Only if You Pay Me More: Field Experiments Support Compensating Wage Differentials Theory[*]

Claus C Pörtner
Department of Economics
Albers School of Business and Economics
Seattle University, P.O. Box 222000
Seattle, WA 98122
work@clausportner.com
www.clausportner.com
&
Center for Studies in Demography and Ecology
University of Washington

Nail Hassairi
Department of Economics
University of Washington

Michael Toomim
Unaffiliated

October 2015

[*] This research was conducted under IRB approval from University of Washington and Seattle University. This research was financed by National Science Foundation award number 1110965, "Interaction Economics: Instruments that Measure Social-Computational Systems." Partial support for this research also came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, 5R24HD042828, to the Center for Studies in Demography and Ecology at the University of Washington. We would like to thank Shamma A. Alam for research assistance. We would also like thank Yoram Barzel, Niels-Hugo Blunch, Lawrence Kahn, Brian Kelly, Terra McKinnish, David Ribar, and seminar participants at the DAEiNA workshop, Society of Labor Economists' annual meeting, Seattle University, University of Colorado Boulder, and University of Washington for helpful comments and discussions.

Abstract

Compensating wage differentials is Adam Smith's idea that wage differences equalize differences in job and worker characteristics. 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. We run experiments in an online labor market, randomizing offered pay and job characteristics, thereby overcoming both problems. We find, as predicted by our 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, supporting the theory.

JEL codes: J3, J2

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 counterbalance 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 Smith (1776, Book I, Chapter X, Part I)

The theory of compensating wage differentials originates with Adam Smith's idea that observed wage differences equalize differences in the characteristics of both work and workers. It is one of the central tenets of labor economics, the value of statistical life (VSL) literature, and urban economics (Rosen, 1986; Viscusi and Aldy, 2003; Roback, 1982). 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]

The main use of the compensating wage differentials idea has been as the basis of a theory of labor supply to jobs differentiated by various job characteristics (Rosen, 1986). Workers treat job characteristics as consumption goods and trade off between wage and job amenities. A company offering favorable working conditions can pay a lower wage than a company offering less favorable working conditions, and the difference in pay is then a measure of how much workers value the difference in working conditions.

However, attempts to estimate the "price" associated with different job characteristics often fail; the only job characteristic consistently found to affect wages is risk of death (Rosen,

[1] Ashenfelter (2006) discusses how speed limits are set. Powell and Shan (2012) analyze the effect of taxation on distortions of the wage-job amenity trade-off.

1986).[2] The basic econometric problem is that workers self-select into specific jobs based on unobservable worker or job characteristics (Brown, 1980). 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, Brown 1980, Duncan and Holmlund 1983, Hwang, Reed, and Hubbard 1992, Kniesner, Viscusi, Woock, and Ziliak 2005, and Bonhomme and Jolivet 2009).

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

Normally, compensating wage differentials papers estimate the marginal worker's willingness to pay for job characteristics because that is what is captured by observed wages. We instead are interested in the differential between marginal and average worker's willingness to pay. 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.

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

[2] The literature is too large to fully review here. See Kniesner and Leeth (2010) for a recent review.

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 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 (Brown, 1980; Duncan and Holmlund, 1983; Elliott and Sandy, 1998). Even when job characteristics are not self-reported, measurement errors occur because some industry specific

[3] Adam Smith's idea of the amount of trust required corresponds closely to the current idea of efficiency wage in modern labor markets (Shapiro and Stiglitz, 1985). 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: Brown (1980), Duncan and Holmlund (1983), Goddeeris (1988), and Kostiuk (1990). For cost of learning: Rosen (1972), Marder and Hough (1983), Weiss (1986), and Barron, Berger, and Black (1999). For availability: Adams (1985), Li (1986), Hamermesh and Wolfe (1990), Moretti (2000), and Averett, Bodenhorn, and Staisiunas (2005). Finally, for probability of success: King (1974), Rosen (1981), and Hartog and Vijverberg (2007)

[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.

job characteristics may not be relevant for all workers in that industry (Viscusi and Aldy, 2003; Kniesner, Viscusi, Woock, and Ziliak, 2012). 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 (Rosen, 1974; Bartik, 1987a,b; Ekeland, Heckman, and Nesheim, 2002). 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 (Eberts and Stone, 1985; Viscusi and Aldy, 2003; Ashenfelter, 2006). 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 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,

[5] Higher wages lead to both significantly higher probability of working and higher number of tasks performed in both experiments. We examine labor supply elasticity estimates in detail in a separate paper (Pörtner and Hassairi, 2015).

and selection substantially lowers the estimated compensating wage differentials in longitudinal data.

