on one of the 3rd-party forums, so that other workers can avoid such an unfair requester.

The worker demographics has been studied by posting surveys to Mechanical Turk itself [138]. 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.

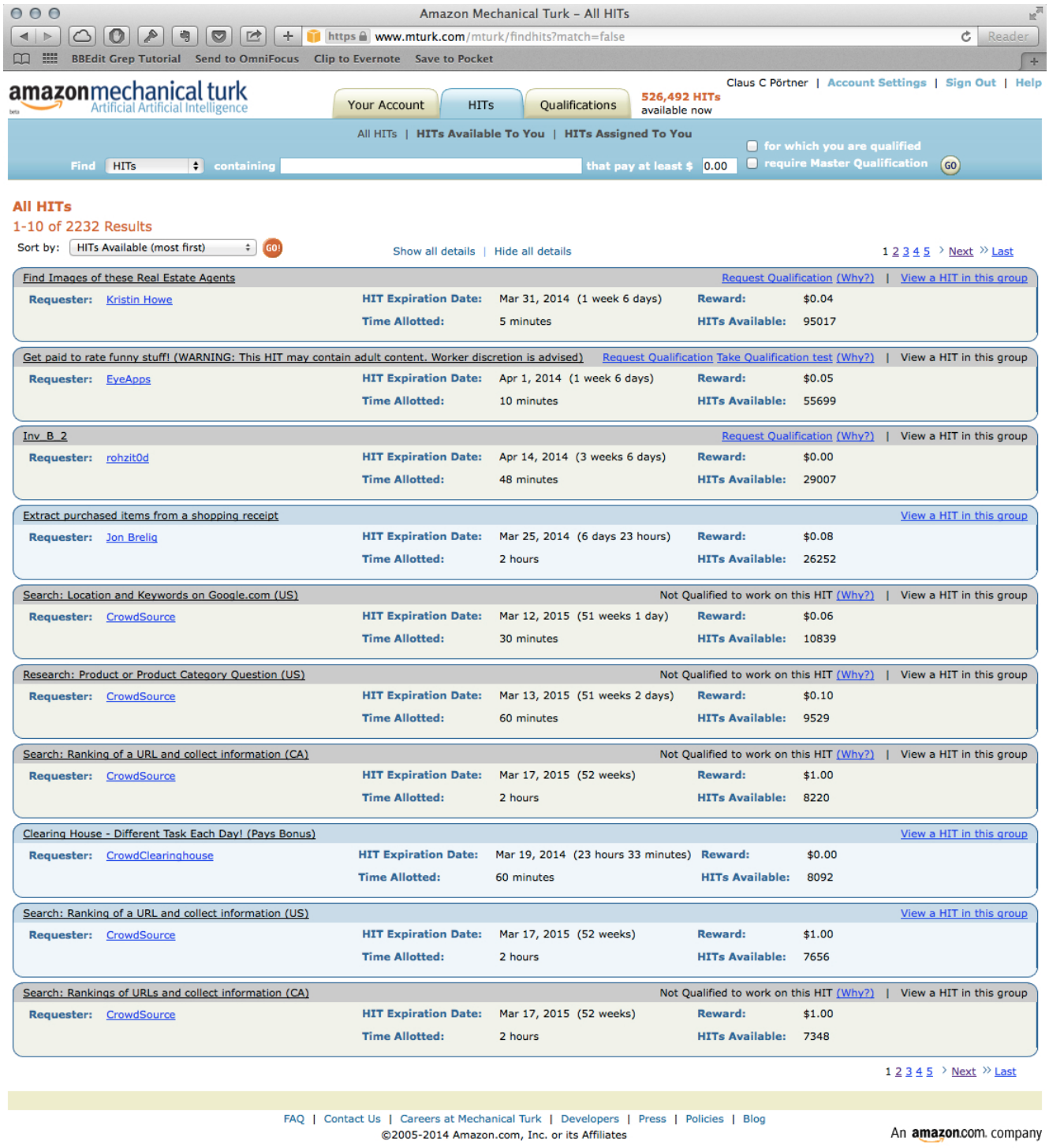
Mechanical Turk is clearly not like "off-line" labor markets. There are no explicit contracts, no set working hours, no commuting, and clothing is entirely optional. Is it, however, similar to the market for freelance or independent contractor work, which rapidly is becoming more and more important in the US economy. A recent estimate is that there are 17.7 million independent workers in the US, making close to $ 1.2 trillion in total income in 2013 and these numbers are been increasing over time [184].[5] 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.

---

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

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

Figure 1.1: Listing of jobs on Mechanical Turk



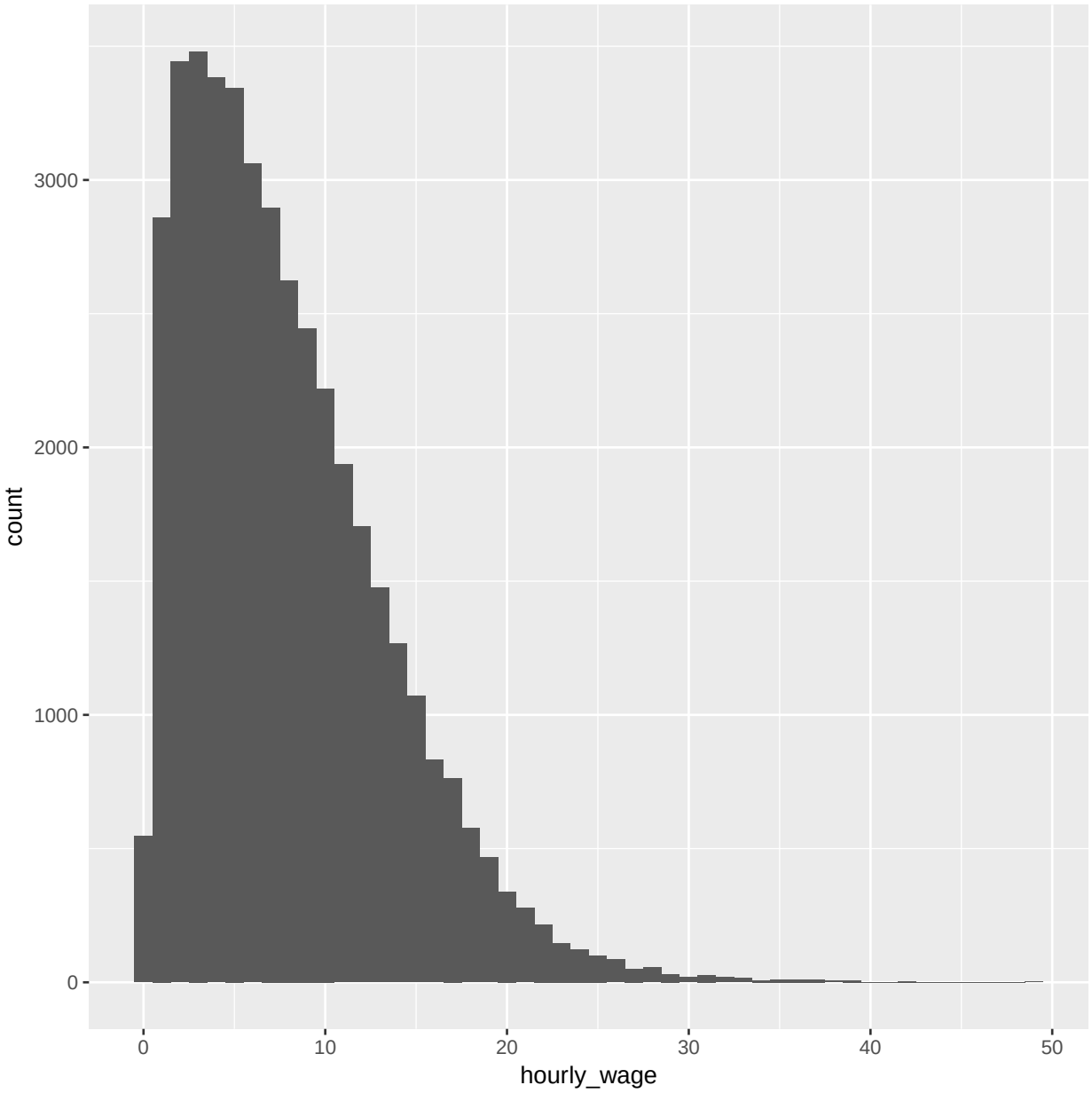


Figure 1.2: Histogram of Hourly Wages (in $)

1.2.2 Image Tagging Job

The data for this paper comes from an image tagging experiment conducted to demonstrate workers' willingness to pay for job attributes ([207]). 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. For the purposes of this paper the experiment is useful in that it provides the main variables relevant to the efficiency wage theory – effort, monitoring, and wages. Additional treatment conditions were part of the experiment given the experiment's original purpose and these other conditions will be described here and controlled for in the analysis.

Once a worker clicks on the job, the experimental software selects and displays five pictures. For each image the worker is asked to provide five tags or keywords, in addition to clicking a radio button indicating whether each of the images is appropriate for a general audience (to prevent disagreeable images from being seen by an audience that could be hurt by seeing them – See Figure 3.1 for a screenshot of the user interface).

The quality/effort/productivity measure collected in the data is based on the latter – the indicator/rating of suitability for sensitive audiences . The number of disagreeable images was varied experimentally. This attribute was central to [207] to elicit valuation of this job attribute. For this paper, its usefulness is in having a baseline judgement of the appropriateness of these images against which to judge workers' rating provided via the radio button. The quality/effort/productivity measure then consists of a count of how many times workers' judgement agreed with the researchers' in terms of rating the appropriateness or inappropriateness of these images. This measure is referred in the analysis variously as "correct appropriatness ratings", "correct_ratings" or simply "ratings" (See Figure 1.3 and Figure 1.4 for histograms of this effort/productivity measure).

There could be anywhere between 0 and 5 inappropriate images in the HIT. In the data, "disagreeableness" is expressed as a percentage rate of the disagreeable images out of the total number of 5 images, resulting in values between 0 and 100 (see Figure 1.5 for a histogram). The number of disagreeable pictures does not change between HITs on a given day, 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

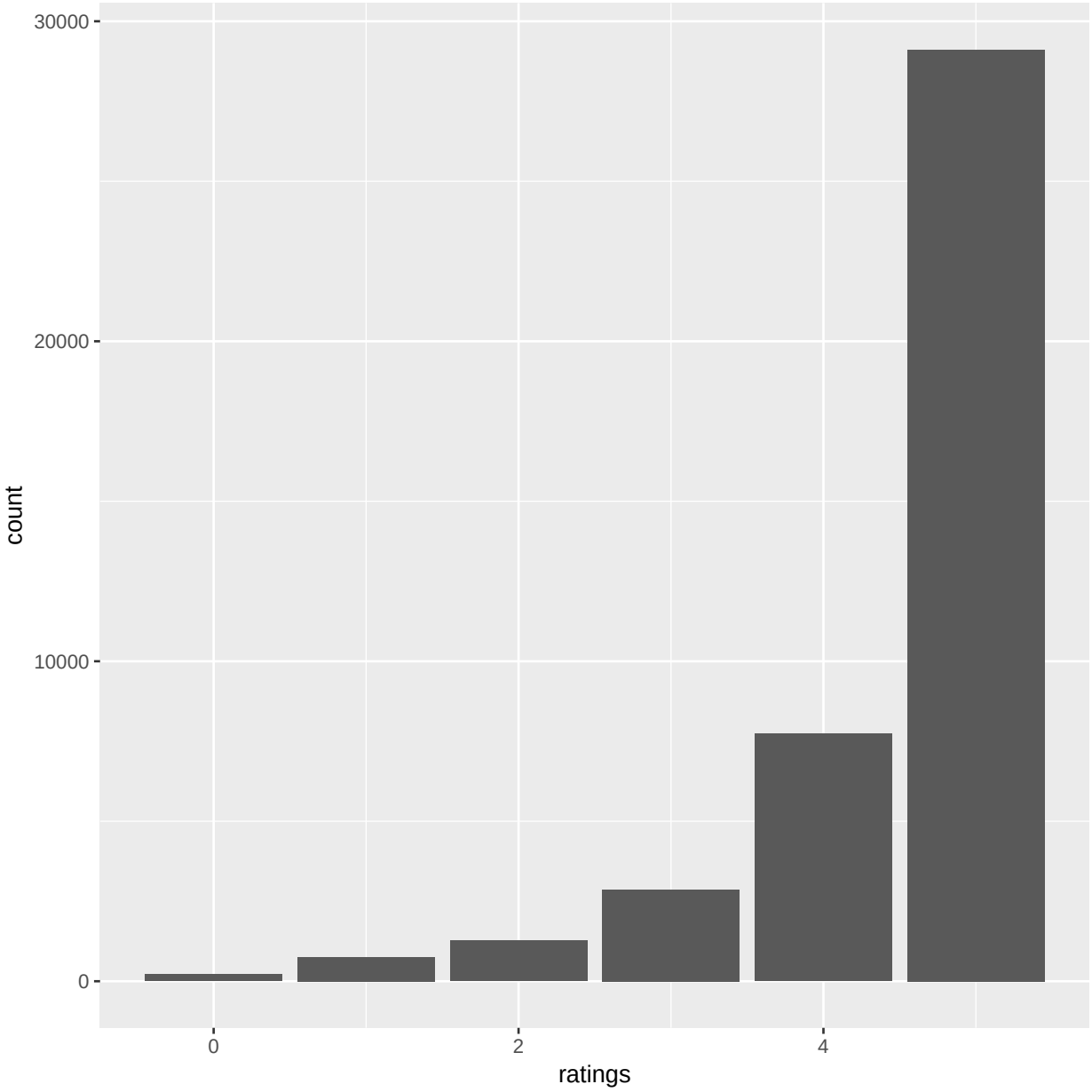


Figure 1.3: Histogram of Correct Ratings

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

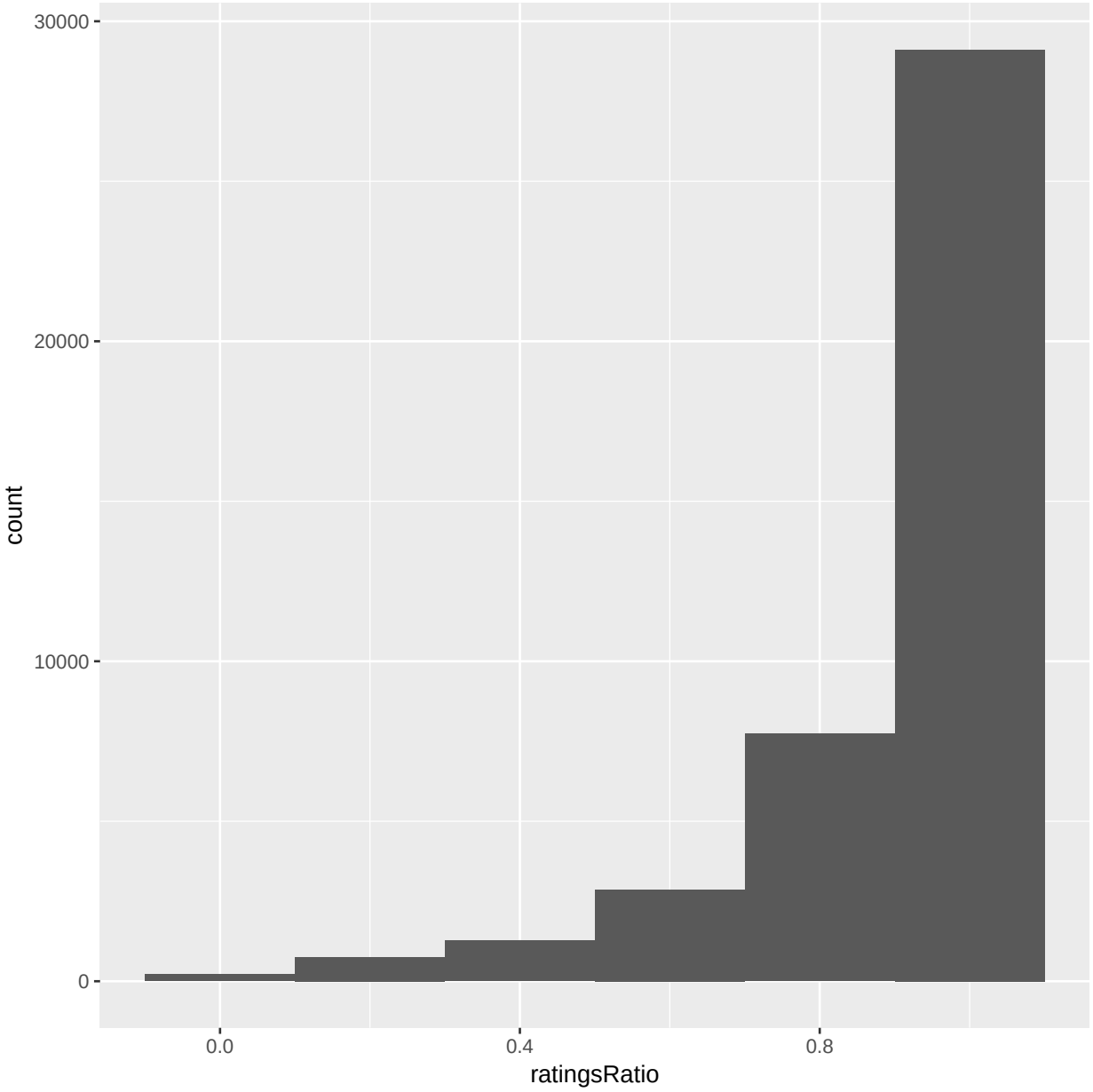


Figure 1.4: Histogram of the Proportion of Correct Ratings

we deleted false positives.[6] This process is, of course, open to cultural biases in what is

[6] 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.

considered disagreeable, but certain responses are more likely biological responses and we aim at those. The conclusions in [207] show that workers were willing to pay substantially to avoid working on images designated by the researchers as inappropriate/disagreeable. 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 random and therefore should only affect the size of the standard errors of the estimates.

There is no standardized way to conceptualize monitoring in the empirical literature on the efficiency wage theory. For example, [102] use the proportion of managerial staff and other workers as a proxy for monitoring intensity. In this paper's experiment the monitoring was conceptualized as "the approval rate." In [207], this was one of the job attributes whose valuation was to be elicited. However, in this paper, it takes on a special significance as a measure of monitoring. The approval/success rate communicated to the workers the bar they have to pass to get paid. Instructions were provided as to how to tag images and rate inappropriateness. Workers were informed that their submissions will be audited and paid only upon being found in agreement with the instructions. The approval rate provided the workers with a running rate, at which submissions were being judged as satisfactory (and paid for) on a given day – giving a sense of how hard they have to try to satisfy the enforced work standards. Figure 1.13 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 (see Figure 1.6 for a histogram of this experimental variable). The approval/success rate was never equal to 100% and never lower than 49% (see Table 1.1 for summary statistics of the variables described in this section). 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. For ethical reasons, all workers were paid for all the work irrespectively of the assigned approval rate. The approval rate then

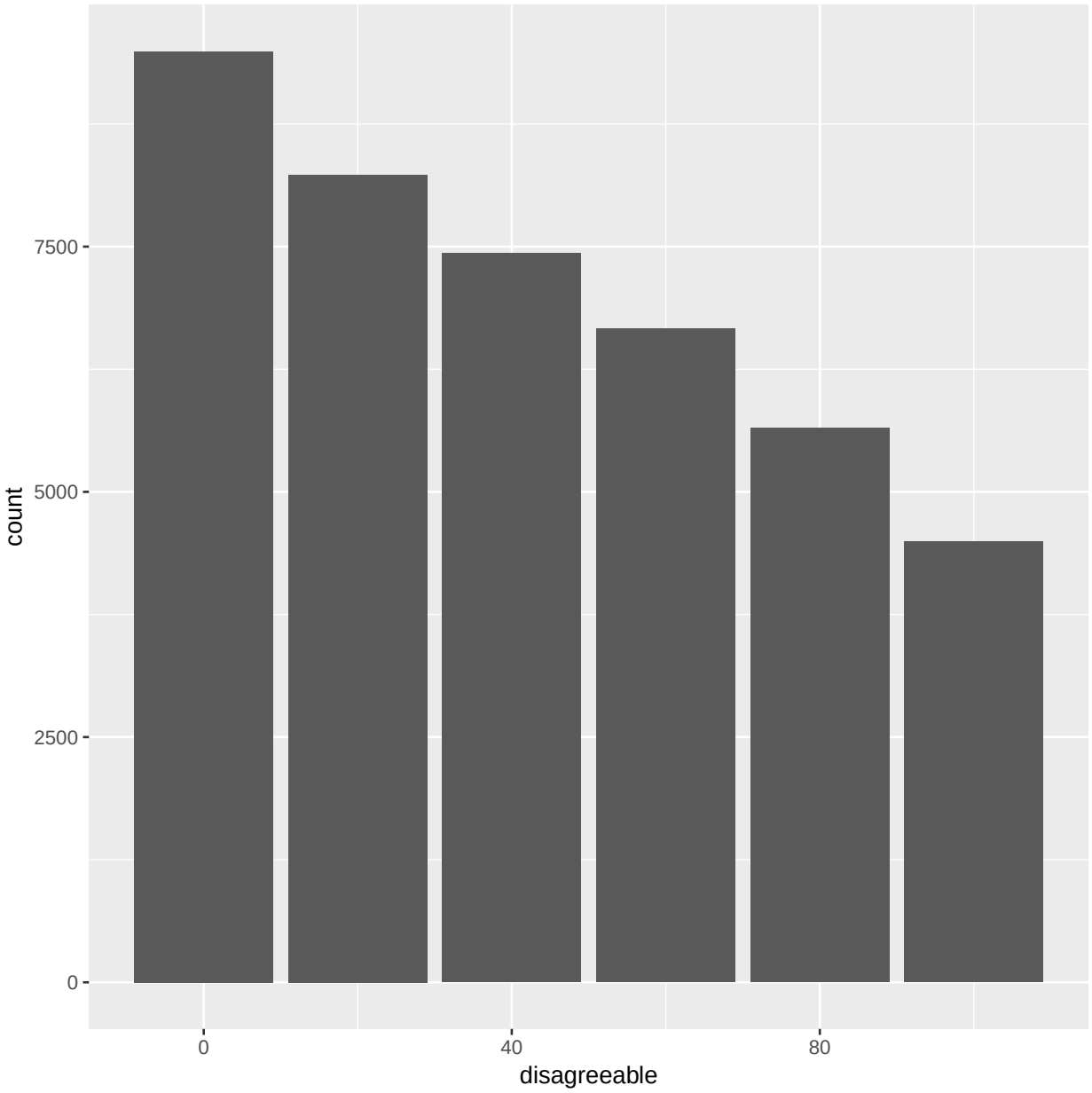


Figure 1.5: Histogram of Percentage of Inappropriate Images in a HIT

represented an empty threat ([207] demonstrates that this threat was taken seriously). This is empty threat might have been responsible for 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.

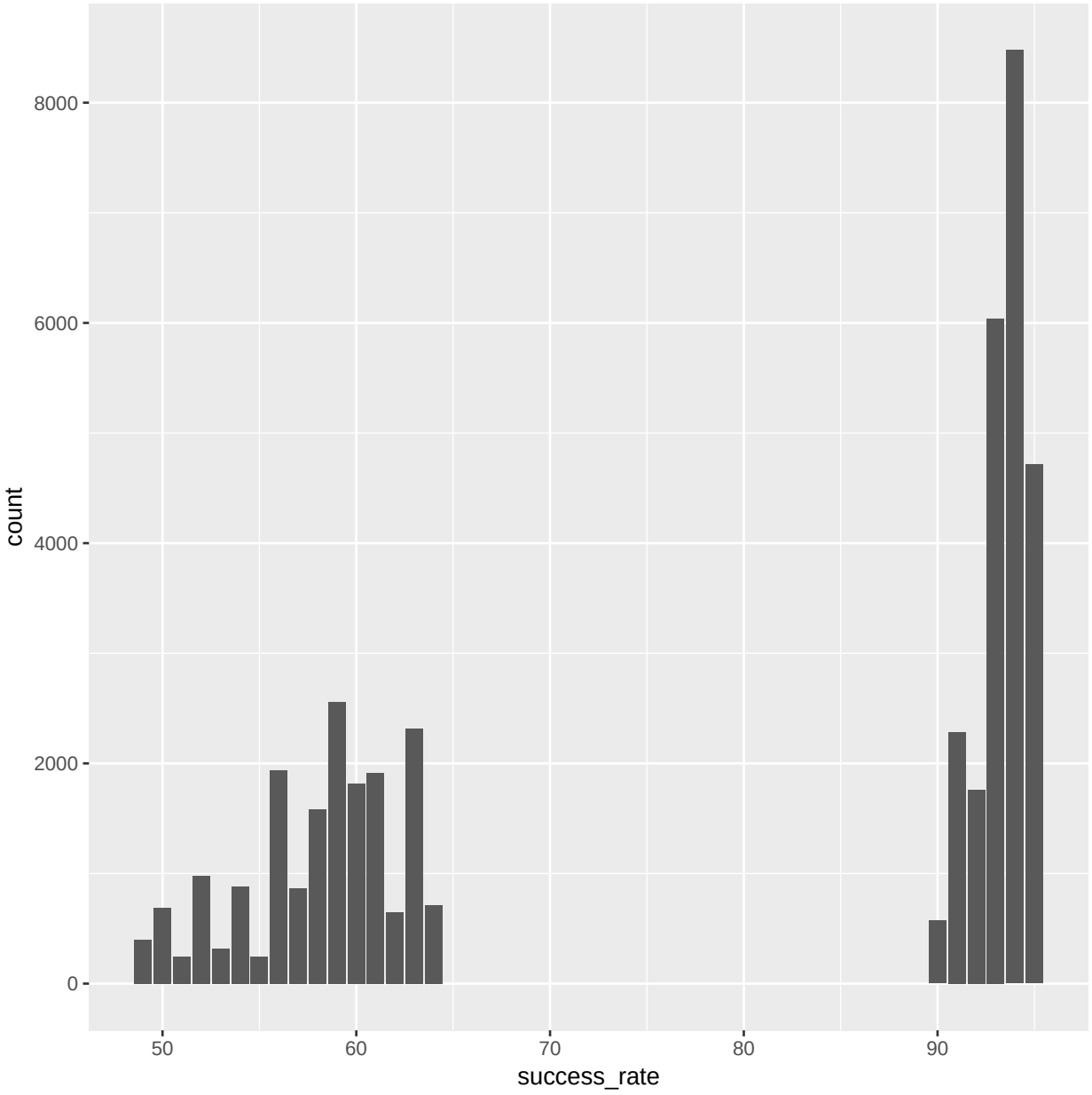


Figure 1.6: Histogram of the Success Rate

The crucial 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 (see Figure 1.7 for a histogram of HIT pay for submitted HITs). Figure 1.13 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 image tagging task 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.[7] Within the job, four job characteristics and the pay offered are randomized. The experiment uses a full factorial design [85]. Experimental conditions are created by systematically varying the levels of each job characteristics 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, **(author?)** 264 and **(author?)** 56). 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. 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.