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 (Rosen, 1986). 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

$$U = u(c) + v(l; d), \tag{1}$$

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]

[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.

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

$$c = hw + I \tag{2}$$

$$T = l + h, \tag{3}$$

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:

$$\frac{v_l}{u_c} = w, \tag{4}$$

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

$$(u_c + hwu_{cc})\mathrm{d}w + (w^2 u_{cc} + v_{ll})\mathrm{d}h + wu_{cc}\mathrm{d}I - v_{ld}\mathrm{d}d = 0. \tag{5}$$

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

$$\frac{\mathrm{d}h}{\mathrm{d}I} = -\frac{wu_{cc}}{w^2 u_{cc} + v_{ll}} < 0 \tag{6}$$

$$\frac{\mathrm{d}h}{\mathrm{d}w} = -\frac{u_c + hwu_{cc}}{w^2 u_{cc} + v_{ll}} \gtrless 0. \tag{7}$$

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 the substitution effect 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.

$$\frac{\mathrm{d}h}{\mathrm{d}d} = \frac{v_{ld}}{w^2 u_{cc} + v_{ll}} < 0 \tag{8}$$

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.

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 (Buhrmester, Kwang, and Gosling, 2011). 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

[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 A.1 shows an example of a job listing on Mechanical Turk.

completed each day (Ipeirotis, 2010). 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 (Ipeirotis, 2008). 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 (Horton, 2011).

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 of the four job characteristics and the pay offered. Both experiments use a full factorial design (Fisher, 1935). 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

[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.

as other approaches (see, for example, Wu and Hamada 2011 and Collins, Dziak, Kugler, and Trail 2014). 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 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 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 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.

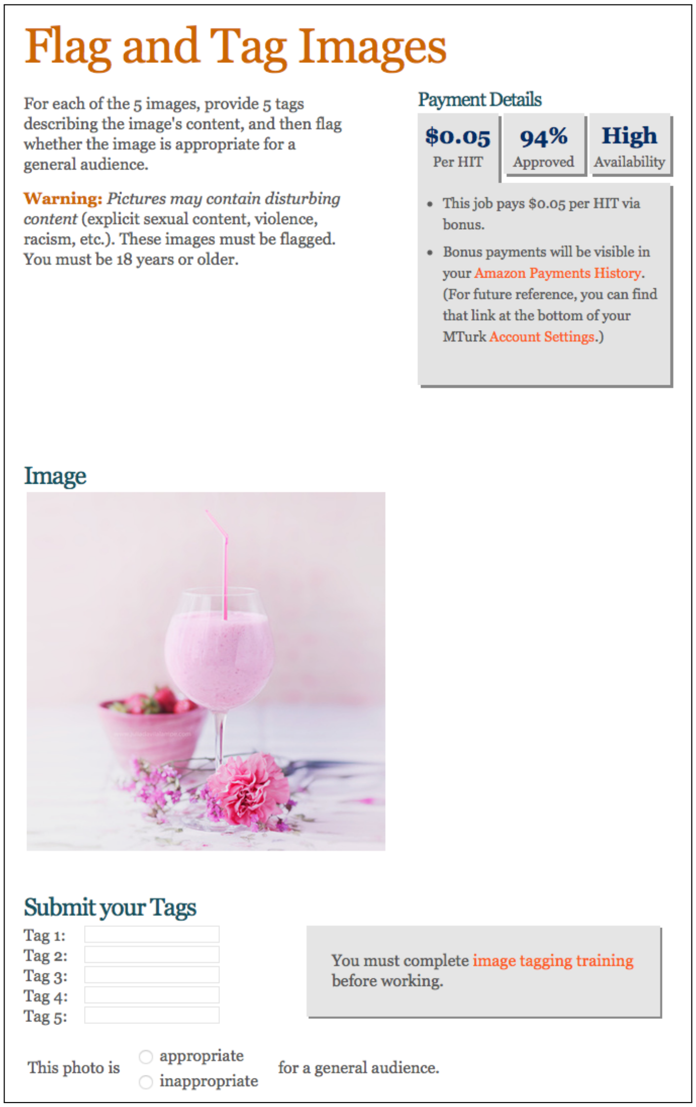
Employers can only contact workers they have paid in the past. We therefore paid all new workers a $0.25 "bonus". The bonus allows us to contact workers for a survey that we ran after the experiments independently of whether they completed any real HITs or not. We do this only the first time a worker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

[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.