Once a worker clicks on the job in the list of available jobs, data collection begins. The data is collected at the HIT level. The resulting panel has dimensions i (worker) and t (HIT). t runs from 1 to the last HIT worker submits (T_i – the subscript indicates that workers in the sample have varying tenure). Different workers submit a different amount of HITs, resulting in an unbalanced panel. Figure 1.8 shows a histogram of the amount of HITs submitted by workers over the course of the six day experiment. The number of HITs submitted (tenure) is used in the analysis to control for the effect of unobserved worker-HIT match heterogeneity,

[7] 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.

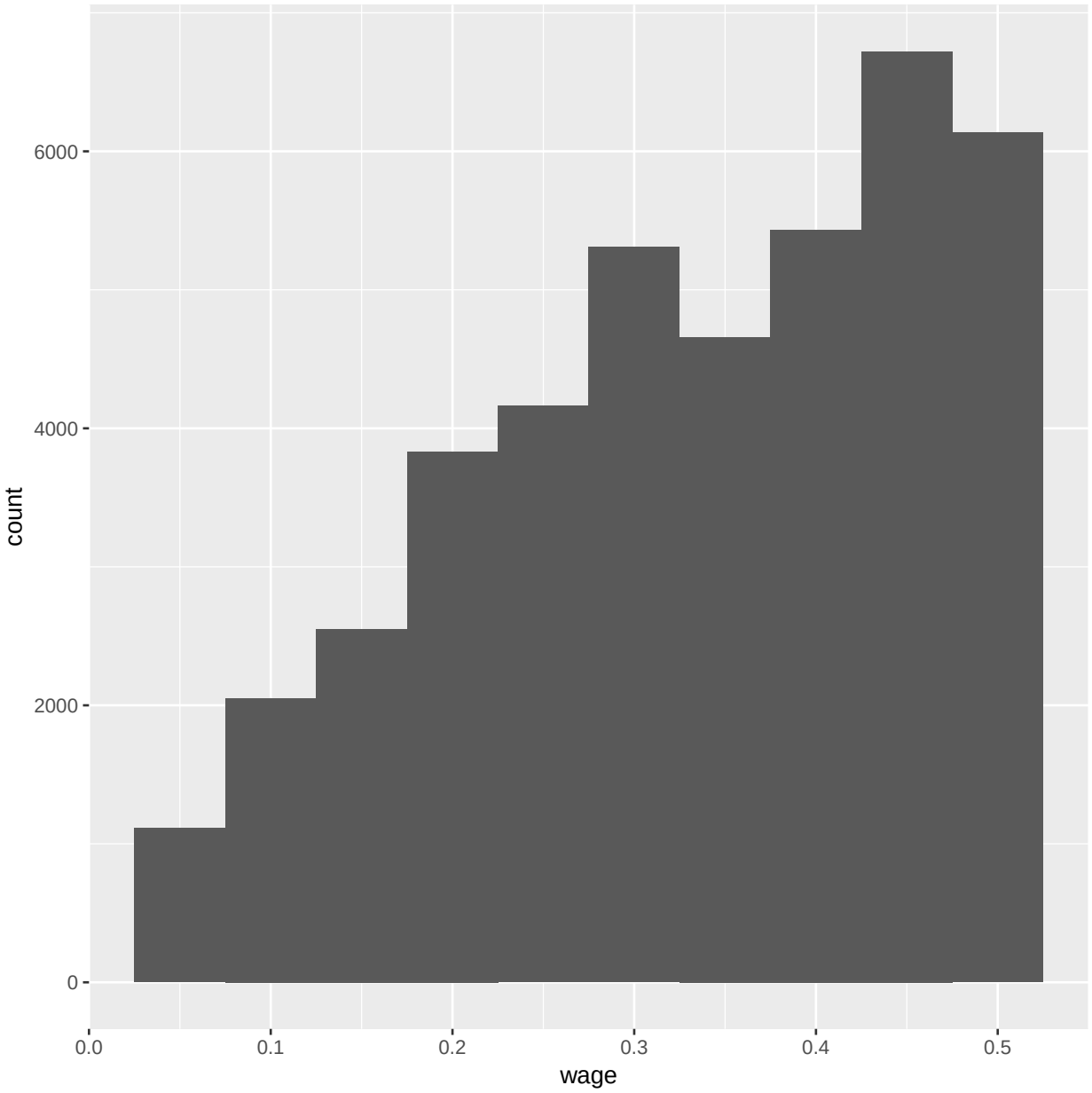


Figure 1.7: Histogram of the HIT Pay

which as one can see from the figure is quite large. While treatments vary on a day-to-day basis, other variables vary on a HIT-by-HIT level – experience with the task/fatigue (t_i), time left before workers stop working ($T_i - t_i$) – and will allow to control for HIT-by-HIT

dynamics that would otherwise obscure results in a worker-day analysis[8].

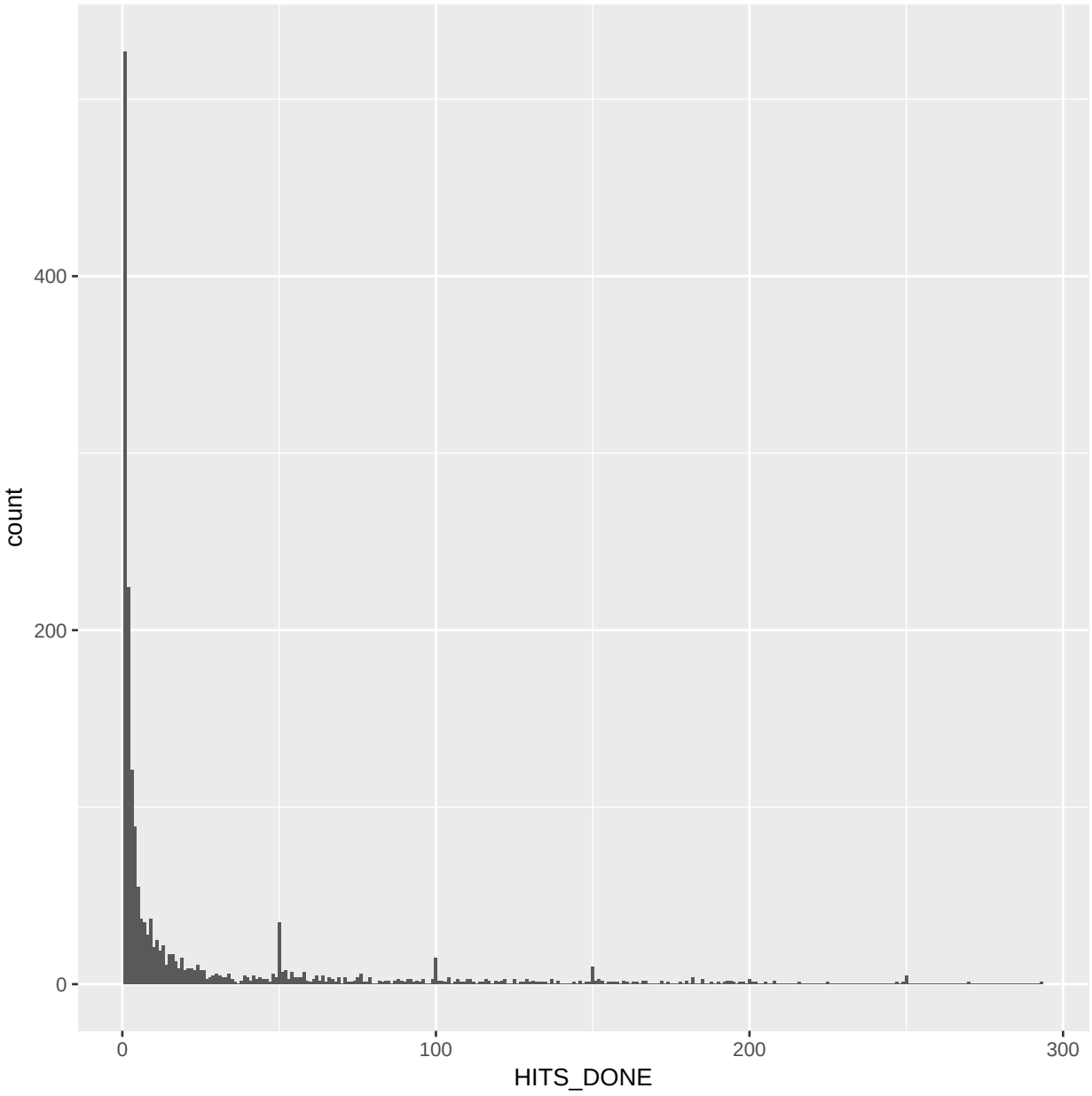


Figure 1.8: Histogram of Total Work Submitted by Workers

[8] The analysis has been performed on worker-day level as well, however, the results are not very illuminating; obscured by un-captured HIT-by-HIT dynamics. Section 1.3 provides more details on the respective results obtained from the worker-HIT and worker-day models.

[258] 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 this experiment, therefore, acceptance wages are decoupled from actual wages to test this aspect of the SSW model. As noted above, wages varied experimentally every day 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 them. To construct the measure of acceptance wage, I have looked at all wages that given worker worked for over the course of the experiment and took the minimum of these accepted wages as 'minimum wage accepted', a proxy for opportunity cost/reservation wage/acceptance wage (see Figure 1.9 for a histogram). This minimum accepted wage will play the screening role described in [258], while the actual wage paid for a given HIT would play the incentive role of wages implied by the shirking model.

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 depend on 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.

The experimenters act as a regular employer on Mechanical Turk. Worker is not informed that the offered jobs are part of an experiment and on a given day is always presented with the same set of circumstances based on their unique worker ID number assigned by Mechanical Turk. Workers were not informed that they were part of an experiment to rule out an observer effect, where workers change behavior to conform to certain perceived social norms in response to being observed (and judged) as part of an experiment. Workers do know that their output is monitored and the level of this monitoring is one of the experimental conditions. The experiments were conducted exclusively through computers ruling out any

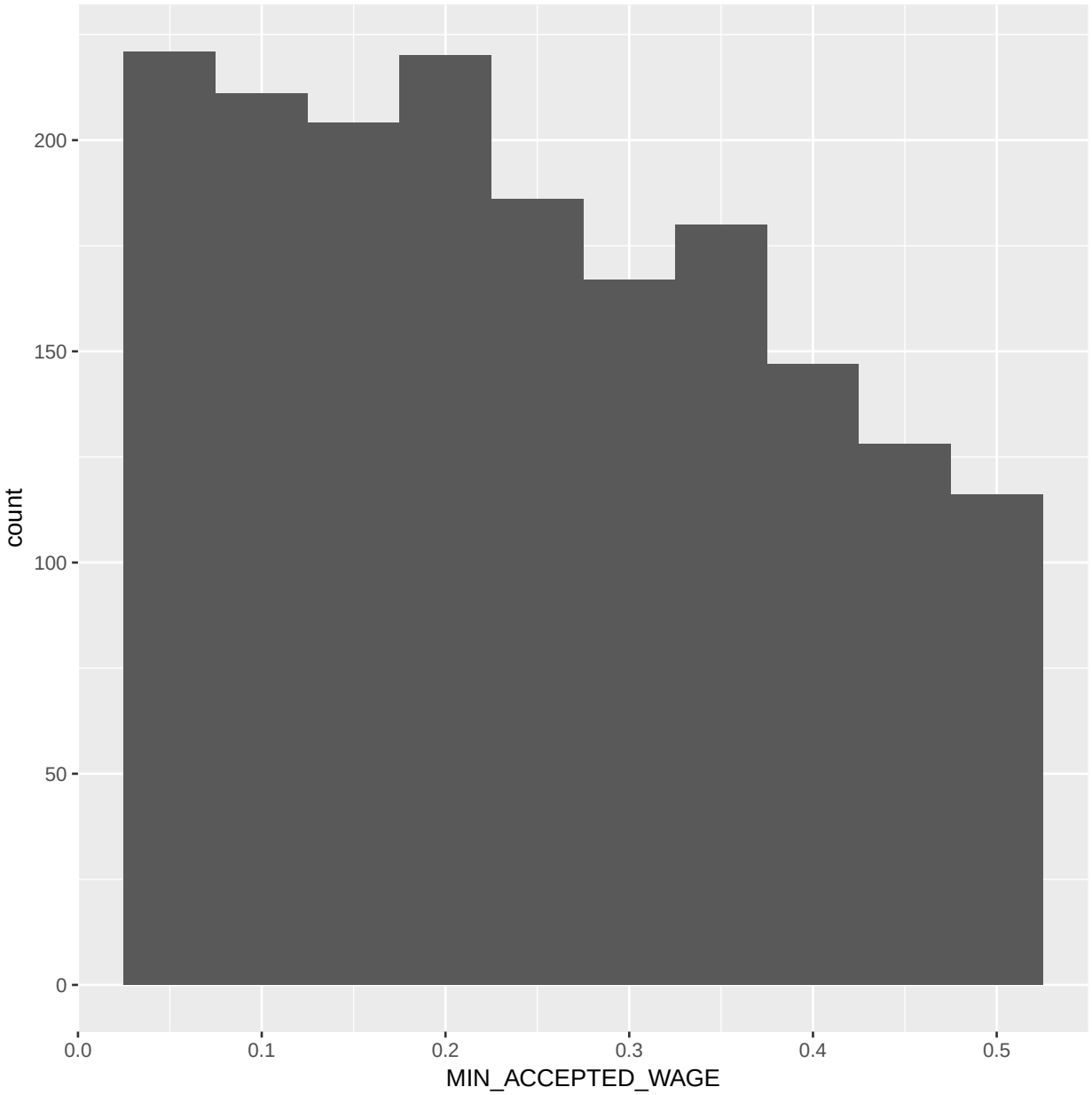


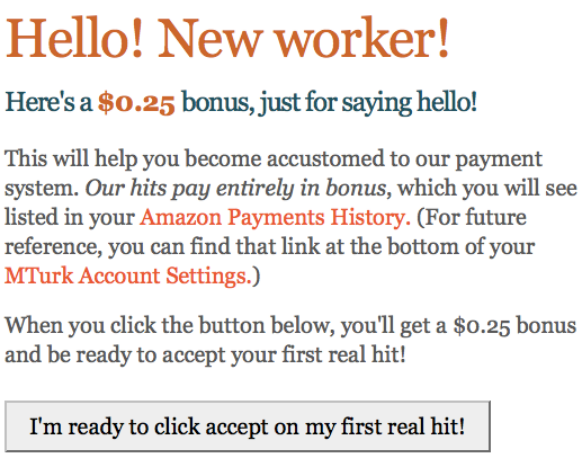
Figure 1.9: Histogram of Worker Min Acceptance Wages

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 1.10. 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. Figure 3.1 shows part of the page presented once a worker accepts the HIT, including one image. 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.

Figure 1.10: Letters to Prisoners Experiment—New Worker "Bonus"



Mechanical Turk allows requestors to require skills and "certifications" of workers. The only requirement for this experiment was that the computer accessing the HITs must be in the US. This allows for estimation of 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.

Cost of learning is another job attribute that featured prominently in [207]. 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 1.12 shows the guidelines and the test questions. This experimental treatment will not feature prominently in the current study; however, a variable controlling for this aspect will be included in the analysis.

Table 1.1: Summary Statistics of the Data

Statistic	N	Mean	St. Dev.	Min	Max
day	41,963	4.317	1.504	1	6
disagreeable	41,963	42.018	33.179	0	100
training	41,963	0.445	0.497	0	1
wage	41,963	0.332	0.127	0.050	0.500
success_rate	41,963	78.198	17.671	49	95
time_on_hit	41,963	253.086	303.469	30	3,617
HITsDone	41,963	53.924	50.859	1	293
hourly_wage	41,963	7.903	5.591	0.056	48.649
min_hourly_wage	41,963	7.903	5.591	0.056	48.649
ratings	41,963	4.490	0.942	0	5
ratingsRatio	41,963	0.898	0.188	0.000	1.000
totalHITs	41,963	105.876	69.973	1	293
HITsLeft	41,963	51.953	50.516	0	292
min_accepted_wage	41,963	0.201	0.116	0.050	0.500
ratingsPerHour	41,963	108.775	68.917	0.000	529.412

1.2.3 External Validity

[173] critically discusses the usefulness of experiments. So far, we have described thoroughly

the process that ensured the results have causal interpretation in the desired direction and that Hawthorne effect can be avoided. What remains to be discussed is how these results apply outside of the Mechanical Turk platform, the original experimental environment – this criterion is termed external validity by [173]. How does the unique experimental environment bear on the external validity of the results. The Mechanical Turk environment is very similar to the model described in [258], 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.

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 ([164]). 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. The 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 1.11: Image Tagging Experiment Page View

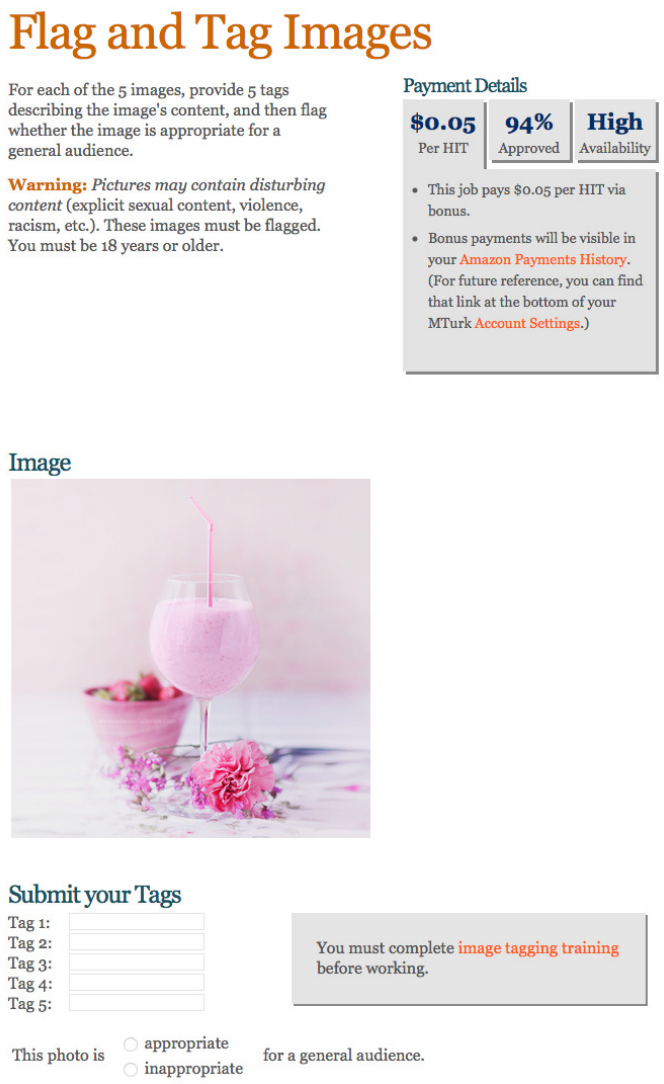


Figure 1.12: Image Tagging Experiment—Training and Test

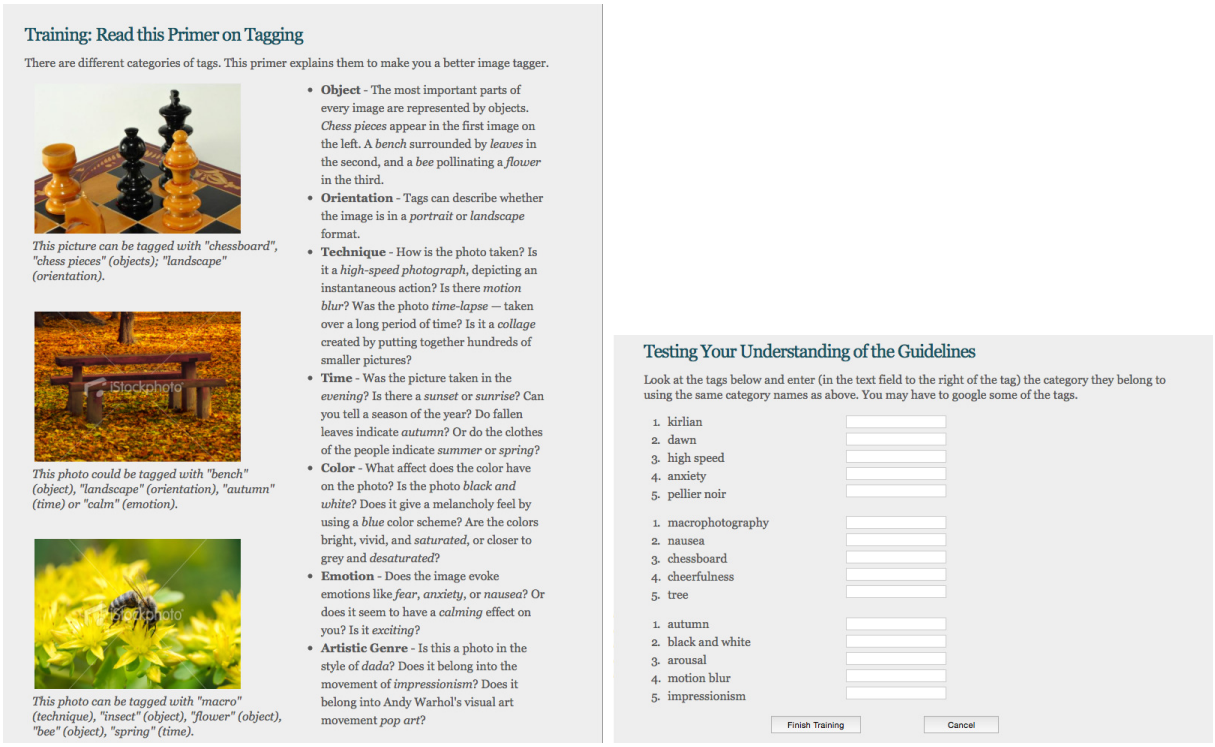
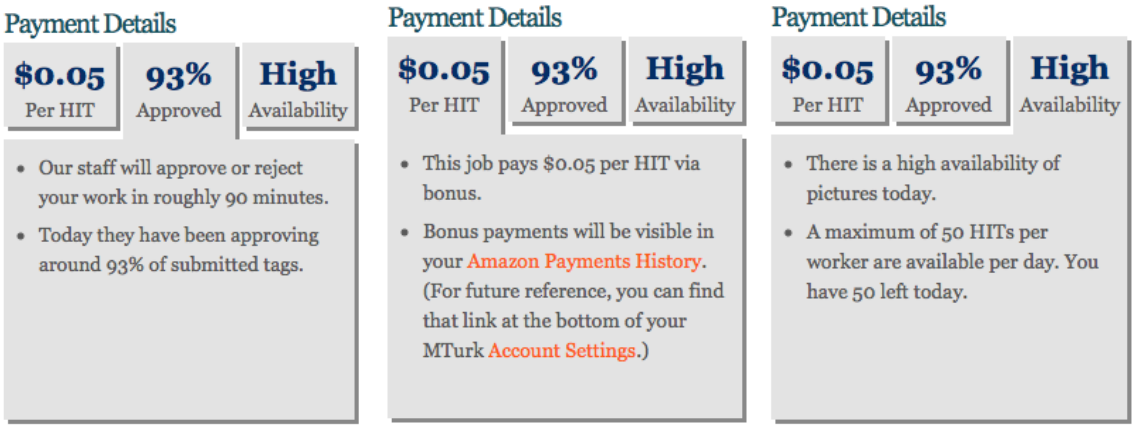


Figure 1.13: Image Tagging Experiment—Approval rate, pay, and availability



1.3 Identification Strategy

1.3.1 Testing the Efficiency Wage Theory against the Alternative Competitive Model

This paper has two main purposes: a) test the shirking model against the competitive model and b) compare the various efficiency wage models against each other. In this section, the first aim will be realized. I construct a simple model, in which principal's (employer's) behavior is taken as exogenous (since it is experimentally varied) and the agent (worker) chooses her optimal effort level given the wage offered, stated level of monitoring, her reservation wage, experience, and preference for given job attributes. Lower effort level will not lead to an increase in the probability of being fired (as is the case in [231]), 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 [258]). The model then combines the aspects of moral hazard from [231] as well as the notion that higher reservation wage correlates with higher unobserved productivity proposed in [258]. 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 distract the exposition from the effort decision. I am assuming that reservation wage of a worker is constant throughout the experiment (or if it is not constant then at least that it is autocorrelated and that the snapshot I use in the analysis is somewhat representative of its time series behavior). Negative job attribute enters the worker's cost of effort function. The worker's payoff function has the following form:

$$u(e) = [1 - P(e, p, n)]U(w) - C(e, \bar{w}, J) \tag{1.1}$$

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 1.1. The first order condition for this problem is:

$$-P'(e,p,n)U(w) - \frac{\partial C}{\partial e}(e,\bar{w},J) = 0 \tag{1.2}$$

While this cannot yield explicit solution for optimal effort without choosing functional 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 the shirking model of the efficiency wage theory (incentives):

$$\frac{\partial e^*}{\partial w} = \frac{-P'(e^*)U'(w)}{P_{ee}(e^*)U(w) + C_{ee}} \tag{1.3}$$

The following assumptions will be made to make conclusions about the signs of the comparative statics effects in this section:

- $C_{e\bar{w}} \geq 0$ (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

In this analysis, the efficiency wage hypothesis is being tested against the competitive model (the alternative hypothesis). The competitive model does not allow for effort to be a choice variable. It is not something the agent wills into existence or has the capacity to change; it is a constant attribute. The agent cannot help themselves but provide equal level of effort under any circumstances. This implies that under the competitive model:

$$\frac{\partial e^*}{\partial w} = 0 \tag{1.4}$$

As can be seen above, one can obtain the results from the competitive model if in equilibrium either $P'(e^*) = 0$ or $U'(w) = 0$. If any of these holds in equilibrium, then the efficiency wage theory is not a relevant model of equilibrium unemployment.