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. Figure 1 shows part of the page presented once a worker accepts the HIT.

Figure 1: Image Tagging Experiment Page View



Flag and Tag Images

For each of the 5 images, provide 5 tags describing the image's content, and then flag whether the image is appropriate for a general audience.

Warning: *Pictures may contain disturbing content* (explicit sexual content, violence, racism, etc.). These images must be flagged. You must be 18 years or older.

Payment Details

$0.05	94%	High
Per HIT	Approved	Availability

- This job pays $0.05 per HIT via bonus.
- Bonus payments will be visible in your Amazon Payments History. (For future reference, you can find that link at the bottom of your MTurk Account Settings.)

Image

Submit your Tags

Tag 1:
Tag 2:
Tag 3:
Tag 4:
Tag 5:

You must complete image tagging training before working.

This photo is ○ appropriate ○ inappropriate for a general audience.

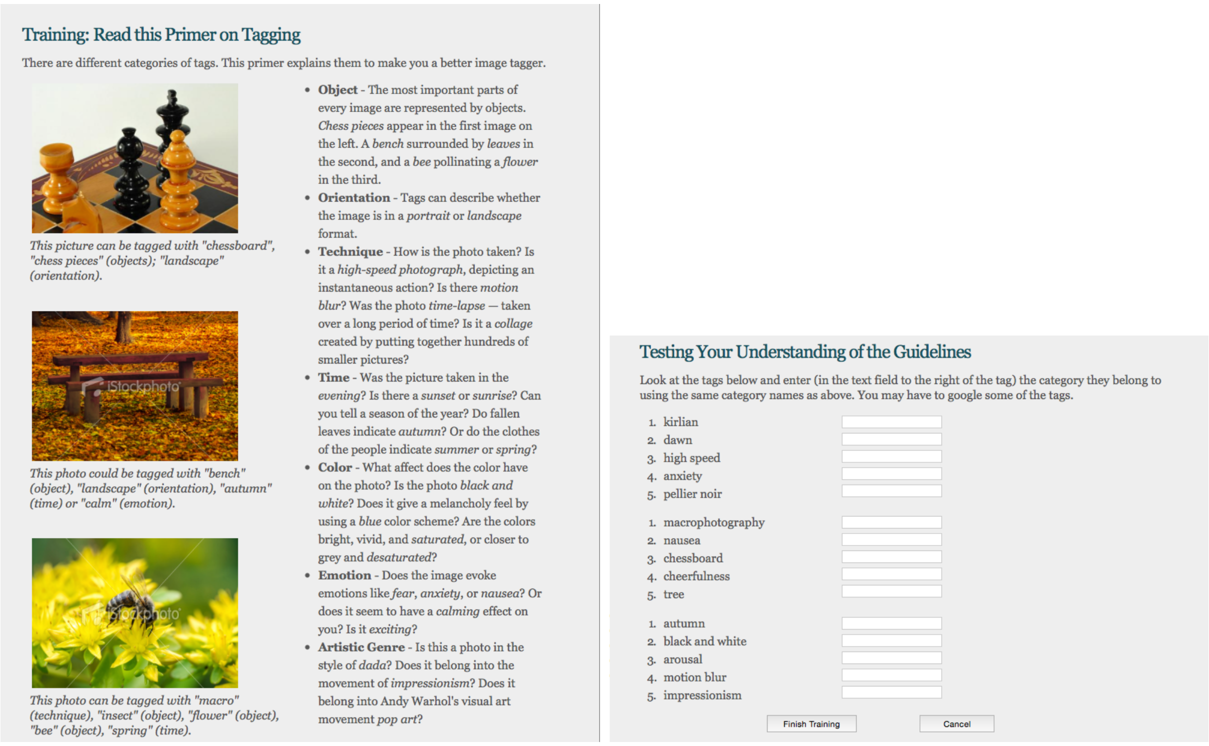
We change the job's agreeableness by varying the number of disagreeable images. There are six levels in the experiment, corresponding to 0, 1, 2, 3, 4, or 5 disagreeable pictures per HIT. In our data disagreeableness is expressed as a ratio between 0 and 1. The number of disagreeable pictures do not 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.[10] 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 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. Figure 2 shows the guidelines and the test questions.