Efficiency wage theory, on the other hand, provides an alternative hypothesis:

$$\frac{\partial e^*}{\partial w} > 0 \tag{1.5}$$

To test the efficiency wage theory against the alternative competitive model, the following regression equation will be used to estimate the size of the above comparative statics effects:

$$correctRatings^* = \beta_0 + \tag{1.6}$$

$$\beta_1 \log(wage) + \tag{1.7}$$

$$\beta_2 log(MinAcceptedWage) + \tag{1.8}$$

$$\beta_3 SuccessRate + \tag{1.9}$$

$$\beta_4 HITsDone + \tag{1.10}$$

$$\beta_5 Disagreeable + \epsilon \tag{1.11}$$

This reduced form equation will map in the following way to the comparative statics result regarding effort and wages:

$$\frac{\partial e^*}{\partial w} = \frac{-P'(e^*)U'(w)}{P_{ee}(e^*)U(w) + C_{ee}} = \beta_1 \tag{1.12}$$

β_1 will be the coefficient of interest in terms of testing the shirking model against the competitive model, all other variables in this specification are controls in this particular context. In the next section, the other variables will come into focus in their own right. $\hat{\beta}_1 = 0$ is the null hypothesis of the competitive model; $\hat{\beta}_1 > 0$ is the alternative efficiency wage hypothesis being tested against the null hypothesis. Table 1.2 shows results from

the test of this hypothesis. The regression was estimated using two models – OLS and censored regression model. The motivation for the censored regression model can be seen from Figure 1.3, where we can see that the outcome is censored at the value of 5. Maximum of 5 images could be rated in terms of appropriateness of for sensitive audiences and the data suggests that the task was easy enough that many workers would have provided even better performance if they had a chance (more images were to be rated within a single HIT). The OLS model is included for comparison. Regressions are run with the outcome as an absolute number of correct ratings, as well as with the outcome as a percentage share of the correct ratings out of total number of images for easier interpretation. Standard error correction is applied at the worker level since shocks are likely to be correlated for HITs submitted by the same worker ([263]). All regressions convincingly reject the null hypothesis of the competitive model – $\hat{\beta}_1$ is estimated to be 0.086 in the censored regression and 0.045 using OLS. The censored model using proportion of correctly rated images as an outcome provides us with the conclusion that a 1% increase in wages leads to 1.7% increase in correctly rated images. This implies that effort is elastic in wages ($\epsilon > 1$).

1.3.2 Sorting versus Incentives

The previous subsection spoke to the baseline validity of the most popular shirking model of the efficiency wage theory. It has been established there that in a labor market that involves routine, low-skill work, workers' effort responds elastically to financial incentives.

Some crude policy implications, such as employment subsidy, affect the labor market in the same way no matter the exact behavior driving the efficiency wage result of the nexus between wages and productivity (nexus running in the opposite direction than would be suggested by the competitive model). Understanding the actual mechanism at work, however, could help design more targeted policies to improve labor markets' efficiency. If there is asymmetry of information exists in the labor market, it may be possible to mitigate it – provide the less informed side of the market with more information. If that were the policy course considered, the question is what kind of information is lacking in the demand

Table 1.2: Regression of Effort on Wages, reservation wages, and tenure

	Dependent variable:			
	ratings		ratingsRatio	
	censored regression	*OLS*	*censored regression*	*OLS*
	(1)	(2)	(3)	(4)
log(wage)	0.086***	0.045***	0.017***	0.009***
	(0.026)	(0.009)	(0.005)	(0.002)
log(min_accepted_wage)	0.195***	0.065***	0.039***	0.013***
	(0.021)	(0.007)	(0.004)	(0.001)
HITsLeft	0.001***	0.001***	0.0002***	0.0001***
	(0.0003)	(0.0001)	(0.0001)	(0.00002)
HITsDone	0.003***	0.001***	0.001***	0.0001***
	(0.0003)	(0.0001)	(0.0001)	(0.00002)
success_rate	−0.003***	−0.001***	−0.001***	−0.0003***
	(0.001)	(0.0002)	(0.0001)	(0.00005)
disagreeable	−0.039***	−0.013***	−0.008***	−0.003***
	(0.0004)	(0.0001)	(0.0001)	(0.00002)
day	0.070***	0.029***	0.014***	0.006***
	(0.010)	(0.003)	(0.002)	(0.001)
training	0.060**	0.018**	0.012**	0.004**
	(0.025)	(0.008)	(0.005)	(0.002)
logSigma	0.674***		−0.935***	
	(0.007)		(0.007)	
Constant	7.939***	5.105***	1.588***	1.021***
	(0.090)	(0.029)	(0.018)	(0.006)
Observations	41,963	41,963	41,963	41,963
R^2		0.216		0.216
Adjusted R^2		0.216		0.216
Residual Std. Error (df = 41954)		0.834		0.167

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

side of the labor market. The two main information asymmetry efficiency wage hypotheses are informational asymmetry regarding agent's behavior (hidden action) and informational asymmetry regarding agent's type (hidden type). The labor market already has institutions designed to combat both types of asymmetries – suggesting their practical importance. To combat hidden action type of information asymmetry, employers have performance reviews, peer reviews, contracts tying pay to performance. However, no study has demonstrated to what extent these systems are effective in their stated goal. And given that there has not been a convincing evidence regarding the effect of pay on performance, these institutions have sprung up independently of rigorous scientific evidence, being motivated rather by conventional wisdom or personal experience. This also means that firms have no data to go by in terms of deciding how much resources to invest in screening and monitoring of their job applicants and employees and that these investments may be at inefficient levels.

To shed light on these two informational asymmetries, the model from the previous section will be used. The model nests both informational asymmetries and allows for the estimation of both of these effects. The first order condition from the agent's optimization problem was:

$$-P'(e,p,n)U(w) - \frac{\partial C}{\partial e}(e,\bar{w},J) = 0 \tag{1.13}$$

In the previous section, comparative statics effect for change in wage offered was derived:

$$\frac{\partial e^*}{\partial w} = \frac{-P'(e^*)U'(w)}{P_{ee}(e^*)U(w) + C_{ee}} \tag{1.14}$$

Similarly, comparative statics effect for the effect of higher reservation wage on optimum effort can be derived:

$$\frac{\partial e^*}{\partial \bar{w}} = \frac{-C_{e\bar{w}}}{P_{ee}(e^*)U(w) + C_{ee}} \tag{1.15}$$

In the like manner, one can derive the comparative statics effect of a reservation wage. The comparative statics effect of an actual wage paid is linked to the incentive effect and the

shirking model of the efficiency wage theory. The comparative statics effect of the reservation wage on effort, on the other hand, is linked to the selection effect. The reservation wage is the only screening device for ability, since no information about workers other than ID is available to the experimenter (author of this paper). While the experimenter did not in fact screen on this reservation wage, the comparative statics effect of reservation wage on performance will show us, whether such screening would increase productivity. Starting from the first order condition of the effort maximization problem (Equation 1.2) I derived Equation 1.15, which shows how much more productive workers with higher reservation wages are. Our regression equation is an explicit equation for effort as a function of wages and reservation wages. Our theoretical model, eschewing making parametric assumptions, allows us to solve only implicitly for effects of wages and reservation wages on equilibrium effort. This allows me to map the comparative static effect from Equation 1.15 on the coefficient β_2 from Equation 1.6:

$$correctRatings^* = \beta_0 + \tag{1.16}$$

$$\beta_1 \log(wage) + \tag{1.17}$$

$$\beta_2 log(MinAcceptedWage) + \tag{1.18}$$

$$\beta_3 SuccessRate + \tag{1.19}$$

$$\beta_4 HITsDone + \tag{1.20}$$

$$\beta_5 Disagreeable + \epsilon \tag{1.21}$$

$$\frac{\partial e^*}{\partial \bar{w}} = \beta_2 \tag{1.22}$$

Let us replicate here the assumptions on worker behavior:

- $C_{e\bar{w}} \geq 0$ (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
- in accordance with the law of diminishing marginal productivity $P_{ee}(e^*) < 0$
- it also commonplace in the literature to assume that $C_{ee} > 0$

It's not possible from this model, and using data from this experiment, to unambiguously sign these effects, however, we can look at the ratio of these two effects – since they share the same denominator and the denominator is what prevents the unambiguous determination of the sign of these affects in equilibrium.

$$\frac{\frac{\partial e^*}{\partial \bar{w}}}{\frac{\partial e^*}{\partial w}} = \frac{C_{e\bar{w}}}{P'(e^*)U'(w)} = \frac{\beta_2}{\beta_1} \tag{1.23}$$

The intuition behind Equation 1.24 is that heterogeneity is only important if $C_{e\bar{w}}$ is different from zero. In other words, for heterogeneity to matter, ability should not only affect levels of disutility of effort, but also the slope of the disutility function as the effort increases. For an individual with less steep effort-disutility relationship, incentives are more effective; individuals with less steep effort-disutility relationship could be thought of as "high types", high-productivity individuals. Selection and incentives are then not necessarily mutually exclusive, it's easier to incentivize a high-productivity individual than a low-productivity one. Equation 1.24 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.

The results from Table 1.2 provide an estimate of this ratio. $\hat{\beta}_1$ is estimated to be .086 and $\hat{\beta}_2$ is 0.195. The ratio of the two then is:

$$\frac{\frac{\partial e^*}{\partial \bar{w}}}{\frac{\partial e^*}{\partial w}} = \frac{C_{e\bar{w}}}{P'(e^*)U'(w)} = \frac{\beta_2}{\beta_1} = \frac{.195}{.086} = 2.23 \tag{1.24}$$

The conclusion of the previous two section then is that a) efficiency wage theory has strong support in the data, b) both shirking and sorting models of the efficiency wage theories have support in the data, and c) in the institutional context where the study took place sorting was in fact the stronger one of the two effects. Some context is in order. The institutional environment of Mechanical Turk presents employers with very little in terms of reducing the informational asymmetries studied in this paper. Some existing labor market institutions in offline labor markets may mitigate both of these effects – interviewing applicants, collecting resumes etc. Our estimates are useful in so far as to say that these informational frictions are real and in the absence of mechanisms to mitigate them they will pose substantial deviations from optimality in the labor market. Mechanical Turk is a good example of unfettered free market for labor. Our study provides an indication of welfare implications of having unregulated institution-free market for labor.

1.4 Conclusion

This paper represents a culmination of efforts to evaluate a central notion of the efficiency wage theory – that pay causally affects effort, and consequently productivity. The contribution of this paper is in clean experimental design coupled with real-world field-experiment environment and a large sample size. The evidence is in support of the efficiency wage theory, showing that effort is elastic in pay. This finding justifies continued effort in developing unemployment models based around the efficiency wage hypothesis. In addition to this main finding, the microeconomic (nano-economic even) nature of the dataset used in this paper allowed for the examination of mechanisms driving the efficiency wage nexus between pay and productivity. Two main models of the efficiency wage theory were examined – the heterogeneous agent model and the shirking model. Evidence strongly supports the contention that both of these models have a strong role to play in the labor market, supporting early theories suggesting that confluence of various efficiency wage mechanisms are contributing to the persistence of higher-than-equilibrium wages. As a result of this confluence, hypothetical labor market institutions (such as performance bonds) are insufficient to eliminate the wage premia in the labor market.

By reaffirming the central tenet of the efficiency wage theory, this paper can contribute to the revival of the efficiency wage theory, which has most recently been merged with the most recent search theory of labor market behavior. I believe this a fruitful direction of research, since it combines the search theory that incorporates most recent empirical findings regarding job-switching and job-searching behavior with the most popular hypothesis that gives rises to real wage rigidity and involuntary unemployment. While the Nash bargaining models of search theory have relaxed somewhat the assumption of perfect information by allowing for incomplete information, such models still lead to the conclusion that allow unemployment is voluntary, which has no basis in facts.

Chapter 2

ONLY IF YOU PAY ME MORE: USING FIELD EXPERIMENTS TO DERIVE WILLINGNESS-TO-PAY FOR JOB CHARACTERISTICS

2.1 Introduction

> The five following are the principal circumstances which, so far as I have been able to observe, make up for a small pecuniary gain in some employments, and counter-balance a great one in others: first, the agreeableness or disagreeableness of the employments themselves; secondly, the easiness and cheapness, or the difficulty and expence of learning them; thirdly, the constancy or inconstancy of employment in them; fourthly, the small or great trust which must be reposed in those who exercise them; and, fifthly, the probability or improbability of success in them.
>
> Adam **(author?)** [236, Book I, Chapter X, Part I]

That workers treat job characteristics as consumption goods and trade off between wage and job amenities is one of the central tenets of labor economics, the value of statistical life (VSL) literature, and urban economics [220, 256, 214]. If correct, it allows us to draw inferences about preferences and technology from wage data and helps us understand wage structures in the economy. These, in turn, affect policy in areas as diverse as highway speed limits and taxation.[1] However, attempts to estimate the "price" associated with different job characteristics often fail; the only job characteristic consistently found to affect wages is

[1] [13] discusses how speed limits are set. [203] analyze the effect of taxation on distortions of the wage-job amenity trade-off.

risk of death [220].[2]

The basic econometric problem is that workers self-select into specific jobs based on unobservable worker or job characteristics [32]. Worker characteristics can broadly be divided into productivity and preferences. A substantial portion of the literature argues that unobserved productivity differences are the main reason for the lack of observed effects of job characteristics on wage in cross-sectional data, and examines various ways of accounting for unobserved productivity, although with varying success (see, for example, **(author?)** 32, **(author?)** 67, **(author?)** 136, **(author?)** 154, and **(author?)** 29).

To further complicate matters, we have little idea how these unobserved worker characteristics are distributed across workers. This leads to two problems. First, because of the self-selection into specific job it makes it difficult to establish whether workers do, indeed, trade off between wages and job amenities, despite this idea's substantial common sense appeal. Second, since public policy is currently based on marginal worker estimates, it is of interest to understand how far away that worker's willingness to pay is from the average. These values can be far apart if there are substantial unobserved differences in productivity, or there are groups of workers who have substantially different preferences from the rest of the population and only a small number of jobs that have a set of characteristics that match those preferences. On one hand, if workers are clustered over a relatively short range of a specific characteristics, then a (unbiased) estimate of willingness to pay is likely close to the average willingness to pay. On the other hand, if the distribution is asymmetrical or very wide, it become substantially more difficult to arrive at a solid estimate of the willingness to pay for a specific job characteristics.

The purpose of this paper is twofold. First, to show that workers exhibit substantial willingness to pay for job characteristics. Second, to provide evidence on the distribution of workers' willingness to pay [note: this part is not developed yet].

We take a different approach from the previous literature. Instead of trying to infer

[2] The literature is too large to fully review here. See [153] for a recent review.

trade-offs between job characteristics and wage using observed wages, we run experiments that allow us to directly examine trade-offs using estimated labor supply functions. We offer jobs, randomly allocating arriving workers to different combinations of job characteristics and pay within each job, and observe workers' decision on whether to work or not and amount of work supplied. We show that worker behavior supports the labor supply version of the compensating wage differentials theory and that workers exhibit substantial willingness to pay for job characteristics.

This work is made possible by the emergence of online labor markets for micro-tasks. We use Amazon's Mechanical Turk (www.mturk.com), which allows us to control all aspects of the jobs offered, such as job type, job characteristics, and pay. We offer two separate jobs at different points in time: One asks workers to tag images with keywords and the other asks them to write letters. Each job requires different skills and appeals to workers with different interests, thereby providing more general validity to our results. We vary four job characteristics that broadly correspond to four of the "principal circumstances" set out by Adam Smith: agreeableness of the task, cost of learning, availability, and probability of success.[3] In each job/experiment, we randomize the level of each job characteristic and the pay offered. For example, for agreeableness we randomly assign workers who look at our offered work to either an "agreeable" version of the task or a more "disagreeable" version of the same tasks.

Mechanical Turk has three major advantages when we want to understand the trade-off between job characteristics and wage. First and foremost, conditional on workers looking at our offered jobs, self-selection is not an issue. In fact, the beauty is that we can follow the sorting process. We observe whether a worker accepts or rejects a job offer, and the effort

---

[3] Adam Smith's idea of the amount of trust required corresponds closely to the current idea of efficiency wage in modern labor markets [233]. The analyses required to test this differ substantially from the other four circumstances and we therefore plan to do that as a separate paper.

Some examples of prior research or surveys that have examined job characteristics broadly consistent with each "circumstance" are for agreeableness: [32], [67], [95], and [157]. For cost of learning: [216], [181], [260], and [20]. For availability: [3], [170], [109], [187], and [15]. Finally, for probability of success: [152], [218], and [116]

supplied if the job is accepted. The randomization of pay and job characteristics ensures that both are orthogonal to worker characteristics and preferences. This allows us to recover workers' willingness to pay for individual job characteristics, and thereby understand whether workers behave as predicted by the theory.[4]

Second, there is substantially less scope for measurement error than in prior studies. Part of the previous literature relied on self-reported job characteristics, which are prone to reporting errors because workers with different preferences likely report identical job characteristics differently [32, 67, 73]. Even when job characteristics are not self-reported, measurement errors occur because some industry specific job characteristics may not be relevant for all workers in that industry [256, 155]. We know exactly what conditions workers were exposed to because we control all aspects of the offered job.

Finally, we avoid the econometric problems associated with estimating hedonic models [217, 21, 22, 72]. In regular labor markets, observed wages may change—despite no change in worker preferences or productivity—because firms' cost of providing a set of job characteristics change [70, 256, 13]. We do not have to worry about the demand side of the job market because we control it and all workers in each experiment see the same basic job.

We expand the standard labor supply theory with disutility of working to also allow for disutility of job disamenities. We show that the likelihood of working and the amount of labor supplied, if working, are always decreasing in worse job disamenities. Hence, if workers adjust hours worked, but this is not captured in data, this is an additional reason why identifying the trade-off between job characteristics and wage is difficult in regular wage data.

Our main finding is, as predicted by our model, that increasing job disamenities significantly reduces the likelihood of working and the amount of work supplied for agreeableness, cost of learning, and probability of success.[5] Correspondingly, the wage increases necessary

[4] We cannot, however, fully recover each individual worker's willingness to pay because we do not observe the reservation wages for specific combinations of job characteristics. We are working on experiments that will allow us to do that.

[5] Higher wages lead to both significantly higher probability of working and higher number of tasks

to compensate workers for worse job disamenities are substantial. Depending on experiment and job disamenity, the increases are in the order of 60 to 335% of the average offered wage for the extensive margin and 30 to 190% for the intensive margin. These effects only show up consistently when we control for selection. Using only workers who self-select into the jobs we find mostly no effect of job disamenities, and even when there is an effect, it is substantially lower than when controlling for selection. We further illustrate the effects of selection using, first, information on workers' tenure on Mechanical Turk and, second, longitudinal data from the image tagging experiment. Length of experience does little to change our main results, and selection substantially lowers the estimated compensating wage differentials in longitudinal data.

2.2 Theory

The standard theoretical framework for compensating wage differentials treats job characteristics as a consumption good, and examines the trade-off between market consumption and job characteristics [220]. Rosen's model is appropriate if hours are fixed and there are no unearned income or outside options. On Mechanical Turk, however, workers decide both whether to work on a given job and how much to work. The essence of these decisions can be captured by expanding the standard labor supply theory with disutility of work to also include disutility of job disamenities.

Assume that vector d captures job disamenities, where a higher d corresponds to worse job characteristics. Each worker's preferences are defined over a market consumption good, c, disutility of work, h, or equivalently utility of leisure, l, and disutility of job disamenities. To ease exposition we assume that work and job disamenities do not affect the utility of consumption, so the utility function is

$$U = u(c) + v(l; d), \tag{2.1}$$

---

performed in both experiments. We examine labor supply elasticity estimates in detail in a separate paper [**?**].

where $u_c > 0$, $u_{cc} < 0$, $v_l > 0$, $v_{ll} < 0$. An increase in job disamenities makes a job less attractive, $v_d < 0$. We assume that $v_{ld} > 0$, so that worse job amenities makes leisure more attractive and work less attractive (the marginal utility of leisure—or equivalently the marginal disutility of work—goes up as job amenities become worse).[6]

Workers maximize their utility subject to a budget and a time constraint.

$$c = hw + I \tag{2.2}$$

$$T = l + h, \tag{2.3}$$

where w is wage per hour, I is unearned income, T is total number of hours available, and l is leisure. Substituting in the constraints, so that the maximization problem is expressed in terms of work, and solving leads to the standard first order condition:

$$\frac{v_l}{u_c} = w, \tag{2.4}$$

the ratio of marginal utility of leisure to marginal utility of consumption is equal to the wage. Total differentiating, assuming an interior solution, and rearranging leads to

$$(u_c + hwu_{cc})\mathrm{d}w + (w^2u_{cc} + v_{ll})\mathrm{d}h + wu_{cc}\mathrm{d}I - v_{ld}\mathrm{d}d = 0. \tag{2.5}$$

Normally the effects of unearned income and wage on hours are of interest:

$$\frac{\mathrm{d}h}{\mathrm{d}I} = -\frac{wu_{cc}}{w^2u_{cc} + v_{ll}} < 0 \tag{2.6}$$

$$\frac{\mathrm{d}h}{\mathrm{d}w} = -\frac{u_c + hwu_{cc}}{w^2u_{cc} + v_{ll}} \gtrless 0. \tag{2.7}$$

---

[6] An open question is the sign of the double derivative with respect to disamenities. On one hand, if v_{dd} is negative then there potentially would be a level of disamenities that could not be reached because the disutility would be infinitely high. We can think of this as a "cumulative" effect of disamenities, where each additional disamenity seem worse and worse. On the other hand, if v_{dd} is positive we would have a "habituation" effect, where increasing disamenities would be less and less "costly" as they increased.

Increasing unearned income always reduces hours worked. As usual, the effect of increasing wage on hours depends on whether the substitution or the income effect dominates. If substitution dominates, higher wage leads to more hours worked, while if the income effect dominates higher wage leads to fewer hours worked—what is known as the backward bending labor supply curve.

What we are interested in here is the effect of changing job disamenities.

$$\frac{\mathrm{d}h}{\mathrm{d}d} = \frac{v_{ld}}{w^2 u_{cc} + v_{ll}} < 0 \tag{2.8}$$

Increasing job disamenities, holding wage and unearned income constant, unambiguously reduces time spent working. A corollary is that higher job disamenities means that a worker is less likely to work at all.

Our formulation of workers labor supply suggests that the most direct way to understand how workers respond to differences in job characteristics is to randomly allocate workers to combinations of job characteristics and offered pay and observe whether there are statistically significant differences in the probability that workers accept the job for the same pay and the amount of work that they decide to do. The model predicts that, holding wage constant, increasing job disamenities lowers the likelihood of a worker accepting a job and reduce the amount of work done if working. From the estimated labor supply we can then recover the value of job characteristics holding effort constant.

2.3 Experimental Design

Amazon's Mechanical Turk is the largest of the emerging micro-task markets with over 100,000 registered workers from over 100 countries [34]. Workers have to be 18 years or older, but otherwise there are few restrictions on participation. Work is paid per task rather than per hour—the corresponding hourly wage is lower than the average U.S. wage, but is close to the U.S. minimum wage. Individual tasks in a job are called HITs (Human Intelligence Tasks) and workers choose jobs from a list on the website that can be sorted by

criteria such as pay per HIT and posting date.[7] Workers can preview a job before accepting, and abort it without penalty at any time. Between 5,000 and 30,000 HITs are completed each day [141]. The Mechanical Turk labor market is built to be low friction for workers, allowing them to quickly move between jobs and work as much or as little as they desire on a given job.

Anyone can register to post jobs on Mechanical Turk. Examples of jobs include transcribing audio recordings into text, reviewing products, rewriting paragraphs, labeling images, searching for information, data entry, and answering surveys. Mechanical Turk allows requestors to require skills and "certifications" of workers. Our only requirement is that the computer accessing our jobs must be in the U.S. This allows us to estimate consistent labor supply functions, while achieving a sufficient sample size. U.S. Mechanical Turk workers are similar to the U.S. Internet population, and the income distribution closely follows the distribution for the overall U.S. population [**?**]. It is possible to circumvent our 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. Employers can reject HITs for subpar work. Having HITs rejected negatively affect workers because employers can exclude workers based on past rejection rates.

We offered the image tagging and letter writing jobs at different points in time. We chose these jobs for two reasons. First, they allow us to change job characteristics without altering the job itself. Second, we wanted a set of jobs that were relatively familiar to workers on Mechanical Turk and simple to explain.[8] In each experiment/job we randomize the levels

---

[7] The tagline for Amazon's Mechanical Turk is "Artificial Artificial Intelligence" to emphasize that these are jobs that are done by people. Appendix Figure **??** shows an example of a job listing on Mechanical Turk.

[8] A subset of other possible jobs that we considered were: 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.

of the four job characteristics and the pay offered. Both experiments use a full factorial design [**?**]. Experimental conditions are created by systematically varying the levels of each job characteristics 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, **?** and **(author?)** 56). With a factorial design, we can estimate main effects of the various job characteristics by "recycling" observations, without having to run individual experiments for each job characteristics.[9]

Data collection begins as soon as a worker clicks on our job in the job listing. To ensure that workers who show up at different times of the day are equally likely to be presented with all job characteristics, we listed all possible combinations in random order. Each worker is automatically assigned the next combination in the list. We observe whether the worker accepts the job and, if so, how many HITs are performed. Workers are not informed that the offered jobs are part of an experiment and are always presented with the same set of circumstances based on their unique worker ID number assigned by Mechanical Turk. We do not inform workers that they are part of an experiment to rule out an observer effect, where workers change behavior in response to being part of an experiment. Workers do, however, know that their output is potentially being monitored, but this monitoring is identical across the experiments and akin to what one would find in any job. The experiments are conducted exclusively through computers ruling out any experimenter bias.

Employers can only contact workers they have paid in the past. We therefore paid all new workers a \$0.25 "bonus". The bonus allows us to contact workers for a survey that we ran after the experiments independently of whether they completed any real HITs or not. 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 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

[9] It is also, in principle, possible to estimate interaction effects between different job characteristics, although our experiments were not powered to do that. We 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 we considered unlikely to achieve.

is not a concern here since the new worker bonus does not vary systematically across the different conditions and we are only interested in the differences between conditions.

2.3.1 Image Tagging Job

The image tagging job is similar to other tagging jobs on Mechanical Turk, where employers have workers go through images before deciding which ones to license. Once a worker clicks on the job, our program 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 the image is appropriate for a general audience.[10]

We change the job's agreeableness by varying the number of disagreeable images. 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 expressed as a ratio between 0 and 1. The number of disagreeable pictures do not change between HITs, but the ordering is random, so that a worker with, say, one disagreeable image per HIT 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.[11] This process 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 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 levels of disagreeableness. This variation is, however, completely random and therefore only make the estimated standard

[10] Appendix Figures 3.1 through 1.13 show the different parts of the page presented once a worker accepts the HIT.