The probability of success is captured by our "approval" rate for tags, a high of 93% and a low of 56%. Figure 3 shows an example. Because experiment run over multiple days, the actual number displayed is drawn from an uniform distribution with a mean equal to either the low or high approval rate. This ensures that returning worker do not see exactly the same number over multiple days. We pay everybody for all work irrespectively of the assigned approval rate. Furthermore, we never reject HITs.

[10] 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.

Figure 2: Image Tagging Experiment—Training and Test



Training: Read this Primer on Tagging

There are different categories of tags. This primer explains them to make you a better image tagger.

This picture can be tagged with "chessboard", "chess pieces" (objects); "landscape" (orientation).

This photo could be tagged with "bench" (object), "landscape" (orientation), "autumn" (time) or "calm" (emotion).

This photo can be tagged with "macro" (technique), "insect" (object), "flower" (object), "bee" (object), "spring" (time).

- **Object** - The most important parts of every image are represented by objects. *Chess pieces* appear in the first image on the left. A *bench* surrounded by *leaves* in the second, and a *bee* pollinating a *flower* in the third.
- **Orientation** - Tags can describe whether the image is in a *portrait* or *landscape* format.
- **Technique** - How is the photo taken? Is it a *high-speed photograph*, depicting an instantaneous action? Is there *motion blur*? Was the photo *time-lapse* — taken over a long period of time? Is it a *collage* created by putting together hundreds of smaller pictures?
- **Time** - Was the picture taken in the *evening*? Is there a *sunset* or *sunrise*? Can you tell a season of the year? Do fallen leaves indicate *autumn*? Or do the clothes of the people indicate *summer* or *spring*?
- **Color** - What affect does the color have on the photo? Is the photo *black and white*? Does it give a melancholy feel by using a *blue* color scheme? Are the colors bright, vivid, and *saturated*, or closer to grey and *desaturated*?
- **Emotion** - Does the image evoke emotions like *fear*, *anxiety*, or *nausea*? Or does it seem to have a *calming* effect on you? Is it *exciting*?
- **Artistic Genre** - Is this a photo in the style of *dada*? Does it belong into the movement of *impressionism*? Does it belong into Andy Warhol's visual art movement *pop art*?

Testing Your Understanding of the Guidelines

Look at the tags below and enter (in the text field to the right of the tag) the category they belong to using the same category names as above. You may have to google some of the tags.

1. kirlian
2. dawn
3. high speed
4. anxiety
5. pellier noir

1. macrophotography
2. nausea
3. chessboard
4. cheerfulness
5. tree

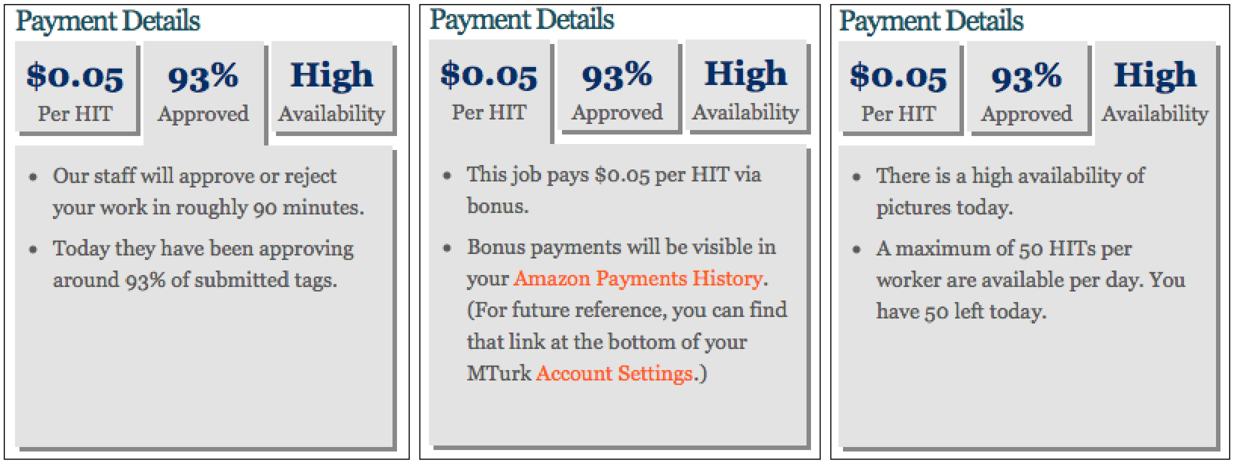
1. autumn
2. black and white
3. arousal
4. motion blur
5. impressionism

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 3 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.

Figure 3: Image Tagging Experiment—Approval rate, pay, and availability



- Our staff will approve or reject your work in roughly 90 minutes.
- Today they have been approving around 93% of submitted tags.

- This job pays $0.05 per HIT via bonus.
- Bonus payments will be visible in your Amazon Payments History. (For future reference, you can find that link at the bottom of your MTurk Account Settings.)

- There is a high availability of pictures today.
- A maximum of 50 HITs per worker are available per day. You have 50 left today.

workers are on Mechanical Turk. 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.

3.2 Letter Writing Job

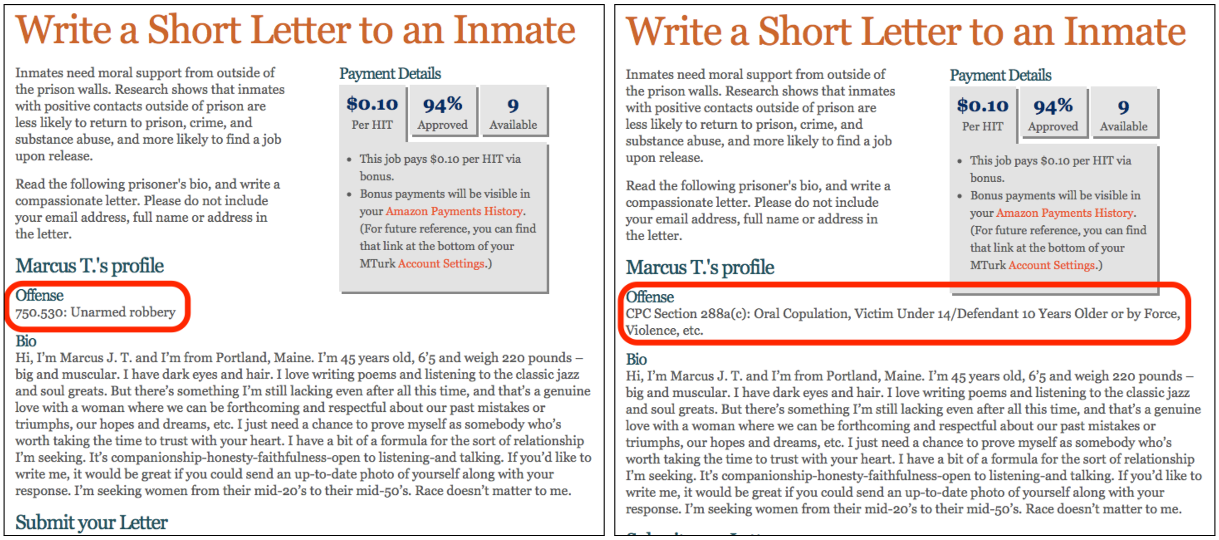
In the letter writing job the task is to write a positive and supportive letter to a prison inmate. All names and profiles of the inmates are fictitious, but based loosely on one or more real inmate profiles from prison pen pal sites. We created 90 profiles and for each arriving worker our program creates a randomized list of the profiles. If a worker left a letter undone the worker see the same prison 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. Figure 4 shows an example of the two levels of disagreeableness.

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. Figure 5 shows the guidelines and one of the test questions.

The probability of success is shown by our "acceptance" rate for letters, although we pay

Figure 4: Letter Writing Experiment—Agreeable vs Disagreeable



Write a Short Letter to an Inmate

Inmates need moral support from outside of the prison walls. Research shows that inmates with positive contacts outside of prison are less likely to return to prison, crime, and substance abuse, and more likely to find a job upon release.

Read the following prisoner's bio, and write a compassionate letter. Please do not include your email address, full name or address in the letter.

Payment Details

$0.10 Per HIT | 94% Approved | 9 Available

- This job pays $0.10 per HIT via bonus.
- Bonus payments will be visible in your Amazon Payments History. (For future reference, you can find that link at the bottom of your MTurk Account Settings.)