[11] 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.

errors larger.

We alter the cost of learning through a "training component" with or without a "test." All workers read a description of different categories of tags and examples of each. Those in the "training" condition answer 15 questions, categorizing tags based on what they just read, and cannot work until all are answered correctly. Workers not selected for "training" are asked to click a button indicating that they had read and understood the content.

The probability of success is captured by our "approval" rate, which is the percent submitted HITs that we claim are approved. Hence, in the absence of other information, worker treats this as the perceived likelihood of being paid. We, however, pay everybody for all work irrespectively of the assigned approval rate. Furthermore, we never reject HITs. Because we ran the experiment over multiple days, the actual number displayed is drawn from an uniform distribution with a mean equal to either the low, 56%, or high, 93%, approval rate. This ensures that returning worker do not see exactly the same number over multiple days.

We implement the effect of availability outside of the factorial experiment, assigning 7% of all arriving workers to a special "low availability" condition where workers at specific times are told to wait for more HITs to become available. Workers in this condition are assigned only agreeable images, not asked to take the test, shown a high approval rate, and paid \$0.25 per HIT. Because this setup is different from the other conditions, we present all results both with and without workers assigned to the low availability condition.

The final part of the experiment is the pay offered. Workers are 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 1.13 shows an example of pay and availability. All workers can complete up to 50 HITs per day. This limit ensures that we do 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 is when fewest U.S. workers are on Mechanical Turk. Workers cannot skip HITs to, for example,

avoid specific images; if a worker leaves a HIT unfinished and returns before 07.58 GMT, the worker will see the same HIT again. This set-up allows us to both look at initial choice about labor supply, and what determines the decisions to return and amount of work to provide on subsequent days.

2.3.2 Letter Writing Job

In the letter writing job the task is to write a positive and supportive letter to a prison inmate. All names and profiles of the inmates are fictitious, but based loosely on one or more real inmate profiles from prison pen pal sites. We created 90 profiles and for each arriving worker our program creates a randomized list of the profiles. As for the image tagging experiment, workers cannot avoid specific prisoner profiles; a returning worker will see the latest, unfinished prisoner profile upon return.

We use different types of offenses to capture disagreeableness. One half of the workers were shown sexual related offenses and the other half crimes that could be perceived as less disagreeable. [12]

As in the image tagging experiment, cost of learning is captured with a training component with or without a "test." Everybody is asked to read the guidelines. Those selected for "training" condition got two questions to answer. Workers could not go on until they had answered both correctly.

The probability of success is shown by our "acceptance" rate for letters, although we pay everybody who submits acceptable letters. Either 94% or 51% are listed as accepted and the left-hand panel of Figure

Availability is modeled by varying the limit on the number of HITs available to the worker. Either 90 or 9 HITs were available.

Pay varies in \$0.1 increments from \$0.1 to \$1.0 per letter written. The letter experiment ran only through one 24-hour segment.

---

[12] Appendix Figures 3.2 through **??** show the different parts of the page presented once a worker accepts the HIT.

2.4 Estimation Strategy

Our experimental setup allows us to examine how job characteristics affect the amount of work, H, supplied. Job characteristics and pay are, however, only truly random the first time a worker visits a job. This is not an issue for the letter writing experiment, since it only ran for one day, but for the image tagging experiment we initially focus only on the first day a worker was observed, and return to what can be learned from longitudinal data below.[13]

Because we observe all workers who reject our jobs and all who accept, we can directly model the selection into work and amount of work supplied. We first estimate the effect of offered wage and job characteristics on the decision to work:

$$1[H_i > 0] = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i, \tag{2.9}$$

where $1[H_i > 0]$ is an indicator variable that takes the value 1 if the worker complete at least one HIT and 0 otherwise, w_i is observed wage per HIT for worker i, and $\mathbf{c}$ is a vector of job characteristics.

We next turn to the intensive margin. To show what the intensive margin results would look like for regular labor market data with no control for self-selection based on unobserved worker characteristics, we estimate the effects of wage and job characteristics on the number of HITs completed, conditional on workers completing at least one HITs:

$$H_i = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i \,\text{if}\, H_i > 0. \tag{2.10}$$

We estimate this using a censored regression model that takes into account upper bound censoring.

Finally, we use that we observe all workers—whether they reject or accept our offered job—and estimate a censored regression model on the number of HITs performed using all

[13] The first day observed is not necessarily the first day the experiment ran, but rather the first day we observed the worker in the image tagging experiment.

workers. The censored regression model takes into account both lower-bound censoring at zero HITs from people who reject our job and the upper-bound censoring built into the experiment.[14] The model implicitly requires two assumptions: that wages are observed for all workers independent of whether they work or not, and that wages are exogenous to the workers' labor supply. Neither assumption would be acceptable in standard labor market data, but are appropriate here. The experimental design provides an offered wage for all workers, whether they work or not, and this wage is by design exogenous to the labor supply because of randomization. The censored regression model also implies an assumption of no fixed costs associated with participation. In our case there are no fixed costs of work, or rather, the worker has already incurred them by joining Mechanical Turk (buying computer and internet connection and signing up for Mechanical Turk) and there are no fixed costs specific to our job.[15]

In addition to understanding the effects of job disamenities on labor supply, we are interested in the additional pay required to compensate for increasing job disamenities, holding labor supply constant. We calculate the average compensating wage for the different job disamenities from the extensive and intensive margin results, holding constant the probability of working and the number of HITs performed.

2.4.1 Longitudinal Analyses

Any differences between the results using all workers and the results restricting to only those who work illustrate the effects of self-selection. In the prior literature, with no experimental data available, fixed effects estimations have been suggested as a way of overcoming the selection problem [see, for example, 32, 67, 252]. The idea is that observing the same worker in multiple jobs allows us to eliminate unobservable worker traits that drive selection into jobs with different characteristics. There are, however, three drawbacks to this approach.

[14] In the cases where there are only one lower and one upper bound censoring point, the results will be the same as that from a Tobit model.

[15] For a more detailed discussion of the three assumptions see **(author?)** [27].

First, it requires workers that move between jobs with different characteristics. Second, fixed effects exacerbate any measurement errors in the data. Finally, if those who move between jobs are a non-random sample of workers, selection effects can still bias the results.

We ran the image tagging experiment over six days, where workers were presented with a randomly allocated set of conditions and pay each day they visited the job. Our setup means that workers are automatically presented with a variety of job characteristics and pay levels and that we have minimal measurement errors, eliminating two of the problems with fixed effects. Any differences between our experimental first visit results and fixed effects results will therefore be due to selection of worker *over time*. The selection happens because, although the conditions and wage that a worker face are randomized anew each day, prior conditions may affect a worker's likelihood of looking at our offered job again, and this likelihood depends on the worker's characteristics. Take two workers, one who intensely dislikes the disagreeable images and one who does not mind them as much, but otherwise they are identical. It is much more likely that we will see the worker who does not mind the disagreeable images again on a subsequent day than that we will see the worker who intensely dislike those images. In regular labor markets the selection over time can come about, for example, when workers learn over time about the job they work in or where there is sorting into different jobs over time based on unobserved productivity differences.

We first estimate how a given visit's job characteristics affect the probability that a worker will return:

$$V_i = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i, \tag{2.11}$$

where V_t is an indicator variable that takes the value 1 if a worker visits our offered job on a subsequent day and 0 otherwise. Days here are defined on the basis of the worker, not the experiment. A worker who, for example, looks at our offered job on the second day of the experiment will have that visit counted as the first visit and V then takes the value 1 if we observe the worker again and 0 otherwise. Observations from the last day of the experiment are dropped because we cannot observe whether the workers would have returned or not.

We estimate equation (2.11) for second through sixth visit.

Second, to compare with the first visit results, we repeat the estimations of how job characteristics affect amount of work done using fixed effects. We estimate the extensive margin:

$$1[H_{it} > 0] = \alpha + \beta_1 w_{it} + \mathbf{c}_{it}\beta_2 + \mu_i + \epsilon_{it}, \tag{2.12}$$

where i is the individual worker, t is visit number, and μ_i is a time invariant worker fixed effect, using information from all days where a worker looked at our job. We then estimate the intensive margin model using only those days where a worker complete at least one HIT:

$$H_{it} = \alpha + \beta_1 w_{it} + \mathbf{c}_{it}\beta_2 + \mu_i + \epsilon_{it} \,\text{if}\, H_{it} > 0. \tag{2.13}$$

Finally, we estimate the intensive margin model using all worker-visit observations, including those where a worker completed no HITs. Neither of these two intensive margin models take into account the censoring at zero HITs or the upper level censoring in the experiment.[16]

2.5 Results

During the image tagging experiment's six 24-hour segments, 4,311 workers visited the job.[17] The letter writing experiment ran for one 24-hour segment and 2,111 workers visited. As mentioned, we initially use only the first day a worker shows up for each experiment and cover longitudinal analyses below.[18] Many workers looked at our offered jobs but decided not to work. For the image tagging experiment 63% did not work, leaving 1,605 workers

[16] There are methods that allow for fixed effects in censored regression models, but the purpose of this paper is to evaluate the standard models used to examine the compensating wage differentials theory, rather than evaluate the different methods available. See **(author?)** [68] for a comparison of different selection correction models for panel models.

[17] We tried to run the image tagging experiment about seven months prior, but aborted it within hours because of server load issues. Removing workers who showed up for both has no effect on our results. The long period between the aborted attempt and the final run was partly because of the time required to design and run load testing programs for the servers and partly to minimize contamination between the aborted run and the final experiment.

[18] Appendix Figure 3.3 shows the distribution of work done in each experiment.

who completed one or more HITs on the first day they visited the job. For the letter writing experiment 73% did not work, leaving 578 workers who completed one or more HITs. In total, 4,366 letters were written and 60,695 images tagged—equal to 303,475 keywords on the first day. The payouts to workers were $3,055 and $3,808.

In the letter writing experiment, workers in the low availability condition were not allowed to work more than 9 HITs, whereas all others had an upper limit of 90 HITs. In the image tagging experiment, the low availability was not implemented as a fixed cut-off, so the only visible limit is the maximum of 50 HITs. Almost 100 workers reached the maximum on their first day working on the image tagging experiment.

Table 3.1 shows estimated effects of wage and job characteristics on extensive and intensive margins for the two experiments. For each experiment, the first column show extensive margin results, the second column intensive margin result for workers who completed at least one HIT, and the final column shows intensive margin results using all workers. The extensive margin estimations use a linear probability model with the dependent variable equal to 1 if a worker completed 1 or more HITs, and 0 otherwise. The intensive margin estimations use a censored regression model; for the "worked" sample there is only right-censoring, whereas for the "full" sample model there is censoring both at zero and at the maximum number of HITs a worker can perform.[19]

Both experiments show the importance of job characteristics on the decision to work, the extensive margin. Disagreeableness, learning cost, and low probability of success all have statistically significant negative effects on the probability of working and the reductions associated with less attractive characteristics are substantial. Increasing disagreeableness reduces the probability of working by more than 10 percentage points for the image tagging experiment and 7 percentage points for the letter writing experiment. This is equivalent to

[19] Alternative specifications are shown in Appendix Tables **??** and **??**. These include, for each experiment, a Logit model of participation (extensive margin) and an OLS model of the intensive margin for the "worked" sample. Each table shows results using wage and log wage separately. Finally, Appendix Table **??** shows the results for the image tagging experiment when excluding workers assigned to the low availability condition. In all cases the results are close to identical across specifications.

Table 2.1: Effects of Job Characteristics on Extensive and Intensive Margins

	Image Tagging Experiment			Letter Writing Experiment		
	Extensive Worked = 1	Intensive HITs Performed		Extensive Worked = 1	Intensive HITs Performed	
	LPM	Censored		LPM	Censored	
Sample	Full[a]	Worked[b]	Full[c]	Full[a]	Worked[d]	Full[e]
Log wage	0.048***	2.820***	3.277***	0.073***	4.962***	6.175***
	(0.011)	(0.514)	(0.484)	(0.014)	(1.074)	(0.934)
Disagreeableness	−0.115***	−4.058***	−6.215***	−0.074***	0.673	−3.717***
	(0.022)	(1.028)	(0.961)	(0.019)	(1.384)	(1.232)
Learning cost	−0.161***	−0.649	−6.133***	−0.110***	0.502	−5.789***
	(0.015)	(0.702)	(0.660)	(0.019)	(1.393)	(1.237)
Low probability of success	−0.077***	−1.090	−3.413***	−0.054***	1.628	−2.127*
	(0.015)	(0.693)	(0.653)	(0.019)	(1.380)	(1.228)
Low availability	−0.027	−3.162*	−2.421	−0.014	−5.406***	−3.415***
	(0.036)	(1.701)	(1.593)	(0.019)	(1.386)	(1.228)
Intercept	0.623***	14.523***	4.651***	0.462***	13.466***	−3.012*
	(0.023)	(1.036)	(1.010)	(0.025)	(1.623)	(1.547)
Observations	4,311	1,605	4,311	2,111	578	2,111
Dependent variable mean	0.372	7.6	2.8	0.274	7.6	2.1

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[b] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 1,605 workers who worked on the first day they were observed, 92 were right-censored observations.
[c] Of the 4,311 observations, 2,706 were left-censored observations, 1,513 uncensored observations, and 92 right-censored observations.
[d] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 578 workers, 68 were right-censored.
[e] Of the 2,111 observations, 1,533 were left-censored observations, 510 uncensored observations, and 68 right-censored observations.

a reduction of one third of the average probability of working for both experiments. Having to take the test before working has an even larger impact on the likelihood of working: for the image tagging experiment the reduction is 16 percentage points and for the letter writing experiment it is 11 percentage points. Being told that there is a low probability of success reduces the likelihood by 8 percentage points and 5 percentage points for the image tagging experiment and letter writing experiment.

For the intensive margin, the samples that correspond to normal labor market data—where you only observe those who work—provide a way to examine how self-selection can affect the estimated effects of job characteristics. Conditional on working, job characteristics have only small and statistically insignificant effects on completed number of HITs in both experiments; for the letter writing experiment, the estimates are even the wrong sign.[20]

[20] The large, negative, and statistically significant effect for availability in the letter writing experiment

The exception is disagreeableness in the image tagging experiment, where going from 0 to 5 disagreeable images reduces the number of completed HITs by 4 on average.[21]

If we instead use the full sample, which allows us to correct for selection, job characteristics in both experiments have strong, statistically significant, negative effects on number of HITs.[22] The effects of job disamenities on number of HITs are large. Going from 0 to 5 disagreeable images reduces the completed number of HITs by more than 6 for the image tagging experiment and close to 4 for the letter writing experiment. Given that the average number of HITs supplied across all workers are only at 2.8 and 2.1 these effects are substantial. Learning costs lead to a reduction of around 6 HITs for both experiments, whereas the lower probability of success leads to a reduction of 3.4 and 2.1.

Comparing the intensive margin results between the "full" and "worked" samples, the smallest difference in estimated effect of job disamenities is for image tagging disagreeableness, and even here the "full" sample estimate is more than 50% higher than the "worked" sample estimate. Hence, there is strong evidence that people do respond as predicted by our model. Holding the wage constant, worse job disamenities lead to less labor supplied, which supports the labor supply part of the compensating wage differentials theory. Importantly, this effect either disappears or is substantially muted if we do not control for self-selection into working. This confirms that prior research's failure to consistently find effects of job characteristics on wage comes from inadequate control for selection on unobservables.

The statistically significant, but nonetheless underestimated, effect of disagreeableness

---

is mechanical. Workers exposed to this condition had the number of available HITs limited to 9, whereas everybody else could complete 90 HITs before running out of available HITs.

[21] We believe there are two reasons behind the statistically significant, negative effect of disagreeableness on number of HITs. First, the image tagging HITs were designed to be completed quicker than the letter writing HITs. Shorter duration lowers the cost of trying a HIT and workers uncertain about their reaction to the disagreeable condition are therefore more likely to try a HIT in the image tagging than the letter writing experiment. Second, not all of the disagreeable images had exactly the same level of disagreeableness and the ordering of the disagreeable images were randomized for each worker. Hence, some workers saw less disagreeable images on the first HIT(s), making them more likely to work and when they encountered more disagreeable images they stopped working.

[22] The one exception is the low availability condition in the image tagging experiment, which is negative but not statistically significant.

among those who work, may parallel the effect of risk of death on wage in the literature.[23] Obviously, there is a big difference between disagreeable images and risk of death, but we have here a case where it looks like there is a substantial, statistically significant effect when not controlling for selection, but that effect is still far from the "true" value. If the relation between risk of death and observed wages is used to estimate value of life, but selection is not completely accounted for, then those estimates are probably substantially too low. As **(author?)** [13, p. C12] argues: "In reality, the vast majority of studies settle for providing estimates of V [value of life] among those people who accept risks." Hence, just because a job characteristic shows a statistically significant effect on wage that does not imply that this point estimate is unbiased. In fact, based on our results, it may be substantially underestimated.

2.5.1 Compensating Wages for Job Disamenities

We have shown that increasing levels of job disamenities have statistically significant and large effects on labor supplied, but what are the increases in wages necessary to compensate for worse job disamenities? These cost estimates depend, of course, on the exact job and job characteristic, but Table 3.1 allows us to calculate the increase in pay required to keep the average worker's probability of working constant—the extensive margin results—and the increase in pay necessary to keep the number of HITs supplied constant—the intensive margin results. Table 2.2 shows the results as both absolute changes in wages and percent changes in wages; both evaluated at the mean offered wage.[24] We focus here on the job disamenities that showed statistically significant effects on labor supply.

We begin with the extensive margin compensating wage differentials. Going from least to most disagreeable is worth between 56 and 66 cent per HIT, if the probability of working has

---

[23] See, for example, [245], [24], [109], [254], [256], [13], and [155].

[24] Appendix Table **??** show the compensating wage differentials for other specifications of the labor supply function. In addition, Appendix Section **??** shows an alternative approach where we treat observed wages as outcomes and directly estimate the association between job characteristics and wage.

Table 2.2: Compensating Wage Differentials
for
Job Disamenities Based on Estimated
Labor Supply Functions

	Image Tagging		Letter Writing	
	\$[a]	%	\$[b]	%
Extensive — LPM				
Disagreeableness	0.66	240	0.56	101
Learning cost	0.92	335	0.83	151
Low probability of success	0.44	160	0.41	74
Low availability	0.15	56	0.11	19
Intensive — "Worked"				
Disagreeableness	0.40	144	−0.04	−14
Learning cost	0.06	23	−0.03	−10
Low probability of success	0.11	39	−0.09	−32
Low availability	0.31	112	0.30	109
Intensive — "Full"				
Disagreeableness	0.52	190	0.33	60
Learning cost	0.51	187	0.52	94
Low probability of success	0.29	104	0.19	34
Low availability	0.20	74	0.30	55

Note. All results are based on Table 3.1. See that table for significance levels. The necessary increase in wage to compensate for a worse job disamenity, c, is $-\frac{\beta_c}{\beta_w} \times w\Delta c$, where wage is evaluated at the mean offered wage.
[a] Evaluated at the mean offered wage, \$0.275.
[b] Evaluated at the mean offered wage, \$0.55.

to remain constant.[25] These costs are large relative to the offered wage; the average offered wage is 27.5 cents for the image tagging experiment and 55 cents for the letter writing experiment. The needed increases in wages are equivalent to a 240% premium for the image tagging experiment and a 101% premium for the letter writing experiment.

The wage increase needed to compensate for learning cost are even larger than for disagreeableness at between 83 cents (151%) for the letter writing experiment and 92 cents (335%) for the image tagging experiment. Why are the costs of the test so high? First there the time involved; in our testing the time required is equivalent to completing between one or two HITs (after having read the instructions, which everybody waw required to do).

[25] If the probability of working has to remain constant, dividing the point estimate for the job characteristic by the point estimate for the wage, or $-\frac{\beta_c}{\beta_w} \times w\Delta c$, will approximate the required change in pay. For the image tagging experiment this is $\frac{0.115}{0.048} \times 0.275 = \$0.66$, whereas it is $\frac{0.074}{0.073} \times 0.55 = \$0.56$ for the letter writing experiment.

Second, workers may be uncertain about whether taking the test is worth it. Workers get to see the HIT, but will not be able to enter any information until they have been through the learning section. As Figure 3.3 show, many workers do relatively few HITs, which increases the cost of taking the test, especially if there is uncertainty about whether they will find the task worthwhile.

A lower probability of success requires just over 40 cents extra for both experiments. As expected this number is larger than the differences in expected payout. There is at most a 50 percentage points differences in the probability of success, but only for the letter writing experiment is the premium close to that at 74%. For the image tagging experiment it is substantially larger at 160%. There are two possible explanations for the larger compensation. First, workers are risk adverse and need to be compensated sufficiently for the extra risk associated with the lower probability of success. This extra risk includes both the payment for the job itself and the indirect cost that comes from potentially worse access to jobs if their HIT approval rate falls. Second, workers estimate how many HITs they are going to do and need to be compensated sufficiently. Since the average number of completed HITs are larger than one, that would suggest that the compensating wage from risk in the intensive margin should be lower than for the extensive margin.

The intensive margin results do, indeed, show lower increases needed to compensate for job disamenities than the extensive margin results. Based on the "full" sample, the extra pay required to have workers supply the same number of HITs when faced with the disagreeable condition instead of the not disagreeable condition is between 33 cent (60%) for the letter writing experiment and 52 cents (190%) for the image tagging experiment. Having to take the test requires just over 50 cents more for both experiments, equivalent to 187% increase for image tagging and 94% for letter writing. For lower probability of success the increases are 20 cents (34%) for letter writing and 30 cents (104%) for image tagging. The increase necessary to compensate for the low probability of success in the letter writing experiment is especially of interest since the difference is only 34%, which is less than the expected difference in pay. It is possible that workers were sufficiently confident that they would be

able to do the job satisfactorily that this particular job disamenity was less important.[26]

When we compare the "worked" and the "full" samples, the main result that stands out is the confirmation of the large effect of selection: disagreeableness, learning cost, and low probability of success all have the wrong sign for the letter writing experiment. For the image tagging experiment, learning cost and low probability of success have substantially lower estimated compensation in for the "worked" sample than for the "full" sample. The only job disamenities that is close between the two are disagreeableness, which is still about 1/4 less when not controlling for selection than when controlling.

In sum, the estimated effects of job characteristics are consistent across the two experiment, despite little overlap in the two sets of workers that looked at our experiments. Enticing workers to tolerate worse job disamenities requires substantial increases in pay. The required increases in pay may seem very large, but keep in mind that we are not estimating the marginal workers willingness to pay for avoiding job disamenities, but rather the average worker's. Our results are all the more strikning in that we here observe between 27 and 37% of potential workers completing at least one HIT. Even observing as high proportion of people working as we do here, we still get estimated compensating differentials that are very low if we do not control for self-selection. In standard labor market, it is difficult to estimate how many potential workers there would be for a given job (workers who could possible do the job, but decided not to because the offered combinations of pay and job characteristics were not attractive), but it is likely that we would observe a substantially lower proportion of people working to potential workers than what do here, further aggravating self-selection problems.

[26] Another possible explanation is that workers considered the job to be worthwhile in itself and therefore cared less about the pay. This, however, runs counter to the higher responsiveness to wages in the letter writing experiment than in the image tagging experiment.

2.6 The Role of Selection

The most obvious level at which selection takes place is the job offer. We address this type of selection above by randomization of wage and job characteristics together with observing all workers. The differences in results between the "worked" and the "full" samples show clearly the importance of self-selection into jobs with different characteristics. In this section, we further assess how selection plays out at different levels and its effect on our main results.[27]

At the labor market level, workers may, over time, change behavior or decide to leave the labor market altogether. Both may change how workers respond to changing job characteristics. We therefore first estimate whether worker experience on Mechanical Turk impact the response to job characteristics. Second, at the job level, initial offered characteristics might change the likelihood not just of whether a worker accepts a job, but also whether a worker even considers that job again. We therefore follow workers over time using the multi-day part of the image tagging experiment and examine how offered job characteristics affect whether a worker returns to visit our job again. This directly affect how useful fixed effects estimations of compensating wage differentials can be, so we also estimate how close to our experimental results we get using fixed effects estimations.

2.6.1 Does Tenure on Mechanical Turk Matter?

Length of tenure on Mechanical Turk can impact our results in two opposing manners. First, there may be labor market wide sorting over time. New workers arrive on Mechanical Turk on a regular basis, but some decide that the pay is too low and/or that they do not like the offered jobs and leave the labor market. Just as workers who work on our jobs are less sensitive to job characteristics than the overall sample of workers, the sorting over time could lead workers who have been on Mechanical Turk longer to be less responsive to job characteristics than more recently arrivals. Second, workers may learn how to behave in an

[27] Appendix Section **??** discusses how selection through survey response can also bias the estimated effects of job characteristics and wage.

optimal manner through work experience or leave the labor market if they do not. This would be equivalent to taxi drivers in New York who do not show optimizing behavior either exiting the profession or learning how to optimize over time [78]. If workers learn over time, we would expect new workers to try most available jobs on Mechanical Turk and be less responsive to wage. The effect would be that more experienced workers would be more responsive to job characteristics than less experienced workers.

Table 2.3: Effects of Job Characteristics on Extensive and Intensive Margins for Letter Writing Experiment Controlling for Experience on Mechanical Turk

	Without Interactions			With Interactions		
	Extensive Worked = 1	Intensive HITs Performed		Extensive Worked = 1	Intensive HITs Performed	
	LPM	Censored		LPM	Censored	
Sample	Full[a]	Worked[b]	Full[c]	Full[a]	Worked[b]	Full[c]
First observed June	−0.007	2.409	0.175	0.044	−1.448	1.629
2013 or before	(0.032)	(2.339)	(2.058)	(0.080)	(5.881)	(5.020)
Log wage	0.073***	4.960***	6.173***	0.075***	5.364***	6.462***
	(0.014)	(1.073)	(0.935)	(0.014)	(1.122)	(0.979)
Log wage				−0.006	−4.801	−2.497
× observed 2013				(0.049)	(3.716)	(3.257)
Disagreeableness	−0.074***	0.720	−3.715***	−0.074***	0.249	−3.857***
	(0.019)	(1.383)	(1.232)	(0.020)	(1.444)	(1.295)
Disagreeableness				−0.007	3.876	0.853
× observed 2013				(0.065)	(4.805)	(4.205)
Learning cost	−0.110***	0.594	−5.784***	−0.105***	−0.120	−5.811***
	(0.019)	(1.394)	(1.238)	(0.020)	(1.452)	(1.298)
Learning cost				−0.033	7.901	0.731
× observed 2013				(0.065)	(5.113)	(4.270)
Low probability of success	−0.054***	1.716	−2.126*	−0.039*	2.108	−1.132
	(0.019)	(1.382)	(1.228)	(0.020)	(1.441)	(1.288)
Low probability of success				−0.147**	−2.587	−9.862**
× observed 2013				(0.065)	(4.887)	(4.176)
Low availability	−0.014	−5.515***	−3.419***	−0.022	−5.013***	−3.630***
	(0.019)	(1.388)	(1.229)	(0.020)	(1.452)	(1.294)
Low availability				0.078	−4.363	1.560
× observed 2013				(0.066)	(4.807)	(4.205)
Intercept	0.463***	13.179***	−3.032*	0.459***	13.550***	−3.103*
	(0.025)	(1.645)	(1.565)	(0.026)	(1.688)	(1.628)
Observations	2,111	578	2,111	2,111	578	2,111
Dependent variable mean	0.274	7.6	2.1	0.274	7.6	2.1

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[b] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 578 workers, 68 were right-censored.
[c] Of the 2,111 observations, 1,533 were left-censored observations, 510 uncensored observations, and 68 right-censored observations.