Marcus T.'s profile

Offense
750.530: Unarmed robbery

Bio
Hi, I'm Marcus J. T. and I'm from Portland, Maine. I'm 45 years old, 6'5 and weigh 220 pounds – big and muscular. I have dark eyes and hair. I love writing poems and listening to the classic jazz and soul greats. But there's something I'm still lacking even after all this time, and that's a genuine love with a woman where we can be forthcoming and respectful about our past mistakes or triumphs, our hopes and dreams, etc. I just need a chance to prove myself as somebody who's worth taking the time to trust with your heart. I have a bit of a formula for the sort of relationship I'm seeking. It's companionship-honesty-faithfulness-open to listening-and talking. If you'd like to write me, it would be great if you could send an up-to-date photo of yourself along with your response. I'm seeking women from their mid-20's to their mid-50's. Race doesn't matter to me.

Submit your Letter

Write a Short Letter to an Inmate

Inmates need moral support from outside of the prison walls. Research shows that inmates with positive contacts outside of prison are less likely to return to prison, crime, and substance abuse, and more likely to find a job upon release.

Read the following prisoner's bio, and write a compassionate letter. Please do not include your email address, full name or address in the letter.

Payment Details

$0.10 Per HIT | 94% Approved | 9 Available

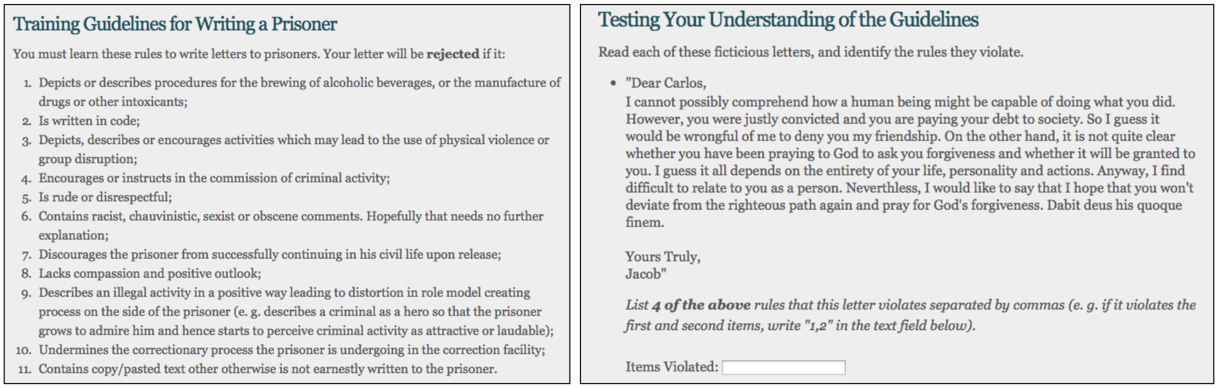
- This job pays $0.10 per HIT via bonus.
- Bonus payments will be visible in your Amazon Payments History. (For future reference, you can find that link at the bottom of your MTurk Account Settings.)

Marcus T.'s profile

Offense
CPC Section 288a(c): Oral Copulation, Victim Under 14/Defendant 10 Years Older or by Force, Violence, etc.

Bio
Hi, I'm Marcus J. T. and I'm from Portland, Maine. I'm 45 years old, 6'5 and weigh 220 pounds – big and muscular. I have dark eyes and hair. I love writing poems and listening to the classic jazz and soul greats. But there's something I'm still lacking even after all this time, and that's a genuine love with a woman where we can be forthcoming and respectful about our past mistakes or triumphs, our hopes and dreams, etc. I just need a chance to prove myself as somebody who's worth taking the time to trust with your heart. I have a bit of a formula for the sort of relationship I'm seeking. It's companionship-honesty-faithfulness-open to listening-and talking. If you'd like to write me, it would be great if you could send an up-to-date photo of yourself along with your response. I'm seeking women from their mid-20's to their mid-50's. Race doesn't matter to me.

Figure 5: Letter Writing Experiment—Cost of Learning



Training Guidelines for Writing a Prisoner

You must learn these rules to write letters to prisoners. Your letter will be **rejected** if it:

1. Depicts or describes procedures for the brewing of alcoholic beverages, or the manufacture of drugs or other intoxicants;
2. Is written in code;
3. Depicts, describes or encourages activities which may lead to the use of physical violence or group disruption;
4. Encourages or instructs in the commission of criminal activity;
5. Is rude or disrespectful;
6. Contains racist, chauvinistic, sexist or obscene comments. Hopefully that needs no further explanation;
7. Discourages the prisoner from successfully continuing in his civil life upon release;
8. Lacks compassion and positive outlook;
9. Describes an illegal activity in a positive way leading to distortion in role model creating process on the side of the prisoner (e. g. describes a criminal as a hero so that the prisoner grows to admire him and hence starts to perceive criminal activity as attractive or laudable);
10. Undermines the correctionary process the prisoner is undergoing in the correction facility;
11. Contains copy/pasted text other otherwise is not earnestly written to the prisoner.

Testing Your Understanding of the Guidelines

Read each of these ficticious letters, and identify the rules they violate.

- "Dear Carlos,
  I cannot possibly comprehend how a human being might be capable of doing what you did. However, you were justly convicted and you are paying your debt to society. So I guess it would be wrongful of me to deny you my friendship. On the other hand, it is not quite clear whether you have been praying to God to ask you forgiveness and whether it will be granted to you. I guess it all depends on the entirety of your life, personality and actions. Anyway, I find difficult to relate to you as a person. Neverthless, I would like to say that I hope that you won't deviate from the righteous path again and pray for God's forgiveness. Dabit deus his quoque finem.

  Yours Truly,
  Jacob"

List ***4 of the above*** *rules that this letter violates separated by commas (e. g. if it violates the first and second items, write "1,2" in the text field below).*

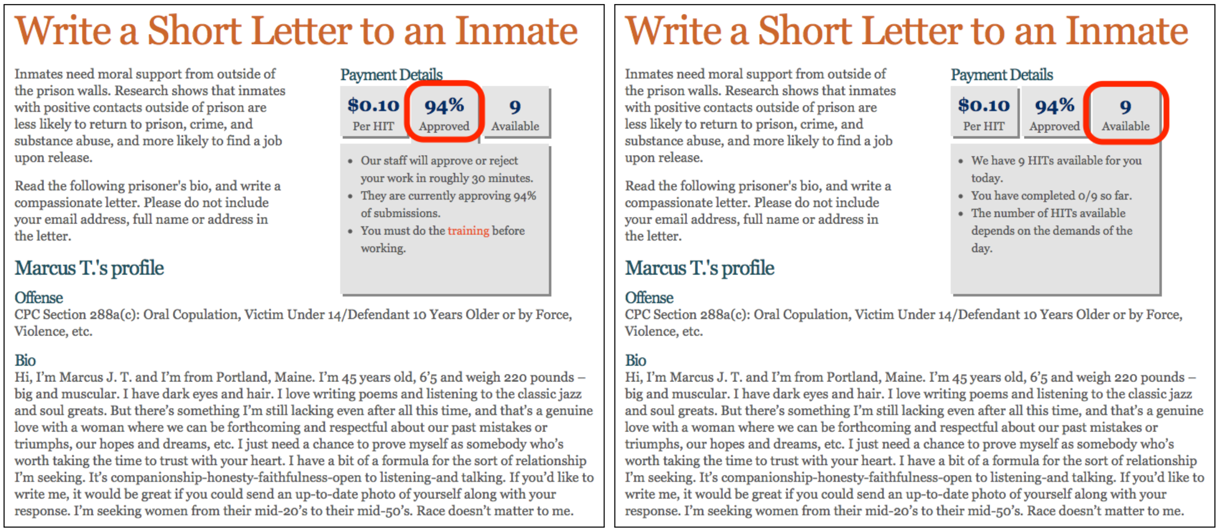
Items Violated: □

everybody who submits acceptable letters. Either 94% or 51% are listed as accepted and the left-hand panel of Figure 6 shows an example of a high approval rate.

Availability is modeled by varying the limit on the number of HITs available to the worker. Either 90 or 9 HITs were available. The right-hand panel of Figure 6 shows a low availability example.

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.

Figure 6: Letter Writing Experiment—Probability of Success and Constancy



Write a Short Letter to an Inmate

Inmates need moral support from outside of the prison walls. Research shows that inmates with positive contacts outside of prison are less likely to return to prison, crime, and substance abuse, and more likely to find a job upon release.