A downside of Mechanical Turk is the lack of background information on workers, including how long their tenure on Mechanical Turk is. We can, however, create measures for how

Table 2.4: Effects of Job Characteristics on Extensive and Intensive Margins for Image Tagging Experiment Controlling for Experience on Mechanical Turk

	Without Interactions			With Interactions		
	Extensive Worked = 1	Intensive HITs Performed		Extensive Worked = 1	Intensive HITs Performed	
	LPM	Censored		LPM	Censored	
Sample	Full[a]	Worked[b]	Full[c]	Full[a]	Worked[b]	Full[c]
First observed June	−0.204***	−1.322	−9.166***	−0.185*	−8.887	−5.355
2013 or before	(0.031)	(1.992)	(1.628)	(0.102)	(6.894)	(5.454)
First observed	−0.226***	2.390	−8.822***	−0.250***	−1.872	−6.107
March/April 2014	(0.024)	(1.535)	(1.220)	(0.075)	(4.594)	(3.822)
Log wage	0.046***	2.807***	3.247***	0.042***	2.858***	3.039***
	(0.011)	(0.515)	(0.485)	(0.011)	(0.532)	(0.511)
Log wage				0.038	−1.123	3.209
× observed 2013				(0.047)	(3.415)	(2.707)
Log wage				0.024	−2.146	1.736
× observed 2014				(0.036)	(2.655)	(1.986)
Disagreeableness	−0.114***	−3.468***	−5.989***	−0.114***	−3.503***	−5.758***
	(0.020)	(0.987)	(0.920)	(0.022)	(1.024)	(0.976)
Disagreeableness				−0.011	9.058	0.254
× observed 2013				(0.090)	(6.293)	(4.773)
Disagreeableness				−0.007	−2.670	−4.557
× observed 2014				(0.069)	(4.796)	(3.645)
Learning cost	−0.164***	−0.943	−6.416***	−0.167***	−1.231*	−6.246***
	(0.014)	(0.681)	(0.646)	(0.016)	(0.706)	(0.683)
Learning cost				0.025	0.018	−2.173
× observed 2013				(0.064)	(4.385)	(3.363)
Learning cost				0.013	5.995*	−1.750
× observed 2014				(0.048)	(3.316)	(2.520)
Low probability of success	−0.077***	−0.894	−3.258***	−0.093***	−1.078	−3.799***
	(0.014)	(0.683)	(0.642)	(0.016)	(0.710)	(0.681)
Low probability of success				0.060	4.869	3.098
× observed 2013				(0.065)	(4.178)	(3.371)
Low probability of success				0.128***	0.361	5.790**
× observed 2014				(0.048)	(3.246)	(2.487)
Intercept	0.653***	14.046***	5.625***	0.656***	14.336***	5.382***
	(0.022)	(1.027)	(0.996)	(0.024)	(1.057)	(1.046)
Observations	4,311	1,605	4,311	4,311	1,605	4,311
Dependent variable mean	0.372	7.6	2.8	0.372	7.6	2.8

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[b] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 1,605 workers who worked on the first day they were observed, 92 were right-censored observations.
[c] Of the 4,311 observations, 2,706 were left-censored observations, 1,513 uncensored observations, and 92 right-censored observations.

long workers have been on Mechanical Turk, based on prior experiments and the experiments here. Tables 3.3 and 3.4 show extensive and intensive margins results, controlling for whether we have observed a worker before and when. Our earliest experiments on Mechanical Turk ran in September 2010 and January 2011 [248]. Of the workers in those experiments, only 45 show up among workers who looked at our letter writing experiment in March 2014, and

only 55 show up at our image tagging experiment in November 2014. We therefore combine these workers with workers we first observed at another experiment that ran in June 2013 to form the dummy variable "First observed June 2013 or before", which has a total of 205 workers in the letter writing experiment and 235 workers in the image tagging experiment. Finally, the letter writing experiment ran in March 2014 and in April 2014 we had an initial run of the image tagging experiment that was aborted within hours of starting because of server load issues. A total of 439 workers in the image tagging experiment were first observed in one of these two experiments. Hence, only about 10% of the workers in the letter writing experiment and 15% of the workers in the image tagging experiment were workers we had seen before.

In the letter writing experiment there is mostly little to no effect of experience. Without interactions there are no statistically significant effects of experience on either of the outcomes. With interactions, only the effect of low probability of success is statistically significant. More experienced workers respond more strongly negatively to being told that there is a lower probability of success than newer workers.

In the image tagging experiment, there are substantial and statistically significant negative effects both on the likelihood of working and the number of HITs completed using the "full" sample in the models without interactions. Having visited one of our previous experiments is associated with a reduction of more than 20 percentage points in the likelihood of working and a reduction of around 8 HITs for the intensive margin. Despite these effects there is little change in the point estimates for the effect of wage or any of the job disamenities. Including interactions between prior experience and job characteristics and wage does little to change the overall picture. Most of the effects of job disamenities are similar to the original effects. The one exception is again probability of success, but here the more experienced workers are more likely than newer workers if offered a low probability of success. A possible explanation for the reversal of the effect of low probability of success could be that workers who have previously seen a similar set-up learn that they are able to successfully complete the job despite the posted probability.

In sum, there is little in these results to strongly support or reject either of the two possible effects of experience. Part of the problem is lack of power, especially in the letter writing experiment. There are few workers who we observe across experiments, and the low number of returning workers makes it difficult to identify any effects of individual job characteristics. Probably the strongest results are for the image tagging experiment without interactions, which shows strong negative effects on labor supply of having more experience, but even here there is little change compared to the results in Table 3.1. Hence, selection at the labor market level over time does not appear to have a strong effect on our main result that workers exhibit a strong willingness to pay for job characteristics.

2.6.2 Selection Between Days

We next turn to the decision to revisit our job by following workers over time using the multi-day part of the image tagging experiment. Over the six days the image tagging experiment ran, we observed 7,954 worker-days, meaning that, on average, we observed each worker slightly less than two times. Over the entire image tagging experiment, a total of 218,030 images were tagged—equal to 1,090,150 keywords—and the total amount paid to workers was $14,346.45.

Table 2.5 shows the effects of last seen job characteristics on workers' probability of returning to look at the job again. We examine whether a worker returns at all, instead of on a specific day, for two reasons. First, workers may not return to Mechanical Turk on specific days because of factors other than the job characteristics. Second, not everybody entered the experiment on the same day. The dependent variable takes the value 1 if we observe the worker looking at our job again and 0 otherwise, independently of whether any work was done, but conditional on at least one day left in the experiment.[28]

[28] For example, if we observe a worker looking at our job for the first time on Tuesday and that worker returns on Thursday that would count as 1 for second day and the last seen job characteristics would be Tuesday's. If we do not observe the worker again after the Thursday visit, the 3rd visit outcome would be zero with Thursday's job characteristics, and the worker would not show up in any of the subsequent visits (4th through 6th).

Table 2.5: Previous Observed Job Characteristics' Effects on Return Visits

	Visit job posting given previous visit's characteristics[a]				
	2nd Visit	3rd Visit	4th Visit	5th Visit	6th Visit
Log wage	0.024**	0.033*	0.047*	−0.037	0.073
	(0.012)	(0.019)	(0.024)	(0.030)	(0.047)
Disagreeableness	−0.047*	−0.095**	−0.062	−0.074	−0.004
	(0.024)	(0.038)	(0.050)	(0.064)	(0.097)
Learning cost	−0.030*	−0.018	0.008	0.040	0.082
	(0.016)	(0.026)	(0.035)	(0.044)	(0.069)
Low probability of success	−0.031*	−0.014	−0.000	0.010	−0.039
	(0.016)	(0.026)	(0.034)	(0.044)	(0.068)
Low availability	0.004	−0.036	0.113	−0.012	0.017
	(0.040)	(0.056)	(0.071)	(0.094)	(0.138)
Intercept	0.574***	0.731***	0.851***	0.784***	1.002***
	(0.026)	(0.040)	(0.052)	(0.068)	(0.105)
Observations	3,863	1,505	656	308	81
Mean of dependent variable	0.486	0.620	0.765	0.834	0.914

Notes. Linear probability model estimates. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Dependent variable takes the value 1 if we observe the worker looking at our offered job and 0 otherwise, conditional on there being at least one day left in the experiment and independently of whether the worker worked on either day. Example: If we first observe a worker looking at our job on Tuesday and that worker returns on Thursday that would count as 1 for second day and the job characteristics exposed to would be those observed Tuesday. If we do not observe the worker again the 3rd visit outcome would be zero and the job characteristics would be Thursday's. The worker would not show up in any of the subsequent visit variables (4th through 6th).

Job characteristics strongly affect selection across days. Higher wage significantly increases the likelihood of a worker visiting the job again for the 2nd through 4th visits. Being exposed to less attractive conditions on the first visit significantly reduces the likelihood of a worker returning a second day. The negative effects of unattractive job characteristics remains for the third visit, although only disagreeableness is statistically significant. The exception is, again, the low availability condition. It may seem surprising that the effect of learning cost is negative, which means that workers asked to do the test were less likely to return. A possible reason is that the sample here is everybody who looked at the job, rather than only those that worked. These effects of job characteristics show up despite our preview page specifically stating "The task and pay change each day, as we find and tune new tasks."

With selection, job characteristics should have less and less of an impact on the decision to revisit the higher the visit number. Sample sizes, however, also become smaller and

smaller, making it difficult to draw strong conclusions. Disagreeableness, for example, show a negative effect on probability of returning for all days—except for the 6th day visit—but the effects are not statistically significant different from each other. Learning costs show the clearest trend with the effect negative and statistically significant on returning for a second day, and then a consistent positive trend. It is, however, never statistically significant for any of the subsequent visits.

What is consistent with selection driving a diminishing effect of job characteristics over visits is the increased probability of returning over time, as shown by the mean of the dependent variable. Of the 3,863 workers who could possibly return for a second visit less than half did. For the third visit the return rate increases to over 60% and continues to increase for later visits. For workers who showed up on the first day of the experiment and looked at the experiment the first five days, more than 90% return to look at it on the experiment's last day.

2.6.3 Worker Fixed Effects Results

Table 3.5 shows fixed effects estimates for both extensive and intensive margin.[29] These results provide us with an indication of how strongly fixed effects estimates are affected by the selection over days. In the absence of substantial selection we should find similar estimates for both labor supply and wage differentials across the analyses using the first visit data and the longitudinal data. Corresponding to our analysis of whether workers return to our job, the selection over time shows up in the higher percentages of people who work compared to the first day analysis. In the panel data, 42% of workers work—up from 37% for the first day analysis—and the average number of HITs performed per worker per day is almost twice as large as in the first day data.

For the extensive margin, three results stand out. First, the effects of wage and the low availability condition are substantially stronger in the fixed effects data than in the first day

[29] Appendix Tables **??** and **??** show additional specifications.

Table 2.6: Effects of Job Characteristics on Extensive and Intensive Margins—Worker Fixed Effects

	Extensive	Intensive	
	Worked = 1	Number of HITs Performed	
	Linear	Linear	linear
Sample	Full	Worked[a]	Full
Log wage	0.113***	7.365***	4.562***
	(0.009)	(0.552)	(0.259)
Disagreeableness	−0.130***	−6.487***	−3.867***
	(0.018)	(1.014)	(0.529)
Learning cost	−0.095***	−0.192	−0.854**
	(0.012)	(0.689)	(0.359)
Low probability of success	−0.056***	−1.758**	−1.099***
	(0.012)	(0.704)	(0.362)
Low availability	−0.135***	−13.025***	−7.021***
	(0.027)	(1.538)	(0.789)
Observations	7,954	3,330	7,954
Number of workers	4,311	1,830	4,311
Mean of dependent variable	0.419	13.095	5.482

Note. Standard errors in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.
[a] This sample has a higher number of people than than the first day results because there are 125 workers that did not work on the first day they visited the job, but did work on a subsequent day. Hence, the first day number of observations for the intensive margin is 1,605, whereas it is 1,830 for the fixed effects estimations on intensive margin.

data. Second, there is a slight increase in the effect of disagreeableness on labor supply, although the differences are not statistically significant. Third, both learning costs and low probability of success matters less in the fixed effects results than the first day results.

For the intensive margin, the "worked" sample results show a substantially larger effect of wage for the fixed effects results than for the first day results.[30] Similarly, the effects of disagreeableness, low probability of success, and low availability all have statistically significant negative effect. Although the point estimates are larger in the fixed effects results than first day results, the effects are smaller when considered as a ratio of the mean number of HITs performed.

For the "full" sample fixed effects model the results are consistent with the extensive margin results. Wage and low availability both have substantially stronger effects in the fixed effects estimation than the first day analysis, and there are weaker, but still statistically

[30] Neither of the intensive margin estimates take into account censoring from below or above.

significant effects of learning costs and low probability of success. The main difference is the lower effect of disagreeableness, which the fixed effects estimate around 40% of the first day estimate.

Table 2.7: Compensating Wage Differentials for Job Disamenities Based on Estimated Fixed Effects Labor Supply Functions for the Image Tagging Experiment

	Extensive—LPM		Intensive—"Worked"		Intensive—"Full"	
	$	%	$	%	$	%
Disagreeableness	0.32	115	0.24	88	0.23	85
Learning cost	0.23	84	0.01	3	0.05	19
Low probability of success	0.14	50	0.07	24	0.07	24
Low availability	0.33	119	0.49	177	0.42	154

Note. All results are based on Table 3.5 for the image tagging experiment. See those tables for significance levels. The necessary increase in wage to compensate for worse job disamenities, c, is $-\frac{\beta_c}{\beta_w} \times w\Delta c$, where wage is evaluated at the mean offered wage, \$0.275.

The effects on labor supply are, however, only one part of the picture. Table 2.7 shows the calculated wage differentials for the four job disamenities. Two things stand out. First, the three job disamenities that had substantial wage differentials using the randomized assignment of job disamenities and pay, disagreeableness, learning costs, and low probability of success, all have substantially smaller differentials based on the fixed effects results. Disagreeableness has the smallest difference and even then the wage differential based on panel data is less than half of what we found using randomization. Second, low availability is the only factor that shows an increase over the first day results.

These results are consistent with selection over time, where those workers who return often, and who care less about job characteristics, show up more frequently in the data and therefore drive the results. This reinforces the conclusion that what is important for self-selected workers may not be important for the average worker. Finally, to the extent that the results here apply to other labor markets, using panel data and worker fixed effects will only help to some extent with establishing whether the compensating wage differentials theory holds.

2.7 Conclusion

We ran two experiments on Mechanical Turk, randomly allocating different wage and job characteristics to workers as they looked at our jobs. These experiments allowed us to estimate labor supply functions, while avoiding the self-selection problems that have plagued the prior literature. Our experimental results show clearly that workers do trade off between wage and job amenities. In line with our theoretical model, workers presented with less attractive job characteristics, holding wage constant, supply significantly less labor. This holds for the job's disagreeableness, cost of learning, and probability of success. What is more, we find similar effects across both experiments even though different groups of workers saw them.

The implied and estimated compensating wage differentials are substantial. Getting the average worker to maintain the same probability of working from the least disagreeable to the most disagreeable condition requires an increase in wage of between 100 and 240% of the average offered wage. Having to take a test before working (learning cost) requires an even larger increase—150 to 330%—while being made to accept a lower probability of success requires an increase of between 75 and 160%. The intensive margin results are smaller, but still very substantial.

These wage differentials are much larger than anything found in the prior literature. There are two reasons for this. First, we estimate wage differentials for the average worker instead of the marginal worker. Second, we do not have the same selection problems as the prior literature. We observe all workers that look at our job, whether or not they decide to work. Indeed, if we only use information from those workers who work, we find mostly no effect of job disamenities on labor supply or wages. The one case where a job disamenity significantly affects labor supply for the self-selected sample, the estimated wage differential is still about 1/3 less than when we control for self-selection into working. We also show that, although using panel data to overcome the self-selection problem is a theoretically appealing approach, the results are still liable to bias from self-selection. In our case, we

observe selection over days, resulting in compensating wage differentials based on panel data that are half or less of what we find using the first day experimental data.

An important question is the internal and external validity of our results. Internal validity comes from our randomization of combinations of job disamenities and pay across arriving workers and that we observe all arriving workers to our jobs. This allows us to assign a causal interpretation to the effects of pay and job disamenities on labor supply. All of our results are conditional on workers looking at our jobs, and we, unfortunately, have no way of knowing the proportion of workers who were on Mechanical Turk when we offered the jobs but decided to not look at our job. We do, however, find little change in our results when we control for worker experience on Mechanical Turk despite that the experience variable is based on substantial different types of experiments. Our results are also consistent across our two experiments. This consistency, even though there was little overlap in the workers, provides additional evidence of internal validity of our results.

When it comes to external validity, 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. We believe that our results have external validity for three reasons, despite these differences. First, workers on Mechanical Turk are people actively looking for work. The pay may not be high, but according to emails that we received and comments on Mechanical Turk workers' discussion forums a large number of people rely on Mechanical Turk as a substantial source of income. Second, surveys show a distribution of worker characteristics that is similar to the general labor market. Finally, advances in computational and communication technology are rapidly pushing labor markets from the traditional form into a more flexible form, where there are fewer permanent jobs and more people working as independent contractors. A sign of the growing importance of freelancing, independent contracting, and consulting work in the U.S. economy is a recent estimate 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 [184].[31]

[31] There is, however, substantial uncertainty about these numbers since the Bureau of Labor Statistics

The main caveat to external validity is that Mechanical Turk has both many workers and many employers. In other words, what we show is that workers behave as predicted in a situation that is close to the standard neoclassical model. What we cannot establish is the extent to which the results would be different if there were only a limited number of employers.

Our results have important implications for policy. First and foremost, standard estimates of compensating wage differentials substantially underestimate the value that workers assign to job characteristics. Even attempts to overcome selection issues, such as using panel data, produce results that have a substantial downward bias. Since estimated wage differentials are used to design policy, we risk assigning a too low value to changing an outcome, thereby making it less likely that the policy will be implemented. One example is speed limits. Presumably we care about the average person's statistical value of life rather than the marginal worker who self-select into a more risky job. To the extent that our results are transferable to the statistical value of life literature, the implication would be that we are placing a too low value on preventing deaths from speeding.

Second, although there is substantial pessimism about the future of workers' rights and ability to secure "fair" wages, our results indicate that with sufficient number of workers and employers, workers are able to exercise choices on labor supply. Even in a situation like Mechanical Turk, which may seem like the quintessential "race to the bottom" labor market, we still observe workers rejecting jobs because the pay is too low for the offered job disamenities. This means that employers are still forced to trade off how fast they want their job done versus any savings that might come from paying a lower salary, even without policy intervention.

An interesting question for future research that arise from the policy discussion is whether workers from states with more restrictive labor laws and higher minimum wages are more likely to be on Mechanical Turk and how those policies affect their behavior. Another

---

does not directly count these types of employment.

potential area of future research is determinants of quality of work. We have, in the interest of space, ignored potential differences in the quality of work, but an important question is whether factors such as wage, likelihood of success, and testing employees improve the quality of work provided. In other words, does the efficiency wage theory hold? Finally, we have shown that workers trade off between wage and job characteristics. This, however, only addresses the labor supply side. To fully understand wage setting in labor markets we now need to better understand employers' decisions making process. Mechanical Turk lends itself well to tackle questions like these and is a promising platform for doing research on a whole host of important questions in labor economics, and, more generally, in applied micro-economics.

Chapter 3

LABOR SUPPLY ELASTICITIES IN A LOW FRICTION LABOR MARKET

3.1 Introduction

A persistent question in economics is the size of labor supply elasticities. Labor supply elasticities are of interest in their own right, but especially because they impact how we evaluate a variety of policies, first and foremost tax policy. Two stylized "facts" emerge from the large literature on labor supply elasticities. First, estimates of labor supply elasticities using micro data are small, especially compared to what is implied by macro models [26, 49, 50, 150]. Second, intensive margin elasticities are substantially lower than extensive margin elasticities [119, 47]. Neither of these conform nicely to economic models, which has led to a large literature trying to explain them away.[1]

One possible explanation for the divergent results may be that using observed wages to estimate labor supply elasticities is associated with a number of potential data and econometric issues that lead to biased estimates of the effects of wages on labor supply. First, individual wages may be endogenous if, for example, unobservable preferences for work are correlated with unobserved productivity and therefore wages [27]. Second, wages are often measured with substantial error, especially if wage per hour is calculated from self-reported hours worked and pay. Third, wage changes may come from shifts in either labor supply or labor demand or both simultaneously, and the driver of these shifts may unobservable [198, 79]. Finally, wages are only observed for workers who decide to work at a given job and for a given wage.[2]

---

[1] The literature is too large to survey here. For a recent review see [151].

[2] For example, when using tax rate kinks to estimate labor supply elasticities micro estimates are identified

We take a different approach from the previous literature. Instead of trying to infer labor supply elasticities from observed wages, we run field experiments that allow us to directly estimate labor supply elasticities. We offer jobs, randomly allocating arriving workers to different wages within each job, and observe workers' decision on whether to work or not and amount of work supplied. There are two main contributions. First, we show how elasticities behave in an environment that is close to the standard neoclassical model. Second, we can estimate labor supply elasticities while controlling for selection because we observe all workers looking at our jobs, whether they work or not. Our experimental setup allows us to estimate extensive margin elasticities and intensive margin elasticities without encountering the econometric problems normally associated with estimating elasticities from regular labor market data. We show that labor supply elasticities are substantial in this market and that intensive margin elasticities are double that of extensive margin elasticities.

This work is made possible by the emergence of online labor markets for micro-tasks. We use Amazon's Mechanical Turk (www.mturk.com), which allows us to control all aspects of jobs offered, including pay. We offer two separate jobs on Mechanical Turk: One asks workers to tag images with keywords and the other asks them to write letters. Each job requires different skill sets and appeal to workers with different interests, thereby providing more general validity to our experiments. As described in [202], in each experiment we randomize the job characteristics that workers are presented with and the pay offered. Here we use the randomized pay to estimate differences in work supplied for different levels of pay.

Mechanical Turk has three major advantages when estimating labor supply elasticities. First and foremost, we can experimentally vary the wage. This means that offered wages are exogenous to worker characteristics and we do not have to worry about supply or demand shifts affecting wages. Second, there is substantially less scope for measurement error than in prior studies. We set the wage and directly observe whether a worker accepts or rejects a

---

from the *ex-post* decision to move to another job, whereas macro estimates are identified from *ex-ante* decisions [49]. In the presence of search costs and other friction micro estimates will therefore be lower than macro estimate, even if using the same tax variation.

job offer, and the amount of work supplied if the job offer is accepted. Finally, we observe offered wage independently of whether a worker decides to work or not.

In addition to data and econometric issues, a number of reasons have been suggested for why estimated labor supply elasticities are small and for why intensive margin elasticities are larger than extensive margin elasticities. The three are investments in human capital, credit constraints, and frictions in labor markets [137, 65, 47]. Mechanical Turk is especially of interest because it is less rigid than most off-line labor markets and therefore allow us to understand how labor supply elasticities perform in a low friction environment. There are low fixed and search costs on both sides and only short-term jobs are offered. Workers who find a job unattractive simply move on to another job without occurring any penalties, and employers engage in either minimal or no initial screening of workers. We can also rule out human capital accumulation as a factor that may impact our estimated elasticities.

Mechanical Turk does have two downsides when estimating labor supply elasticities. First, we have little to no information about individual workers. We do present results based on survey information about basic demographic characteristics, but opposite our overall results who responds depends potentially on unobserved characteristics. Second, we cannot observe overall behavior for workers on Mechanical Turk, only how they perform in our experiment. We return to this later.

This work is closely related to [198] work on labor supply elasticities for stadium workers, the work by [39] and [79] on taxi drivers in New York City, [19] on archeological dig, and [96] work on rural labor supply in Malawi.

We find strong positive elasticities, so little evidence of the negative wage elasticities found by [39] or the negative elasticity of effort in [81].

Our estimates of labor supply elasticities are decidedly short-run. The longest job ran only over 6 days.

It is possible for us to show large effects in what is effectively a standard neoclassical labor market.

3.2 Experimental Design

Amazon's Mechanical Turk is the largest of the emerging micro-task markets with over 100,000 registered workers from over 100 countries [34]. Workers have to be 18 years or older, but otherwise there are few restrictions on participation. Work is paid per task rather than per hour—the corresponding hourly wage is lower than the overall US labor market, but will be close to the U.S. minimum wage. Individual tasks in a job are called HITs (Human Intelligence Tasks) and workers choose jobs from a list on the website that can be sorted by criteria such as pay per HIT and posting date.[3] Workers can preview a job before accepting it, and abort the job without penalty at any time. Between 5,000 and 30,000 HITs are completed each day [141]. The Mechanical Turk labor market is built to be low friction for workers, allowing them to quickly move between jobs and work as much or as little as they desire on a given job. That does not mean that it is costless to move between jobs; there are clearly still search cost, and for some jobs it is more difficult to assess the work burden per HIT up front than others. We will return to this below.

Anyone can register to post jobs on Mechanical Turk. Examples of jobs include transcribing audio recordings into text, reviewing products, rewriting paragraphs, labeling images, searching for information, data entry, and answering surveys. Mechanical Turk allows requestors to require skills and "certifications" of workers. Our only requirement is that the computer accessing our jobs must be in the U.S. This allows us to estimate consistent wage responses, while achieving a sufficient sample size. U.S. Mechanical Turk workers are similar to the U.S. Internet population, and the income distribution closely follows the distribution for the overall U.S. population [**?**]. It is possible to circumvent our 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. Requestors can reject HITs for

[3] The tagline for Amazon's Mechanical Turk is "Artificial Artificial Intelligence" to emphasize that these are jobs that are done by people. Appendix Figure **??** shows an example of a job listing on Mechanical Turk.

subpar work. Having HITs rejected negatively affect workers because requesters can exclude workers based on rejection rates [128].

We offered two separate jobs, which ran at different times during 2013 and 2014. The two jobs were designed to be attractive to different segments within the Mechanical Turk worker community and to require different skill sets. In one job we offered workers a pictures tagging task, and in the other asked them to write short letters. This allows us to compare how labor supply elasticities respond to different types of work and thereby achieve broader validity. The experiments were originally designed to test the compensating wage differentials theory by randomizing offered combinations of job characteristics and pay [202]. Here we use the randomization of pay offered.

Data collection begins as soon as a worker clicks on our offered job in the job listing. To ensure that workers who show up at different times of the day are equally likely to be presented with all job characteristics, we listed all the possible combinations in random order. Each worker that looks at our job is automatically assigned the next combination in the list. We observe whether the worker accepts the job and, if so, how many HITs are performed. Workers were not informed that the offered jobs were part of an experiment and were 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, however, know that their output is potentially being monitored, but this monitoring is identical across the experiments 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". 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 contact workers for our survey independently of whether they completed any real HITs or not. 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 may bias upward our estimates of extensive and intensive margin elasticities. [TK run estimations for those who did not get a new worker bonus and those who did]

The image tagging job is straightforward and similar to many other tagging jobs offered on Mechanical Turk, where requestors have worker go through images before deciding which ones to license. Once a worker accepts the image tagging job, our program selects five pictures and for each image we ask the worker to provide five tags or keywords in addition to clicking a radio button indicating whether each the image is appropriate for a general audience. Figure 3.1 shows part of the page presented once a worker accepts the HIT, including one image.

Figure 3.1: Image Tagging Experiment

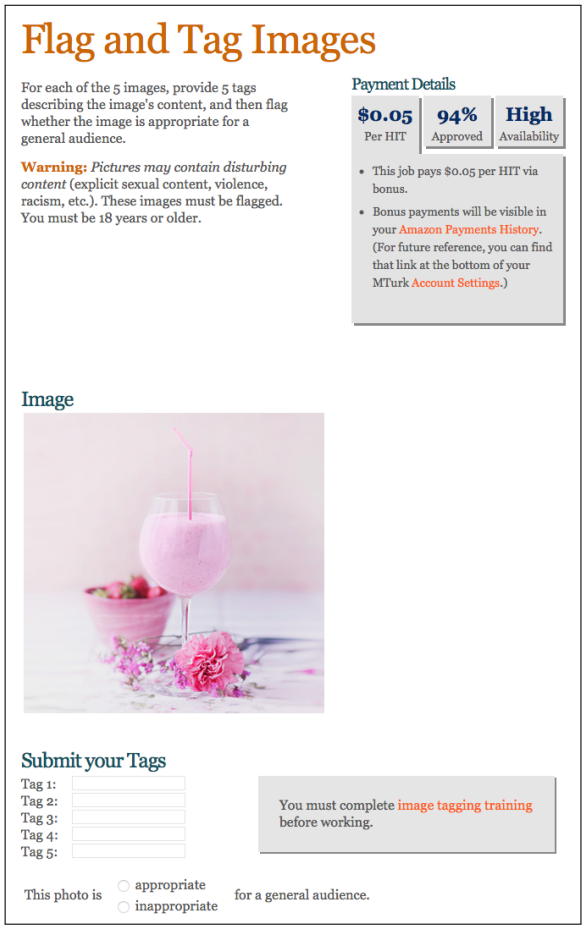
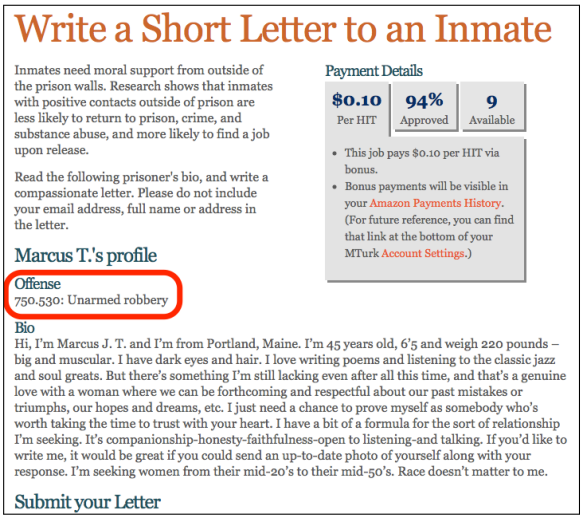


Figure 3.2: Letter Writing Experiment



Workers were randomly assigned to a pay per five images tagged, equal to 25 tags, of between USD 0.05 and USD 0.50 in USD 0.05 increments. The experiment ran over a six day period in 24 hour segments from 07.58 GMT. A worker would see one set of conditions during each 24 hour period and then after 07.58 GMT the worker job conditions and pay would be randomized anew. The randomization on subsequent days does not take account of previous job characteristics or pay that the worker has experienced. We choose 07.58 GMT because that was the time of the day where there were the fewest number of workers on Mechanical Turk.[4] This set-up allows us to both look at initial choice about labor supply and what determines the decisions to return on subsequent days and amount of work provided.

In the letter writing job the basic task is to write a positive and supportive letter to a prison inmate. An example is shown in Figure 3.2. The pay vary in 10 cent increments from USD 0.1 to USD 1.0 per HIT completed. As for the image tagging job this pay remained with the worker throughout the experiment. The letter experiment ran only through one 24 hour segment.

3.3 Estimation Strategy

Our experimental setup allows us to examine how wage affects the amount of work, H, supplied. Because we observe all workers, whether they reject or accept our jobs, we can directly model the selection into work and amount of work supplied. We first estimate the effect of offered wage and job characteristics on the decision to work:

$$1[H_i > 0] = \alpha + \beta_1 \log(w_i) + \mathbf{c}_i\beta_2 + \epsilon_i, \tag{3.1}$$

where $1[H_i > 0]$ is an indicator variable that takes the value 1 if the worker complete at least one HIT and 0 otherwise, w_i is observed pay per HIT for worker i, and $\mathbf{c_i}$ is a vector of job

[4] A worker working around the change point could potentially see one set of condition initially and another set later on.

characteristics.[5] We estimate this extensive margin decision using both a linear probability model and a Logit model.

We next turn to the intensive margin. To show what the intensive margin results would look like for regular labor market data with no control for selection, we estimate the effects of wage and job characteristics on the number of HITs completed, conditional on workers completing at least one HIT:

$$H_i = \alpha + \beta_1 \log(w_i) + \mathbf{c}_i\beta_2 + \epsilon_i \,\text{if}\, H_i > 0. \tag{3.2}$$

We present both OLS results and results from a censored model that takes into account upper bound censoring.

Finally, we estimate a censored regression model that takes into account both the lower bound censoring that occur at zero HITs when workers reject our job and the upper bound censoring built into the experiment.[6] The censored regression model implicitly requires two assumptions: that wages are observed for all workers independent of whether they work or not, and that wages are exogenous to the workers' labor supply. Neither assumption would be acceptable in standard labor market data, but are appropriate here. The experimental design provides an offered wage for all workers, whether they work or not, and this wage is by design exogenous to the labor supply because of randomization. The censored regression model also implies an assumption of no fixed cost of participation. In our case there are no fixed costs of work, or rather, they have already been incurred by the worker by joining Mechanical Turk (buying computer and internet connection and signing up for Mechanical Turk) and there are no fixed costs specific to our job.[7]

We calculate three wage elasticities for each experiment: the extensive margin elasticity, the intensive margin elasticity conditional on working, and the "overall" elasticity for all

[5] We do not show the effects of job characteristics here. See [202] for those results.

[6] In the cases where there are only one lower and one upper bound censoring point, the results from this estimation will be the same as that from a Tobit model.

[7] For a more detailed discussion of the three assumptions see **(author?)** [27].

workers whether they work or not. One question is the extent to which workers consider the offered wage a transitory wage change, in which case the elasticities estimates are Frisch elasticities. If workers see the offered wage as transitory we would expect no substantial income effect. Our expectation is that workers on Mechanical Turk treat each offered job as temporary and that if a job shows up which pays more than expected they treat this as a temporary shock since most active jobs on Mechanical Turk are relatively short-lived. Furthermore, all other jobs on Mechanical Turk remains at the same wage.

The extensive margin elasticity captures the effect of wage changes on the probability of working on our given job. The coefficient on the extensive margin regressions do not directly show the extensive margin elasticity, but we can calculate it as

$$\epsilon_e = \frac{\partial \Pr[H > 0] / \Pr[H > 0]}{\partial w / w} = \frac{\beta_1}{\Pr[H > 0]}, \tag{3.3}$$

using the results from our estimation of equation (3.1) with log wage as the explanatory variable, where $\Pr[H > 0]$ is the probability that the number of HITs performed is greater than zero. This requires picking a number for the probability; here we use the participation rate for each experiment.[8]

Similarly, we calculate the intensive margin elasticities as

$$\epsilon_i = \frac{\partial H / H}{\partial w / w} = \frac{\beta_1}{H}, \tag{3.4}$$

using the results from equation (3.2) with log wage as the explanatory variable, where H is the number of HITs completed. For the intensive margin conditional on working we use the number of HITs completed by those who did at least one HIT for the experiment, whereas for the intensive margin for all workers we use the average over all workers.[9]

A downside of Mechanical Turk is the lack of background information on workers, in-

[8] If wage enter linearly in labor supply function the formula is $\beta_1 \frac{w}{\Pr[H>0]}$.

[9] Dividing by the average number of HITs for the censored regression ignores that the latent number of HITs supplied is lower than the observed when setting those who work to zero.

cluding how long their tenure on Mechanical Turk is. We can, however, create measures for how long workers have been on Mechanical Turk, based on prior experiments and the experiments here. Experience on Mechanical Turk is especially of interest since [79] argues that taxi drivers in New York City are more likely to be optimizers the longer they have been working. Hence, it may be that workers who have been on Mechanical Turk longer may be more "rational" and therefore less likely to try the HIT unless they find it worthwhile. We estimate the same set of models as above using dummies for experience with and without interaction with wage.

3.3.1 Longitudinal Analyses

Fixed effects estimations has been used to overcome problems that arise from unobservable worker characteristics [see, for example, 178, 33, 9, 198]. The idea is that observing the same worker allows us to eliminate unobservable worker traits that drive selection into jobs. There are three drawbacks to this approach. First, it requires workers that receive different wages. Second, if those who change wages is a non-random sample of workers there can still be important selection effects that will bias the results. In our case, although fixed effects will remove time-invariant worker characteristics, we may still have workers who self-select out of work because their reservation wage is above the offered wage. These workers will remain unobserved, potentially leading to biased results even with longitudinal data. We varied our offered wage substantially more than it would be possible to do in a regular off-line labor market to lower the chance of self-selection, but workers may still decide not to work because we are not offering them enough to surpass their reservation wage. Finally, fixed effects exacerbate any measurement errors in the data.

We ran the image tagging experiment over six days to examine whether the fixed effects approach provides comparable results to our experimental results. Workers were presented with a randomly allocated set of conditions and pay each day they visited the job. Although the conditions and wage that a worker face are randomized anew each day, only the conditions on the first day a worker visits can truly be considered random. Conditions may affect a

worker's likelihood of looking at our offered job again. Hence, all data collected after the first day a worker visits are potentially affected by self-selection.

To compare with the first visit results, we estimate how wage affects amount of work done using fixed effects. We first estimate the extensive margin

$$1[H_{it} > 0] = \alpha + \beta_1 \log(w_{it}) + \mathbf{c}_{it}\beta_2 + \mu_i + \epsilon_{it}, \tag{3.5}$$

where i is the individual worker, t is visit number, and μ_i is a time invariant worker fixed effect, for all days where the worker looked at our job using a regular linear fixed effects model and a conditional Logit model. We then estimate the intensive margin model using only those workers who complete at least one HIT on their visit

$$H_{it} = \alpha + \beta_1 \log(w_{it}) + \mathbf{c}_{it}\beta_2 + \mu_i + \epsilon_{it} \,\text{if}\, Y_{it} > 0. \tag{3.6}$$

Finally, we estimate the same model but include all worker-visit observations, including those where the worker did no HITs. Neither of these two models take into account the censoring at zero HITs and the upper level censoring in the experiment.[10]

3.4 Results

Job characteristics and pay are only truly random the first time a worker visits the job. For the image tagging experiment we therefore initially focus only on the first time a worker was observed, and return to what can be learned from the panel aspect below.[11] During the image tagging experiment's six 24 hour segments, 4,311 workers visited the job.[12] The letter

[10] There are methods that allow for fixed effects in censored regression models, but the purpose of this paper is to evaluate the standard models used to examine the compensating wage differentials theory, rather than evaluate the different methods available. See **(author?)** [68] for a comparison of different selection correction models for panel models.

[11] The first day is not necessarily the first day the experiment ran, but rather the first day we observed the worker in the image tagging experiment.

[12] We tried to run the image tagging experiment about seven months prior, but aborted it within hours because of server load issues. Removing workers who showed up for both has no effect on our results. The

writing experiment ran for one 24 hour segment and 2,111 workers visited. Figure 3.3 shows the distribution of work done in each experiments. The panels on the left include those who chose not to work, while the panels on the right are conditional on working to show the distribution of work done more clearly.

Many workers looked at our offered jobs but decided not to work. For the image tagging experiment 63 percent did not work, leaving 1,605 workers who completed one or more HITs on the first day they visited the job. For the letter writing experiment 73 percent did not work, leaving 578 workers who completed one or more HITs The advantage of Mechanical Turk is that we observe all the workers who decide that they do not want to work, and who would not be observed in standard labor markets.

One of the job characteristics tested in the experiment was availability and for the letter writing experiment workers in the low availability condition were not allowed to work more than 9 HITs, whereas all others have an upper limit of 90 HITs. Both show clearly in the histogram. In the image tagging experiment the low availability was not implemented as a fixed cut-off, so the only limit is the maximum of 50 HITs. Almost 100 workers reached the maximum on their first day working on the image tagging experiment.

In total, 4,366 letters were written and 60,695 images tagged—equal to 303,475 keywords on the first day. The pay-outs to workers were $3,055.1 and $3,808.3.

Tables 3.1 show estimated effects of wage on extensive and intensive margins for the two experiments. In each, the first two columns are on extensive margin using a linear probability model and a logit model, the third and fourth columns are on intensive margin using the sample of workers who completed at least one HIT, and the final column show results when taking account of censoring at zero and above using all workers. Each table shows results using wage and log wage separately.

Table 3.2 shows the elasticities from the different models. One of the characteristics

---

long period between the aborted attempt and the final run was partly because of the time required to design and run load testing programs for the servers and partly to minimize contamination between the aborted run and the final experiment.

Table 3.1: Effects of Wage on Extensive and Intensive Margins

	Extensive		Intensive	
	Worked = 1, not = 0		Number of HITs Performed	
	LPM	Logit	Censored	
Sample	Full[a]	Full[a]	Worked[c,e]	Full[d,f]
Image Tagging Experiment—Zero HIT Cut-off				
Log wage	0.048***	0.220***	2.820***	3.277***
	(0.011)	(0.049)	(0.514)	(0.484)
Observations	4,311	4,311	1,605	4,311
Mean of dependent variable	0.372	0.372	7.6	2.8
Image Tagging Experiment—One HIT Cut-off				
Log wage	0.050***	0.315***	3.877***	5.713***
	(0.009)	(0.058)	(0.841)	(0.852)
Observations	4,311	4,311	958	4,311
Mean of dependent variable	0.222	0.222	11.996	2.816
Letter Writing Experiment—Zero HIT Cut-off				
Log wage	0.073***	0.400***	4.962***	6.175***
	(0.014)	(0.076)	(1.074)	(0.934)
Observations	2,111	2,111	578	2,111
Mean of dependent variable	0.274	0.274	7.6	2.1
Letter Writing Experiment—One HIT Cut-off				
Log wage	0.057***	0.496***	8.198***	11.062***
	(0.011)	(0.099)	(1.907)	(1.859)
Observations	2,111	2,111	326	2,111
Mean of dependent variable	0.154	0.154	12.620	2.068

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%. Effects of other job characteristics controlled for, but not shown. See **(author?)** [202] for full results.

[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.

[c] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 1,605 workers who worked on the first day they were observed, 92 were right-censored observations.

[d] Of the 4,311 observations, 2,706 were left-censored observations, 1,513 uncensored observations, and 92 right-censored observations.

[e] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 578 workers, 68 were right-censored.

[f] Of the 2,111 observations, 1,533 were left-censored observations, 510 uncensored observations, and 68 right-censored observations.

Table 3.2: Wage Elasticities of Labor Supply

	Image tagging		Letter writing	
	0 HIT[a]	1 HIT[b]	0 HIT[a]	1 HIT[b]
Extensive margin	0.13	0.23	0.27	0.37
Intensive margin, conditional on working[c]	0.37	0.32	0.65	0.65
Intensive margin, all workers	1.17	2.03	2.94	5.35

Note. Each elasticity calculated at the mean value of the dependent variable.
[a] Based on results from Table 3.1 with zero HITs as cut-off.
[b] Based on results from Table 3.1 with one HITs as cut-off.
[c] All elasticities based on the censored regression models for the sample of workers who worked on the experiment.

of the Mechanical Turk labor market is the low adjustment cost.[13] If a worker dislike the offered set of wages and job characteristics it is easy to move on to another job. Workers do, however, still have imperfect information about the job. They may, for example, not know how long it takes to complete a HIT, and therefore not know whether it is worthwhile to work on it. This suggests that many workers will complete one HIT and then decide whether they want to continue working. We therefore also show estimated elasticities using one HIT as the cut-off point. Extensive margin results are for the probability of working more than one HIT, and intensive margin results are the number of HIT completed minus one.

The extensive margin wage elasticities are 0.13 for the image tagging experiment and 0.27 for the letter writing experiment. For comparison, workers on an archeological dig in Syria in the 1930s were found to have an extensive margin elasticity of 0.035, a labor market experiment in Malawi showed an elasticity of around 0.15, whereas stadium vendors in the US had an elasticity of between 0.55 and 0.65 [19, 96, 198].

Although our extensive margin elasticities are in the range of published results they are lower than other U.S. extensive margin elasticities; both **(author?)** [198] and **(author?)** [47] find elasticities that are more than twice as large as ours. Three features of the Mechanical Turk labor market combine to explain our lower estimates. First, workers do not commit to

[13] See **(author?)** [47] for a discussion of how to estimate elasticities in the presence of adjustment costs and frictions.

a particular number of HITs by working one HIT. Second, fixed entry costs for individual jobs are low once a worker is on the Mechanical Turk platform. Finally, workers likely have imperfect information about how attractive an offered combination of pay and job characteristics is, leading many workers to try the job as part of their search for the jobs with sufficiently high return to effort. If workers routinely do a "trial" HIT before deciding whether to continue working we should see much higher elasticities if we ignore the first HIT. The elasticities using one HIT as the cut-off are, indeed, 10 percentage points higher than the pure extensive margin elasticities.

The commonly held view is that intensive margin elasticities are substantially lower than extensive margin elasticities [119]. For the sample of workers who worked intensive margin elasticities are, however, more than twice as large as our extensive margin elasticities—using zero HITs as the cut-off. The intensive margin elasticities are 0.37 for the image tagging experiment and 0.65 for the letter writing experiment.

The standard available labor market data only allows for the calculation of the intensive margin elasticities for those workers who are working. We, however, have offered wages for all workers independently of whether they decide to work or not. We can therefore also calculate intensive margin elasticities for the sample of all workers who looked at our offered jobs. These intensive margin elasticities are large. The lowest is for the image tagging experiment with zero HITs as the cut-off and even that is 1.2. The equivalent one for the letter writing experiment is almost 3, and the largest is for the one HIT cut-off for the same experiment at 5.4.

(author?) [47, p 1009] argues that the commonly held view that the extensive margin elasticities are larger than intensive margin elasticities may simply be because of frictions and that "[i]n steady state, the intensive elasticity may actually be larger than extensive elasticities..." That our intensive margin elasticities are larger than our extensive margin elasticities supports his argument and, indirectly, our conjecture that the Mechanical Turk labor market is a low friction labor market. Although our elasticities are not directly comparable to other published numbers given the low entry and exit costs of the individual

Mechanical Turk jobs and that a worker does not commit to a particular number of HITs when entering, our results do suggest that workers on Mechanical Turk behave as we would expect given standard labor economics models and that the Mechanical Turk labor market is close to the standard neo-classical model of labor supply.

3.4.1 Worker Characteristics and Their Effects

Tables 3.3 and 3.4 show extensive and intensive margins results, controlling for whether we have observed a worker before and when. Our earliest experiments on Mechanical Turk ran in September 2010 and January 2011 [248]. Of the workers in those experiments, only 45 show up among workers who looked at our letter writing experiment in March 2014, and only 55 show up at our image tagging experiment in November 2014. We therefore combine these workers with workers we first observed at another experiment that ran in June 2013 to form the dummy variable "First observed June 2013 or before", which has a total of 205 workers in the letter writing experiment and 235 workers in the image tagging experiment. Finally, the letter writing experiment ran in March 2014 and in April 2014 we had an initial run of the image tagging experiment that was aborted within hours of starting because of server load issues. A total of 439 workers in the image tagging experiment where first observed in one of these two experiments. Hence, only about 10% of the workers in the letter writing experiment and 15% of the workers in the image tagging experiment were workers we had seen before.

Controlling for experience on Mechanical Turk does little to change the estimated elasticities. Hence, less experienced workers behave as expected and show positive and substantial labor supply elasticities. If anything it appears that less experienced workers show a stronger response to wage changes than workers we have observed before.

3.4.2 Worker Fixed Effects Results

Table 3.5 shows fixed effects estimates for both extensive and intensive margin. The selection over time shows up in the larger mean of the higher percentages of people who work compared

Table 3.3: Effects of Job Characteristics on Extensive and Intensive Margins for Letter Writing Experiment Controlling for Experience on Mechanical Turk

	Without Interactions			With Interactions		
	Extensive Worked = 1	Intensive HITs Performed		Extensive Worked = 1	Intensive HITs Performed	
	LPM	Censored		LPM	Censored	
Sample	Full[a]	Worked[b]	Full[c]	Full[a]	Worked[b]	Full[c]
First observed June	−0.007	2.409	0.175	0.044	−1.448	1.629
2013 or before	(0.032)	(2.339)	(2.058)	(0.080)	(5.881)	(5.020)
Log wage	0.073***	4.960***	6.173***	0.075***	5.364***	6.462***
	(0.014)	(1.073)	(0.935)	(0.014)	(1.122)	(0.979)
Log wage				−0.006	−4.801	−2.497
× observed 2013				(0.049)	(3.716)	(3.257)
Observations	2,111	578	2,111	2,111	578	2,111
Dependent variable mean	0.274	7.6	2.1	0.274	7.6	2.1
Elasticities at mean values	0.266	0.653	2.940	0.274	0.706	3.077

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[b] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 578 workers, 68 were right-censored.
[c] Of the 2,111 observations, 1,533 were left-censored observations, 510 uncensored observations, and 68 right-censored observations.

to the first day analysis. In the panel data, 42 percent of workers work—up from 37 for the first day analysis—and the average number of HITs performed per worker per day is almost twice as large as in the single day data.

Compared to the first day results, the extensive margin results show substantially stronger effects of wage on the likelihood of working. In the linear model, the effect of wage is twice as large in the panel data than in the first day data. For the intensive margin only the linear results are directly comparable with the first day results. Again we see a substantially larger effect of wage, which the estimated effect on number of HITs performed, conditional on working, are 2 to 3 times larger for the fixed effects results than for the first day results.

The fixed effects extensive margin elasticity are just over twice as large as using only the first day data, which puts it close to the one HIT cut-off results. Similarly, the two intensive margin elasticities are also larger than in the first day results. A potential reason they increase less than for the extensive margin may be that the fixed effects estimations do

Table 3.4: Effects of Job Characteristics on Extensive and Intensive Margins for Image Tagging Experiment Controlling for Experience on Mechanical Turk

	Without Interactions			With Interactions		
	Extensive Worked = 1	Intensive HITs Performed		Extensive Worked = 1	Intensive HITs Performed	
	LPM	Censored		LPM	Censored	
Sample	Full[a]	Worked[b]	Full[c]	Full[a]	Worked[b]	Full[c]
First observed June	−0.204***	−1.322	−9.166***	−0.185*	−8.887	−5.355
2013 or before	(0.031)	(1.992)	(1.628)	(0.102)	(6.894)	(5.454)
First observed	−0.226***	2.390	−8.822***	−0.250***	−1.872	−6.107
March/April 2014	(0.024)	(1.535)	(1.220)	(0.075)	(4.594)	(3.822)
Log wage	0.046***	2.807***	3.247***	0.042***	2.858***	3.039***
	(0.011)	(0.515)	(0.485)	(0.011)	(0.532)	(0.511)
Log wage				0.038	−1.123	3.209
× observed 2013				(0.047)	(3.415)	(2.707)
Log wage				0.024	−2.146	1.736
× observed 2014				(0.036)	(2.655)	(1.986)
Observations	4,311	1,605	4,311	4,311	1,605	4,311
Dependent variable mean	0.372	7.6	2.8	0.372	7.6	2.8
Elasticities at mean values	0.124	0.369	1.160	0.113	0.376	1.085

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[b] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment. Of the 1,605 workers who worked on the first day they were observed, 92 were right-censored observations.
[c] Of the 4,311 observations, 2,706 were left-censored observations, 1,513 uncensored observations, and 92 right-censored observations.

not control for censoring they way we did for the frist day results. That said, conditional on working the intensive margin elasticities is 0.562 and using all workers it is 0.832.

3.5 Conclusion

Mechanical Turk is clearly not like "off-line" labor markets. There are no explicit contract, no set working hours, no commuting, and clothing is entirely optional. Is it, however, similar to the market for freelance or independent contractor work. Freelancing, independent contracting, and consulting work is rapidly becoming more and more important in the US economy. A recent estimate is that there are 17.7 million independent workers, making close to USD 1.2 trillion in total income in 2013, and these numbers have been increasing over

Table 3.5: Effects of Wages on Extensive and Intensive Margins—Worker Fixed Effects

	Extensive		Intensive	
	Worked = 1, not = 0		Number of HITs Performed	
Sample	Linear Full[a]	Logit Full[b]	Linear Worked[c]	linear Full
Log wage	0.113***	0.885***	7.365***	4.562***
	(0.009)	(0.076)	(0.552)	(0.259)
Observations	7,954	2,357	3,330	7,954
Number of workers	4,311	719	1,830	4,311
Mean of dependent variable	0.419	0.542	13.095	5.482
Elasticities at mean values	0.270		0.562	0.832

Note. Standard errors in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.
[c] This sample has a higher number of people than than the first day results because there are 125 workers that did not work on the first day they visited the job, but did work on a subsequent day. Hence, the first day number of observations for the intensive margin is 1605, whereas it is 1830 for the fixed effects estimations on intensive margin.

time [184].[14] Finally, Mechanical Turk attracts people actively looking for work, rather than being a sample of undergraduate students participating in a lab experiment.

We find that labor supply elasticities follow the pattern predicted by the standard neoclassical model. Intensive margin elasticities when conditioning on working are about twice as larger as extensive margin elasticities. Because we can observe all workers whether they decide to work or not, we can also estimate intensive margin elasticities for all workers. These elasticities are approximately 10 times what we find for extensive margin elasticities and substantially larger what the previous literature has found.

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

Letter writing experiment

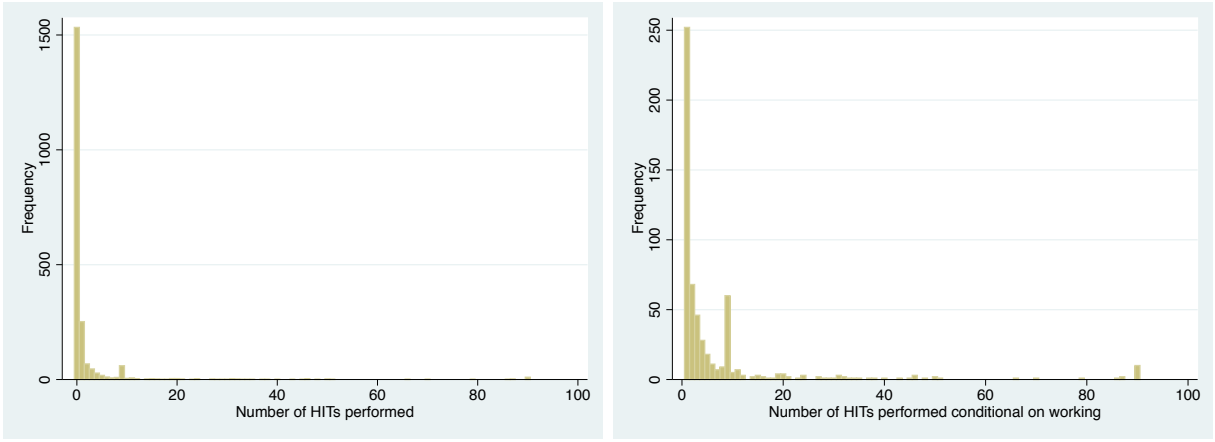


Image tagging experiment

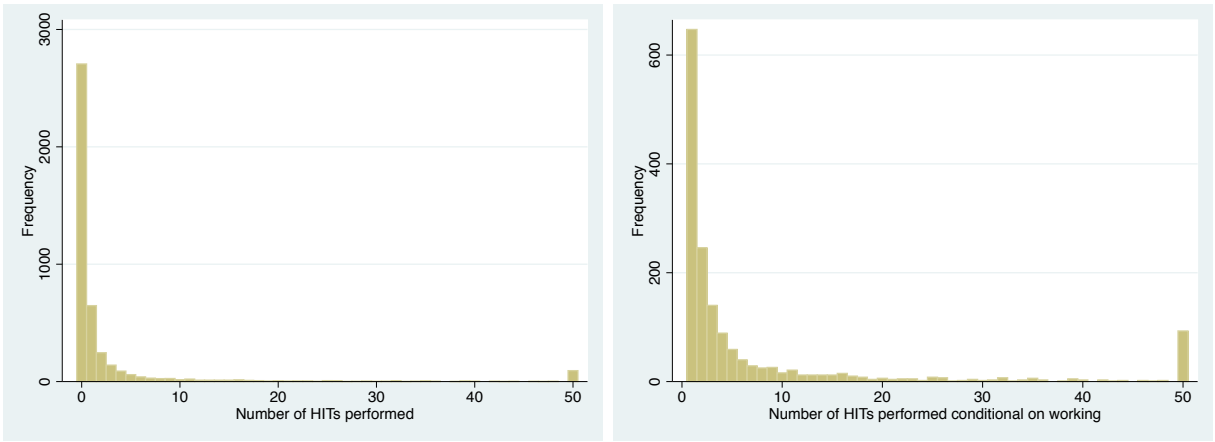


Figure 3.3: Distribution of work done by experiment

BIBLIOGRAPHY

[1] John M Abowd, Patrick Corbel, and Francis Kramarz. The entry and exit of workers and the growth of employment: an analysis of french establishments. *Review of Economics and Statistics*, 81(2):170–187, 1999.

[2] Daron Acemoglu and Andrew F Newman. The labor market and corporate structure. *European Economic Review*, 46(10):1733–1756, 2002.

[3] James D. Adams. Permanent differences in unemployment and permanent wage differentials. *The Quarterly Journal of Economics*, 100(1):29–56, 1985.

[4] Susanne Ackum Agell. Swedish evidence on the efficiency wage hypothesis. *Labour Economics*, 1(2):129–150, 1994.

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

[6] George A Akerlof. Gift exchange and efficiency-wage theory: Four views. *The American Economic Review*, 74(2):79–83, 1984.

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

[8] Armen A Alchian and Harold Demsetz. Production, information costs, and economic organization. *The American economic review*, pages 777–795, 1972.

[9] Joseph G. Altonji. Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy*, 94(3):S176–S215, 1986.

[10] James Andreoni and Charles Sprenger. Risk preferences are not time preferences. *American Economic Review*, pages 3357–3376, 2012.

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

[12] Richard J Arnott and Joseph E Stiglitz. Labor Turnover, Wage Structures, and Moral Hazard: The Inefficiency of Competitive Markets. *Journal of Labor Economics*, 3(4):434–62, October 1985.

[13] Orley Ashenfelter. Measuring the value of a statistical life: Problems and prospects. *The Economic Journal*, 116(510):C10–C23, 2006.

[14] Susan Athey and Scott Stern. An empirical framework for testing theories about complimentarity in organizational design. Technical report, National Bureau of Economic Research, 1998.

[15] Susan Averett, Howard Bodenhorn, and Justas Staisiunas. Unemployment risk and compensating differentials in new jersey manufacturing. *Economic Inquiry*, 43(4):734–749, 2005.

[16] Costas Azariadis. Employment with asymmetric information. *The Quarterly Journal of Economics*, pages 157–172, 1983.

[17] Martin Neil Baily, Eric J Bartelsman, and John Haltiwanger. Labor productivity: structural change and cyclical dynamics. *Review of Economics and Statistics*, 83(3):420–433, 2001.

[18] W. S. Bainbridge. The scientific research potential of virtual worlds. *Science*, 317, 2007.

[19] Tim Barmby and Peter Dolton. What lies beneath? effort and incentives on archaeological digs in the 1930's. Working paper, University of Aberdeen, April 2009.

[20] John M. Barron, Mark C. Berger, and Dan A. Black. Do workers pay for on-the-job training? *The Journal of Human Resources*, 34(2):235–252, 1999.

[21] Timothy J. Bartik. Estimating hedonic demand parameters with single market data: The problems caused by unobserved tastes. *The Review of Economics and Statistics*, 69(1):pp. 178–180, 1987.

[22] Timothy J. Bartik. The estimation of demand parameters in hedonic price models. *Journal of Political Economy*, 95(1):81–88, 1987.

[23] Yoram Barzel. *Economic analysis of property rights*. Cambridge University Press, 1997.

[24] Jeff E. Biddle and Gary A. Zarkin. Worker preference and market compensation for job risk. *The Review of Economics and Statistics*, 70(4):660–667, 1988.

[25] Alan S Blinder and Don H Choi. A shred of evidence on theories of wage stickiness. *The Quarterly Journal of Economics*, 105(4):1003–1015, 1990.

[26] Richard Blundell and Thomas MaCurdy. Labor supply: A review of alternative approaches. In Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, Part A, pages 1559–1695. Elsevier, Amsterdam, 1999.

[27] Richard Blundell, Thomas MaCurdy, and Costas Meghir. Labor supply models: Unobserved heterogeneity, nonparticipation and dynamics. volume 6, Part A of *Handbook of Econometrics*, chapter 69, pages 4667 – 4775. Elsevier, 2007.

[28] Patrick Bolton and Mathias Dewatripont. *Contract theory*. MIT press, 2005.

[29] Stephane Bonhomme and Gregory Jolivet. The pervasive absence of compensating differentials. *Journal of Applied Econometrics*, 24(5):763–795, 2009.

[30] George J. Borjas. *Labor Economics*. McGraw-Hill, New York, NY, seventh edition, 2015.

[31] Samuel Bowles. The production process in a competitive economy: Walrasian, neo-hobbesian, and marxian models. *The American Economic Review*, 75(1):16–36, 1985.

[32] Charles Brown. Equalizing differences in the labor market. *The Quarterly Journal of Economics*, 94(1):113–134, 1980.

[33] Martin Browning and Margaret Irish Angus Deaton. A profitable approach to labor supply and commodity demands over the life-cycle. *Econometrica*, 53(3):503–543, May 1985.

[34] 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.

[35] 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.

[36] James L Butkiewicz, Kenneth J Koford, and Jeffrey B Miller. *Keynes' economic legacy: contemporary economic theories*. Praeger Publishers, 1986.

[37] Guillermo Calvo. Quasi-walrasian theories of unemployment. *The American Economic Review*, 69(2):102–107, 1979.

[38] Guillermo A Calvo and Stanislaw Wellisz. Hierarchy, ability, and income distribution. *The Journal of Political Economy*, pages 991–1010, 1979.

[39] Colin Camerer, Linda Babcock, George Loewenstein, and Richard Thaler. Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441, 1997.

[40] Colin F Camerer and Robin M Hogarth. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1-3):7–42, December 1999.

[41] Colin F Camerer, George Loewenstein, and Matthew Rabin. *Advances in behavioral economics*. Princeton University Press, 2011.

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

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

[44] D. Chandler and A. Kapelner. Breaking monotony with meaning: motivation in crowdsourcing markets. *University of Chicago. Mimeo*, 2010.

[45] Gary Charness. Attribution and reciprocity in an experimental labor market. *Journal of labor Economics*, 22(3):665–688, 2004.

[46] D. Chen and J. Horton. The wages of pay cuts: evidence from a field experiment. *Harvard University, Mimeo*, 2010.

[47] Raj Chetty. Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, 80(3):969–1018, 2012.

[48] Raj Chetty. Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica*, 80(3):969–1018, 05 2012.

[49] Raj Chetty, John N. Friedman, Tore Olsen, and Luigi Pistaferri. Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records. *The Quarterly Journal of Economics*, 126(2):749–804, 2011.

[50] Raj Chetty, Adam Guren, Dayanand S. Manoli, and Andrea Weber. Does indivisible labor explain the difference between micro and macro elasticities? a meta-analysis of extensive margin elasticities. Working Paper 16729, National Bureau of Economic Research, January 2011.

[51] Raj Chetty, Adam Guren, Dayanand S. Manoli, and Andrea Weber. Does Indivisible Labor Explain the Difference Between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities. NBER Working Papers 16729, National Bureau of Economic Research, Inc, January 2011.

[52] Steven NS Cheung. Transaction costs, risk aversion, and the choice of contractual arrangements. *JL & Econ.*, 12:23, 1969.

[53] Steven NS Cheung. The contractual nature of the firm. *Journal of Law and Economics*, pages 1–21, 1983.

[54] Ronald H Coase. The nature of the firm. *economica*, 4(16):386–405, 1937.

[55] Alain Cohn, Ernst Fehr, and Lorenz Goette. Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8):1777–1794, 2014.

[56] 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.

[57] Yves Croissant, Giovanni Millo, et al. Panel data econometrics in r: The plm package. *Journal of Statistical Software*, 27(2):1–43, 2008.

[58] Jean-Pierre Danthine and John B Donaldson. Non-walrasian economies. *Frontiers of business cycle research*, pages 217–242, 1995.

[59] Steven J Davis and John Haltiwanger. Gross job flows. *Handbook of labor economics*, 3:2711–2805, 1999.

[60] Angus Deaton and John Muellbauer. Functional forms for labor supply and commodity demands with and without quantity restrictions. *Econometrica: Journal of the Econometric Society*, pages 1521–1532, 1981.

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

[62] Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. Technical report, National Bureau of Economic Research, 2016.

[63] Jeff DeSimone and Edward J. Schumacher. Compensating wage differentials and aids risk. Working Paper 10861, National Bureau of Economic Research, November 2004.

[64] Peter Doeringer and Michael J Piore. Internal labor markets and manpower adjustment. *New York*, 1971.

[65] David Domeij and Martin Flodén. The labor-supply elasticity and borrowing constraints: Why estimates are biased. *Review of Economic Dynamics*, 9(2):242 – 262, 2006.

[66] Sarah A Donovan, David H Bradley, and Jon O Shimabukuru. What does the gig economy mean for workers? 2016.

[67] Greg J. Duncan and Bertil Holmlund. Was adam smith right after all? another test of the theory of compensating wage differentials. *Journal of Labor Economics*, 1(4):366–379, 1983.

[68] Christian Dustmann and María Engracia Rochina-Barrachina. Selection correction in panel data models: An application to the estimation of females' wage equations. *The Econometrics Journal*, 10(2):263–293, 2007.

[69] B Curtis Eaton and William D White. Agent compensation and the limits of bonding. *Economic inquiry*, 20(3):330–343, 1982.

[70] Randall W. Eberts and Joe A. Stone. Wages, fringe benefits, and working conditions: An analysis of compensating differentials. *Southern Economic Journal*, 52(1):274–280, 1985.

[71] C. C. Eckel and R. K. Wilson. Internet cautions: experimental games with internet partners. *Experimental Economics*, 9:53–66, 2006.

[72] Ivar Ekeland, James J. Heckman, and Lars Nesheim. Identifying hedonic models. *The American Economic Review*, 92(2):304–309, 2002.

[73] Robert F Elliott and Robert Sandy. Adam smith may have been right after all: A new approach to the analysis of compensating differentials. *Economics Letters*, 59(1):127 – 131, 1998.

[74] Moretti Enrico. Local labor markets. *Handbook of labor economics*, 4:1237–1313, 2011.

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

[76] A. Falk and J. J. Heckman. Lab experiments are a major source of knowledge in the social sciences. *Science*, 326, 2009.

[77] Armin Falk. Gift exchange in the field. *Econometrica*, 75(5):1501–1511, 2007.

[78] Henry S. Farber. Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. Working Paper 20604, National Bureau of Economic Research, October 2014.

[79] Henry S. Farber. Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. *The Quarterly Journal of Economics*, 130(4):1975–2026, 2015.

[80] Ernst Fehr, Urs Fischbacher, and Elena Tougareva. Do high stakes and competition undermine fairness? evidence from russia. IEW - Working Papers 120, Institute for Empirical Research in Economics - University of Zurich, 1999.

[81] Ernst Fehr and Lorenz Goette. Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317, 2007.

[82] Ernst Fehr, Erich Kirchler, Andreas Weichbold, and Simon Gächter. When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor economics*, 16(2):324–351, 1998.

[83] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, pages 437–459, 1993.

[84] Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, 42(1):1–34, 1998.

[85] Ronald Fisher. *The Design of Experiments*. Macmillan, 1935.

[86] Robert Forsythe, Joel L. Horowitz, N.E. Savin, and Martin Sefton. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347 – 369, 1994.

[87] James E Foster and Henry Y Wan. Involuntary unemployment as a principal-agent equilibrium. *The American Economic Review*, 74(3):476–484, 1984.

[88] J.K. Galbraith. *The Affluent Society*. A Mariner Book. Houghton Mifflin, 1998.

[89] John E. Garen. Compensating wage differentials and the endogeneity of job riskiness. *Review of Economics and Statistics*, 70:9–16, 1988.

[90] Alex Gershkov and Motty Perry. Dynamic contracts with moral hazard and adverse selection. *The Review of Economic Studies*, 79(1):268–306, 2012.

[91] Robert Gibbons. Four formal (izable) theories of the firm? *Journal of Economic Behavior & Organization*, 58(2):200–245, 2005.

[92] Herbert Gintis and Tsuneo Ishikawa. Wages, work intensity, and unemployment. *Journal of the Japanese and International Economies*, 1(2):195–228, 1987.

[93] Uri Gneezy and John A List. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384, 2006.

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

[95] John H. Goddeeris. Compensating differentials and self-selection: An application to lawyers. *Journal of Political Economy*, 96(2):411–428, 1988.

[96] Jessica Goldberg. Kwacha gonna do? experimental evidence about labor supply in rural malawi. *American Economic Journal: Applied Economics*, 8(1):129–49, 2016.

[97] Jessica A. Goldberg. Kwacha gonna do? experimental evidence about labor supply in rural malawi. *American Economic Journal: Applied Economics*, Forthcoming.

[98] S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John. Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59:93–104, 2004.

[99] Daniel Gottlieb and Humberto Moreira. Simultaneous adverse selection and moral hazard. Technical report, Mimeo, 2013.

[100] Christian Gourieroux, Alberto Holly, and Alain Monfort. Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica: journal of the Econometric Society*, pages 63–80, 1982.

[101] T. Gronberg and W. Reed. Estimating workers' marginal willingness to pay for job attributes using duration data. *Journal of Human Resources*, 29:911–931, 1994.

[102] Erica L Groshen and Alan B Krueger. The structure of supervision and pay in hospitals. *Industrial & Labor Relations Review*, 43(3):134S–146S, 1990.

[103] Roger Guesnerie and Jean-Jacques Laffont. A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics*, 25(3):329 – 369, 1984.

[104] 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.

[105] Robert E Hall. Labor-market frictions and employment fluctuations. *Handbook of macroeconomics*, 1:1137–1170, 1999.

[106] Robert E. Hall. High discounts and high unemployment. *American Economic Review*, 107(2):305–30, February 2017.

[107] Robert E Hall, Aaron Gordon, and Charles Holt. Turnover in the labor force. *Brookings Papers on Economic Activity*, 1972(3):709–764, 1972.

[108] Daniel S Hamermesh and Gerard A Pfann. Adjustment costs in factor demand. *Journal of Economic Literature*, 34(3):1264–1292, 1996.

[109] Daniel S. Hamermesh and John R. Wolfe. Compensating wage differentials and the duration of wage loss. *Journal of Labor Economics*, 8(1):S175–S197, 1990.

[110] Daniel S. Hamermesh and John R Wolfe. Compensating wage differentials and the duration of wage loss. *Journal of Labor Economics*, 8:S175–S197, 1990.

[111] R Lynn Hannan, John H Kagel, and Donald V Moser. Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity

differences, and effort requests on behavior*. *Journal of Labor Economics*, 20(4):923–951, 2002.

[112] Bruce E. Hansen. The new econometrics of structural change: Dating breaks in u.s. labor productivity. *The Journal of Economic Perspectives*, 15(4):117–128, 2001.

[113] John R Harris and Michael P Todaro. Migration, unemployment and development: a two-sector analysis. *The American economic review*, 60(1):126–142, 1970.

[114] Glenn W. Harrison and James C. Lesley. Must contingent valuation surveys cost so much? *Journal of Environmental Economic Management*, 31:79–95, 1996.

[115] Glenn W. Harrison and John A. List. Field experiments. *Journal of Economic Literature*, XLII:1009–1055, 2004.

[116] Joop Hartog and Wim P.M. Vijverberg. On compensation for risk aversion and skewness affection in wages. *Labour Economics*, 14(6):938 – 956, 2007. Education and RiskEducation and Risk S.I.

[117] Jerry A Hausman and David A Wise. Attrition bias in experimental and panel data: the gary income maintenance experiment. *Econometrica: Journal of the Econometric Society*, pages 455–473, 1979.

[118] James J Heckman. Sample selection bias as a specification error (with an application to the estimation of labor supply functions), 1977.

[119] James J. Heckman. What has been learned about labor supply in the past twenty years? *The American Economic Review*, 83(2):116–121, 1993.

[120] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33:62–135, 2010.

[121] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: Behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):pp. 73–78, 2001.

[122] Joni Hersch and W. Kip Viscusi. Cigarette smoking, seatbelt use, and difference in wage-risk tradeoffs. *Journal of Human Resources*, 25:202–227, 1990.

[123] Bengt Holmstrom and Paul Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, pages 24–52, 1991.

[124] Bo E Honoré. Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica: journal of the Econometric Society*, pages 533–565, 1992.

[125] J. Horton. Online labor markets. *Workshop on Internet and network economics*, pages 515–522, 2010.

[126] J. Horton. Online labor markets. *Workshop on Internet and network economics*, pages 515–522, 2010.

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

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

[129] J. Horton and L. Chilton. The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM conference on electronic commerce*, 2010.

[130] J. Horton and L. Chilton. The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM conference on electronic commerce*, 2010.

[131] J. Horton, D. Rand, and R Zeckhauser. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14:399–425, 2011.

[132] J. Horton, D. Rand, and R Zeckhauser. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14:399–425, 2011.

[133] Tzu-Ling Huang, Arnems Hallam, Peter F Orazem, and Eizabeth M Paterno. Empirical tests of efficiency wage models. *Economica*, 65(257):125–143, 1998.

[134] H. Hwang, W. Reed, and C. Hubbard. Compensating differentials and unobserved productivity. *Journal of Political Economy*, 100:835–858, 1992.

[135] Hae-shin Hwang, Dale T Mortensen, and W Robert Reed. Hedonic Wages and Labor Market Search. *Journal of Labor Economics*, 16(4):815–47, October 1998.

[136] Hae-shin Hwang, W. Robert Reed, and Carlton Hubbard. Compensating wage differentials and unobserved productivity. *Journal of Political Economy*, 100(4):835–858, 1992.

[137] Susumu Imai and Michael P. Keane. Intertemporal labor supply and human capital accumulation. *International Economic Review*, 45(2):601–641, 2004.

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

[139] P. Ipeirotis. Mechanical Turk Monitor: Task/HITs/$$$ completed per day. `http://hyperion.stern.nyu.edu/mturk/completed.php`, 2009. Accessed: 9/18/2009.

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

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

[142] Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[143] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on Human Computation (HCOMP 2010)*, pages 64–67. ACM, 2010.

[144] Panos Ipeirotis. The new demographics of mechanical turk. *URL http://behind-the-enemy-lines. blogspot. com/2010/03/new-demographics-of-mechanical-turk. html*, 2010.

[145] Panos Ipeirotis. Crowdsourcing using mechanical turk: quality management and scalability. *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011*, page 1, 2011.

[146] Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):pp. 1449–1475, 2003.

[147] Daniel Kahneman, Jack L Knetsch, and Richard Thaler. Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, pages 728–741, 1986.

[148] Lawrence F Katz. Efficiency wage theories: A partial evaluation. In *NBER Macroeconomics Annual 1986, Volume 1*, pages 235–290. MIT Press, 1986.

[149] Michael Keane and Richard Rogerson. Micro and macro labor supply elasticities: A reassessment of conventional wisdom. *Journal of Economic Literature*, 50(2):464–76, 2012.

[150] Michael Keane and Richard Rogerson. Reconciling micro and macro labor supply elasticities: A structural perspective. *Annual Review of Economics*, 7(1):89–117, 2015.

[151] Michael P. Keane. Labor supply and taxes: A survey. *Journal of Economic Literature*, 49(4):961–1075, 2011.

[152] Allan G. King. Occupational choice, risk aversion, and wealth. *Industrial and Labor Relations Review*, 27(4):586–596, 1974.

[153] Thomas J Kniesner and John D Leeth. Hedonic wage equilibrium: Theory, evidence and policy. Discussion Paper 5076, IZA, Bonn, Germany, July 2010.

[154] Thomas J. Kniesner, W. Kip Viscusi, Christopher Woock, and James P. Ziliak. How unobservable productivity biases the value of a statistical life. Working Paper 11659, National Bureau of Economic Research, October 2005.

[155] Thomas J Kniesner, W Kip Viscusi, Christopher Woock, and James P Ziliak. The value of a statistical life: Evidence from panel data. *Review of Economics and Statistics*, 94(1):74–87, 2012.

[156] Frank H Knight. *Risk, uncertainty and profit*. Courier Corporation, 1922.

[157] Peter F. Kostiuk. Compensating differentials for shift work. *Journal of Political Economy*, 98(5):1054–1075, 1990.

[158] 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.

[159] Sebastian Kube, Michel André Maréchal, and Clemens Puppe. The currency of reciprocity: Gift exchange in the workplace. *The American Economic Review*, 102(4):1644–1662, 2012.

[160] S. Kuznetsov. Motivations of contributors to Wikipedia. *ACM SIGCAS Computers and Society*, 36(2), 2006.

[161] Ekaterini Kyriazidou. Estimation of a panel data sample selection model. *Econometrica: Journal of the Econometric Society*, pages 1335–1364, 1997.

[162] Jean-Jacques Laffont and David Martimort. *The theory of incentives: the principal-agent model*. Princeton university press, 2009.

[163] P Richard G Layard, Stephen J Nickell, and Richard Jackman. *Unemployment: macroeconomic performance and the labour market*. Oxford University Press on Demand, 2005.

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

[165] Thomas Lemieux, W. Bentley MacLeod, and Daniel Parent. Contract form, wage flexibility, and employment. *American Economic Review*, 102(3):526–31, May 2012.

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

[167] S. D. Levitt and J. A. List. Field experiments in economics: the past, the present, and the future. *European Economic Review*, 53:1–18, 2009.

[168] Steven D. Levitt and John A. List. Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1 – 18, 2009.

[169] Tracy R Lewis and David E.M Sappington. Countervailing incentives in agency problems. *Journal of Economic Theory*, 49(2):294 – 313, 1989.

[170] Elizabeth H. Li. Compensating differentials for cyclical and noncyclical unemployment: The interaction between investors' and employees' risk aversion. *Journal of Labor Economics*, 4(2):277–300, 1986.

[171] Assar Lindbeck, Dennis J Snower, et al. The insider-outsider theory of employment and unemployment. *MIT Press Books*, 1, 1989.

[172] John A. List. The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, 119:49–89, 2004.

[173] John A. List and Imran Rasul. Chapter 2 - field experiments in labor economics. volume 4, Part A of *Handbook of Labor Economics*, pages 103 – 228. Elsevier, 2011.

[174] Stuart A. Low and Lee R. McPheters. Wage differentials and risk of death: An empirical analysis. *Economic Inquiry*, 21:271–280, 1983.

[175] W Bentley MacLeod. Can contract theory explain social preferences? *The American economic review*, 97(2):187–192, 2007.

[176] W Bentley MacLeod and James M Malcomson. Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica: Journal of the Econometric Society*, pages 447–480, 1989.

[177] W Bentley MacLeod, James M Malcomson, and Paul Gomme. Labor turnover and the natural rate of unemployment: efficiency wage versus frictional unemployment. *Journal of Labor Economics*, 12(2):276–315, 1994.

[178] Thomas E. MaCurdy. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy*, 89(6):1059–1085, 1981.

[179] James M Malcomson. Unemployment and the efficiency wage hypothesis. *The Economic Journal*, 91(364):848–866, 1981.

[180] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[181] William D. Marder and Douglas E. Hough. Medical residency as investment in human capital. *The Journal of Human Resources*, 18(1):49–64, 1983.

[182] W. Mason, D. J. Watts, , and S. Suri. Conducting behavioral research on amazon's mechanical turk. *SSRN eLibrary*, 2010.

[183] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *Proceedings of the ACM SIGKDD workshop on human computation*, pages 77–85, 2009.

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

[185] Jacob Mincer and Yoshio Higuchi. Wage structures and labor turnover in the united states and japan. *Journal of the Japanese and International Economies*, 2(2):97–133, 1988.

[186] Hajime Miyazaki. Workers norms and involuntary unemployment. *The Quarterly Journal of Economics*, pages 297–311, 1984.

[187] Enrico Moretti. Do wages compensate for risk of unemployment? parametric and semiparametric evidence from seasonal jobs. *Journal of Risk and Uncertainty*, 20(1):45–66, 2000.

[188] Enrico Moretti. Local labor markets. *Handbook of labor economics*, 4:1237–1313, 2011.

[189] Dale T Mortensen. Job search and labor market analysis. *Handbook of labor economics*, 2:849–919, 1986.

[190] Dale T Mortensen and Christopher A Pissarides. New developments in models of search in the labor market. *Handbook of labor economics*, 3:2567–2627, 1999.

[191] Giuseppe Moscarini and Fabien Postel-Vinay. The contribution of large and small employers to job creation in times of high and low unemployment. *American Economic Review*, 102(6):2509–39, May 2012.

[192] Giuseppe Moscarini and Fabien Postel-Vinay. Wage posting and business cycles. *American Economic Review*, 106(5):208–13, May 2016.

[193] Derek Neal. Industry-specific human capital: Evidence from displaced workers. *Journal of labor Economics*, pages 653–677, 1995.

[194] David M. Newbery and Joseph E. Stiglitz. Wage rigidity, implicit contracts, unemployment and economic efficiency. *The Economic Journal*, 97(386):pp. 416–430, 1987.

[195] Douglass C North. *Institutions, institutional change and economic performance*. Cambridge university press, 1990.

[196] Chris Nosko and Steven Tadelis. The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research, 2015.

[197] O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.

[198] Gerald S. Oettinger. An empirical analysis of the daily labor supply of stadium venors. *Journal of Political Economy*, 107(2):360–392, 1999.

[199] A. Pallais. Inefficient hiring in entry-level labor markets. 2010.

[200] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 2010.

[201] Pierre Picard. *Wages and unemployment: a study in non-Walrasian macroeconomics*. Cambridge University Press, 1993.

[202] Claus C Pörtner, Nail Hassairi, and Michael Toomim. Only if you pay me more: Field experiments support compensating wage differentials theory. Working paper, Seattle University, September 2015.

[203] David Powell and Hui Shan. Income taxes, compensating differentials, and occupational choice: How taxes distort the wage-amenity decision. *American Economic Journal: Economic Policy*, 4(1):224–47, 2012.

[204] Canice Prendergast. The provision of incentives in firms. *Journal of economic literature*, 37(1):7–63, 1999.

[205] Canice Prendergast. What trade-off of risk and incentives? *The American Economic Review*, 90(2):421–425, 2000.

[206] Lant Pritchett, Justin Sandefur, et al. Learning from experiments when context matters. *American Economic Review*, 105(5):471–75, 2015.

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

[208] Matthew Rabin. Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302, 1993.

[209] Daniel MG Raff. Wage determination theory and the five-dollar day at ford. *The Journal of Economic History*, 48(02):387–399, 1988.

[210] 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.

[211] James B Rebitzer and Lowell J Taylor. A model of dual labor markets when product demand is uncertain. *The Quarterly Journal of Economics*, 106(4):1373–1383, 1991.

[212] 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.

[213] James B. Rebitzer and Lowell J. Taylor. Chapter 8 - extrinsic rewards and intrinsic motives: Standard and behavioral approaches to agency and labor markets. volume 4, Part A of *Handbook of Labor Economics*, pages 701 – 772. Elsevier, 2011.

[214] Jennifer Roback. Wages, rents, and the quality of life. *Journal of Political Economy*, 90(6):pp. 1257–1278, 1982.

[215] Richard Rogerson, Robert Shimer, and Randall Wright. Search-theoretic models of the labor market: A survey. *Journal of Economic Literature*, 43(4):959–988, 2005.

[216] Sherwin Rosen. Learning and experience in the labor market. *The Journal of Human Resources*, 7(3):326–342, 1972.

[217] Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.

[218] Sherwin Rosen. The economics of superstars. *The American Economic Review*, 71(5):845–858, 1981.

[219] Sherwin Rosen. The Theory of Equalizing Differences. In Orley C Ashenfelter and Richard Layard, editors, *Handbook of Labor Economics*, volume 1, chapter 12, pages 641–692. Elsevier, 1986.

[220] Sherwin Rosen. The theory of equalizing differences. volume 1 of *Handbook of Labor Economics*, chapter 12, pages 641–692. Elsevier, 1986.

[221] P. Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[222] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[223] Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):pp. 97–109, 1982.

[224] E. Elisabet Rutström. Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory*, 27:427–41, 1998.

[225] Bernard Salanié. Testing contract theory. *CESifo Economic Studies*, 49(3):461–477, 2003.

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

[227] Ekkehart Schlicht. Labour turnover, wage structure, and natural unemployment. *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics*, pages 337–346, 1978.

[228] Wendelin Schnedler. Incentives and misdirected effort. 2015.

[229] D. O. Sears. College sophomores in the lab: Influences of narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51:515–530, 1986.

[230] Avner Shaked and John Sutton. Involuntary unemployment as a perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 1351–1364, 1984.

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

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

[233] Carl Shapiro and Joseph E Stiglitz. Equilibrium Unemployment as a Worker Discipline Device: Reply. *American Economic Review*, 75(4):892–93, September 1985.

[234] Robert Shimer. The cyclical behavior of equilibrium unemployment and vacancies. *The American Economic Review*, 95(1):25–49, 2005.

[235] Robert Shimer. Reassessing the ins and outs of unemployment. *Review of Economic Dynamics*, 15(2):127–148, 2012.

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

[237] Vernon L. Smith. An experimental study of competitive market behavior. *The Journal of Political Economy*, 70(2):111–137, 1962.

[238] Vernon L Smith. Effect of market organization on competitive equilibrium. *The Quarterly Journal of Economics*, pages 182–201, 1964.

[239] Joseph E Stiglitz. Alternative theories of wage determination and unemployment in ldc's: The labor turnover model. *The Quarterly Journal of Economics*, pages 194–227, 1974.

[240] Joseph E Stiglitz. The theory of screening, education, and the distribution of income. *The American Economic Review*, 65(3):283–300, 1975.

[241] Joseph E Stiglitz. Prices and queues as screening devices in competitive markets. *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn, D. Gale and O. Hart (eds.), Cambridge: MIT Press*, pages 128–166, 1976.

[242] Joseph E Stiglitz. Theories of wage rigidity, 1984.

[243] Paul Sullivan and Ted To. Job dispersion and compensating wage differentials. *US Bureau of Labor Statistics*, 2013.

[244] Knut Sydsæter, Peter Hammond, and Atle Seierstad. *Further mathematics for economic analysis*. Pearson education, 2008.

[245] Richard Thaler and Sherwin Rosen. The value of saving a life: Evidence from the labor market. In Nestor E. Terleckyj, editor, *Household Production and Consumption*, pages 265–302. NBER, 1976.

[246] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

[247] Michael P Todaro. A model of labor migration and urban unemployment in less developed countries. *The American economic review*, 59(1):138–148, 1969.

[248] Michael Toomim, Travis Kriplean, Claus C Pörtner, and James A Landay. Utility of Human-Computer Interactions: Toward a Science of Preference Measurement. In

Proceedings of the 2011 annual conference on Human factors in computing systems, pages 2275–2284. ACM, 2011.

[249] Stephen J. Trejo. The effects of overtime pay regulation on worker compensation. *The American Economic Review*, 81(4):719–740, 1991.

[250] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[251] Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061, 1991.

[252] Ernesto Villanueva. Estimating compensating wage differentials using voluntary job changes: Evidence from germany. *Industrial and Labor Relations Review*, 60(4):544–561, 2007.

[253] W. Kip. Viscusi. Wealth effects and earnings premiums for job hazards. *Review of Economics and Statistics*, 60:408–416, 1978.

[254] W. Kip Viscusi. The value of risks to life and health. *Journal of Economic Literature*, 31(4):1912–1946, 1993.

[255] W. Kip Viscusi and Michael J. Moore. Workers' compensation: Wage effects, benefit inadequacies, and the value of health losses. *Review of Economics and Statistics*, 69:249–261, 1987.

[256] W.Kip Viscusi and JosephE. Aldy. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, 27(1):5–76, 2003.

[257] 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.

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

[259] Andrew Weiss. *Efficiency wages: Models of unemployment, layoffs, and wage dispersion*. Princeton University Press, 2014.

[260] Yoram Weiss. The determination of life cycle earnings: A survey. volume 1 of *Handbook of Labor Economics*, chapter 11, pages 603 – 640. Elsevier, 1986.

[261] O.E. Williamson. *Markets and hierarchies, analysis and antitrust implications: a study in the economics of internal organization*. A study in the economics of internal organization. Free Press, 1975.

[262] Jeffrey Winking and Nicholas Mizer. Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, 34(4):288–293, 2013.

[263] Jeffrey M Wooldridge. Cluster-sample methods in applied econometrics. *The American Economic Review*, 93(2):133–138, 2003.

[264] C.F.J. Wu and M.S. Hamada. *Experiments: Planning, Analysis, and Optimization*. Wiley Series in Probability and Statistics. Wiley, 2011.

[265] Janet L Yellen. Efficiency wage models of unemployment. *The American Economic Review*, pages 200–205, 1984.

[266] Ayegl ahin, Joseph Song, Giorgio Topa, and Giovanni L. Violante. Mismatch unemployment. *American Economic Review*, 104(11):3529–64, November 2014.

Appendix A

ADDITIONAL RESULTS AND ROBUSTNESS EXERCISES RELATED TO THE FIRST CHAPTER

A.1 Heterogeneity in Response to Incentives

The main analysis has been performed using the worker-HIT panel data. The outcome varies HIT by HIT, while the treatments (wage offered) vary day by day. It may be interesting then to compare the results obtained using the worker-HIT panel with results obtained from a worker-day panel. It should be stated, however, that while experimental treatments are not changing on a HIT by HIT basis, some of the controls are – experience for one (which may induce either fatigue or learning). Another control varying by HIT is $HITsLeft$, the estimate of how long the worker plans to keep working on the HITs overall. The shirking model operates on the idea that there is a stick and a carrot, with termination being the stick. If worker plans to quit in the near future, this stick loses its potency. The variable $HITsLeft$ helps to keep track of this varying stick potency.

Results from Table A.1 are a bit puzzling as far as the coefficient on $log(wage)$ is concerned – .094. How is this possible if the worker-HIT analysis shows a coefficient estimate of the opposite sign? By design, the worker-HIT panel overweighs workers who do more work and the worker-day panel overweighs workers who do less work. The worker-HIT analysis also shows that workers' effort goes down with their tenure. Figure 1.8 showed us that there is great heterogeneity in tenure and the worker-HIT analysis helps control for this by using not only $totalHITs$, but also $HITsLeft$, which change HIT by HIT. The worker-day analysis is not able to control for these dynamic effects, resulting in an outsize effect of workers with short tenures, who are not affected by the stick of job termination, since they are not planning to remain on the job for very long in the first place.

Table A.1: Worker Day Data - Regression of Effort on Wages and reservation wages

	Dependent variable:	
	ratings	
	censored regression	*OLS*
	(1)	(2)
log(wage)	-0.094^{**}	0.008
	(0.041)	(0.028)
log(minAcceptedWage)	0.115^{***}	0.040
	(0.038)	(0.025)
successRate	-0.001	-0.001
	(0.001)	(0.001)
disagreeable	-0.025^{***}	-0.015^{***}
	(0.001)	(0.0004)
training	-0.055	-0.035
	(0.042)	(0.028)
logSigma	0.063^{***}	
	(0.017)	
Constant	6.088^{***}	5.200^{***}
	(0.117)	(0.076)
Observations	3,135	3,135
R^2		0.291
Adjusted R^2		0.289
Log Likelihood	$-3{,}584.496$	
Akaike Inf. Crit.	7,182.991	
Bayesian Inf. Crit.	7,225.344	
Residual Std. Error		0.783 (df = 3129)
Note:	*p<0.1; **p<0.05; ***p<0.01	

This would explain why there would be zero correlation between wages and performance but why is the coefficient estimate on $log(wage)$ negative? 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.

A.2 Opportunity Costs over Time

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.

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. Reservation wage measure so constructed was still positively correlated with the outcome, however, inclusion of this time data seems to have biased the coefficient on $log(wage)$ in the regression, pushing it into negative territory. Appendix A.3 elaborates on why it may not be a good idea to include time in the regression as an independent variable.

Given that not only wages were manipulated but also job attributes, it would be better if the opportunity cost measure reflected this. The minimum hourly wage measure would incorporate this, as working negative job attributes would be reflected in the time it takes to get the job done (as confirmed by the results in Table A.4).

Let us see whether wage of accepted HITs change over the course of the days of the experiment by using the following specification:

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

Results from this regression are presented in Table A.3. We see some changes on day2 but the size of this effect is less than 1 percent of the baseline.

What would happen to the results if reservation was not constant? How much would we imagine it could vary? And would it vary in a continuous fashion, generating an auto-correlated series? If so, using the measure used in this paper would constitute a proxy, a

Table A.2: Regression of Effort on Wages, minimum hourly wages, and tenure

	Dependent variable:	
	ratings	
	(1)	(2)
log(wage)	-0.330^{***}	0.086^{***}
	(0.041)	(0.026)
log(min_hourly_wage)	0.517^{***}	
	(0.033)	
log(min_accepted_wage)		0.195^{***}
		(0.021)
HITsLeft	0.0003	0.001^{***}
	(0.0003)	(0.0003)
HITsDone	0.001^{***}	0.003^{***}
	(0.0003)	(0.0003)
success_rate	-0.003^{***}	-0.003^{***}
	(0.001)	(0.001)
disagreeable	-0.039^{***}	-0.039^{***}
	(0.0004)	(0.0004)
day	0.057^{***}	0.070^{***}
	(0.010)	(0.010)
training	0.091^{***}	0.060^{**}
	(0.025)	(0.025)
time_on_hit	0.0004^{***}	
	(0.0001)	
logSigma	0.664^{***}	0.674^{***}
	(0.007)	(0.007)
Constant	6.239^{***}	7.939^{***}
	(0.138)	(0.090)
Observations	41,963	41,963
Log Likelihood	−38,684.270	−38,883.240
Akaike Inf. Crit.	77,390.540	77,786.480
Bayesian Inf. Crit.	77,485.630	77,872.920
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A.3: 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)
Adjusted R^2	0.002
Note:	*p<0.1; **p<0.05; ***p<0.01

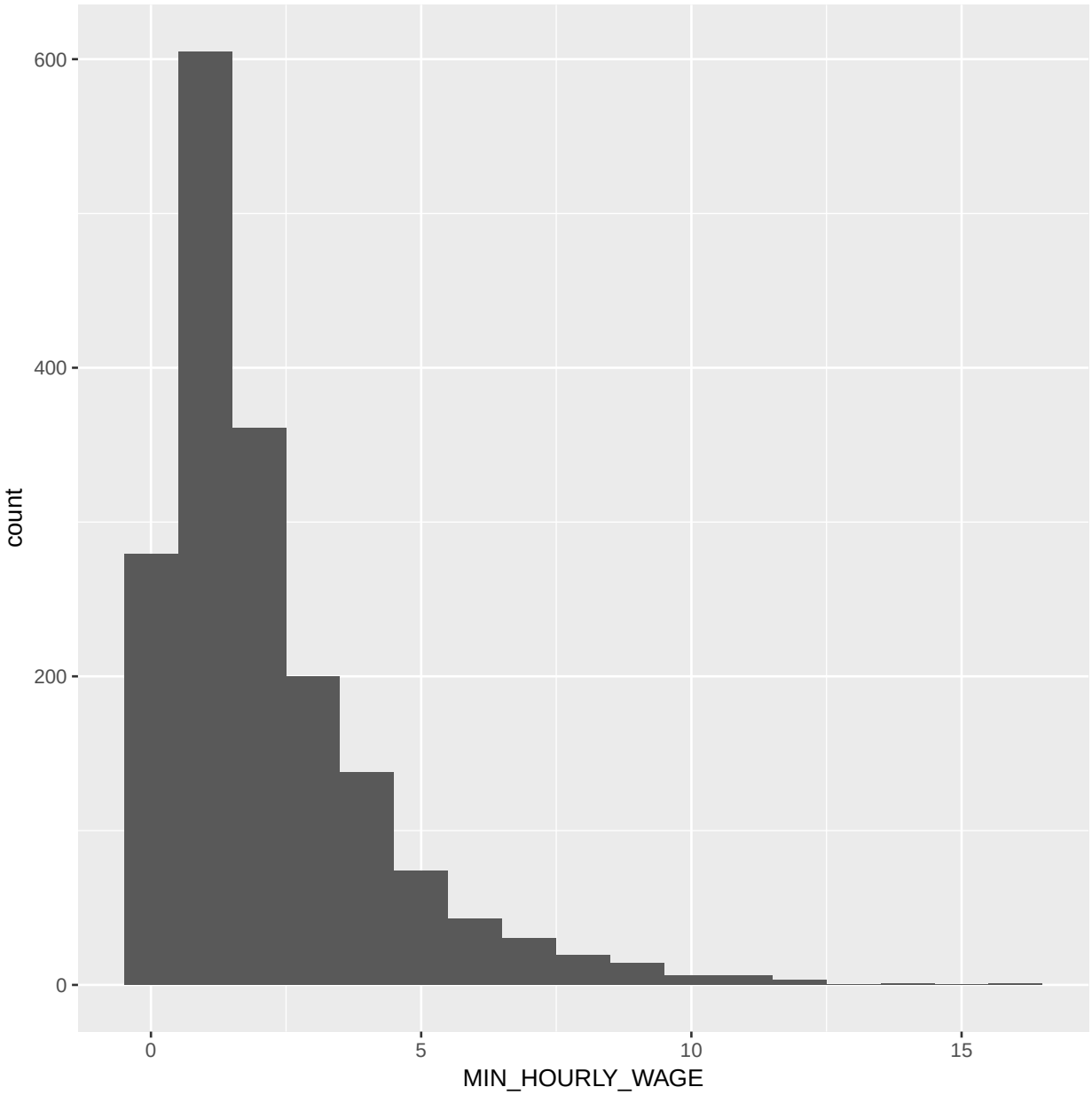


Figure A.1: Histogram of Minimum Hourly Wages

single observation from such a time series of opportunity costs. This would introduce noise, however, would not lead to a systematic bias.

A.3 Analysis Using Time

The nature of the experimental design allowed for the collection of time stamps for individual events, such as HIT submit, page load, etc. It is then possible to construct some measure of "time on HIT." The usefulness of such a measure, however, is not obvious. First, this would likely be a noisy measure, since a worker could be simultaneously working and reading the news on another page, or taking a snack-time break. These would be hard to detect in the data. Some outlier observations in the time data suggest that this is not a hypothetical concern (note the long tail in Figure A.2).

Additionally, it is not obvious what the time on HIT stands for. Arguments can be made that time on HIT is a measure of:

1. effort

2. productivity, or

3. fatigue

These possibilities are in conflict with each other, making the inclusion of the time data in the analysis complicated. This is illustrated in the regression using the $log(timeOnHIT)$ as a measure of effort in Table A.4. The wage elasticity of $log(timeOnHIT)$ is .069, so higher wages offered lead to more time spent on HIT. This results is analogous to the result obtained with the *ratings* as a measure of effort. However, unlike with that analysis, the coefficient on the $log(minAcceptedWage)$ is negative. This would only make sense if time would be a measure of productivity, rather than effort. The fact that $log(timeOnHIT)$ falls with tenure could interpreted both in the context of measure of productivity or fatigue. For these reasons, this variable has been excluded from the main analysis. Time has not even been used as a control, since it really is an outcome (a choice variable of the agent) and as such would likely have common components in its residual with the *ratings* measure of effort, resulting in a correlation between $log(timeOnHIT)$ and the error term.

Table A.4: Regression of Effort on Wages, reservation wages, and tenure

	Dependent variable:
	log(time_on_hit)
log(wage)	0.069^{***}
	(0.010)
HITsLeft	-0.002^{***}
	(0.0002)
log(min_accepted_wage)	-0.087^{***}
	(0.006)
HITsDone	-0.003^{***}
	(0.0001)
success_rate	-0.0002
	(0.0002)
disagreeable	0.001^{***}
	(0.0001)
day	-0.034^{***}
	(0.003)
training	0.069^{***}
	(0.007)
log(wage):HITsLeft	-0.001^{***}
	(0.0001)
Constant	5.452^{***}
	(0.024)
Observations	41,963
R^2	0.063
Adjusted R^2	0.063
Residual Std. Error	0.680 (df = 41953)
F Statistic	315.127^{***} (df = 9; 41953)
Note:	*p<0.1; **p<0.05; ***p<0.01

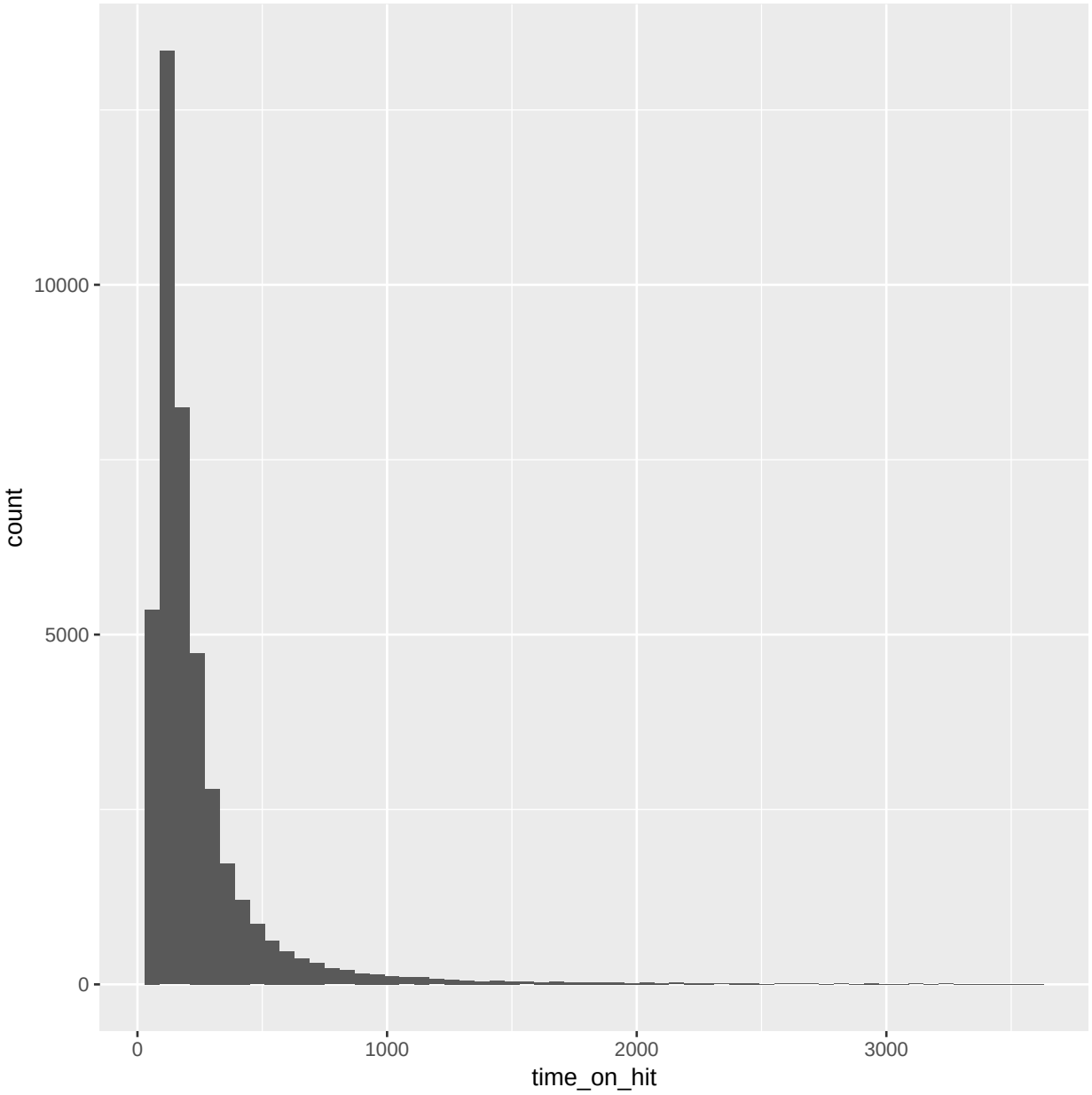


Figure A.2: Histogram of time on HIT (seconds)

A.4 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 ([11]). 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 ([94]). 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.

VITA

Nail Hassairi is a Research Scientist with Childcare Quality and Early Learning Center for Research and Professional Development at the University of Washington, following his completion of the PhD in Economics at the University of Washington. His main research interests are political economics, labor economics, behavioral economics, and experimental economics.