Read the following prisoner's bio, and write a compassionate letter. Please do not include your email address, full name or address in the letter.

Payment Details

$0.10	94%	9
Per HIT	Approved	Available

- Our staff will approve or reject your work in roughly 30 minutes.
- They are currently approving 94% of submissions.
- You must do the training before working.

Marcus T.'s profile

Offense
CPC Section 288a(c): Oral Copulation, Victim Under 14/Defendant 10 Years Older or by Force, Violence, etc.

Bio
Hi, I'm Marcus J. T. and I'm from Portland, Maine. I'm 45 years old, 6'5 and weigh 220 pounds – big and muscular. I have dark eyes and hair. I love writing poems and listening to the classic jazz and soul greats. But there's something I'm still lacking even after all this time, and that's a genuine love with a woman where we can be forthcoming and respectful about our past mistakes or triumphs, our hopes and dreams, etc. I just need a chance to prove myself as somebody who's worth taking the time to trust with your heart. I have a bit of a formula for the sort of relationship I'm seeking. It's companionship-honesty-faithfulness-open to listening-and talking. If you'd like to write me, it would be great if you could send an up-to-date photo of yourself along with your response. I'm seeking women from their mid-20's to their mid-50's. Race doesn't matter to me.

Write a Short Letter to an Inmate

Inmates need moral support from outside of the prison walls. Research shows that inmates with positive contacts outside of prison are less likely to return to prison, crime, and substance abuse, and more likely to find a job upon release.

Read the following prisoner's bio, and write a compassionate letter. Please do not include your email address, full name or address in the letter.

Payment Details

$0.10	94%	9
Per HIT	Approved	Available

- We have 9 HITs available for you today.
- You have completed 0/9 so far.
- The number of HITs available depends on the demands of the day.

Marcus T.'s profile

Offense
CPC Section 288a(c): Oral Copulation, Victim Under 14/Defendant 10 Years Older or by Force, Violence, etc.

Bio
Hi, I'm Marcus J. T. and I'm from Portland, Maine. I'm 45 years old, 6'5 and weigh 220 pounds – big and muscular. I have dark eyes and hair. I love writing poems and listening to the classic jazz and soul greats. But there's something I'm still lacking even after all this time, and that's a genuine love with a woman where we can be forthcoming and respectful about our past mistakes or triumphs, our hopes and dreams, etc. I just need a chance to prove myself as somebody who's worth taking the time to trust with your heart. I have a bit of a formula for the sort of relationship I'm seeking. It's companionship-honesty-faithfulness-open to listening-and talking. If you'd like to write me, it would be great if you could send an up-to-date photo of yourself along with your response. I'm seeking women from their mid-20's to their mid-50's. Race doesn't matter to me.

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.[11]

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:

$$1[H_i > 0] = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i, \tag{9}$$

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

[11] 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.

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:

$$H_i = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i \, \text{if} \, H_i > 0. \tag{10}$$

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 that takes into account both lower bound censoring at zero HITs and the upper bound censoring built into the experiment.[12] 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 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.[13]

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.

[12] 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.

[13] For a more detailed discussion of the three assumptions see Blundell, MaCurdy, and Meghir (2007).

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, Brown, 1980; Duncan and Holmlund, 1983; Villanueva, 2007). 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. 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 intently 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 intently 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:

$$V_i = \alpha + \beta_1 w_i + \mathbf{c}_i \beta_2 + \epsilon_i, \tag{11}$$

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 (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:

$$1[H_{it} > 0] = \alpha + \beta_1 w_{it} + \mathbf{c}_{it} \beta_2 + \mu_i + \epsilon_{it}, \tag{12}$$

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:

$$H_{it} = \alpha + \beta_1 w_{it} + \mathbf{c}_{it} \beta_2 + \mu_i + \epsilon_{it} \text{ if } H_{it} > 0. \tag{13}$$

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.[14]

[14] 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 Dustmann and Rochina-Barrachina (2007) for a comparison of different selection correction models for panel models.

5 Results

During the image tagging experiment's six 24-hour segments, 4,311 workers visited the job.[15] 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. Figure 7 shows the distribution of work done in each experiment. 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% 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% 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. Both show clearly in the histograms. 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 1 shows estimated effects of wage and job characteristics on extensive and intensive margins for the two experiments. For each experiment, the first column is for extensive margin results, the second column intensive margin result using the sample of 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

[15] 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.

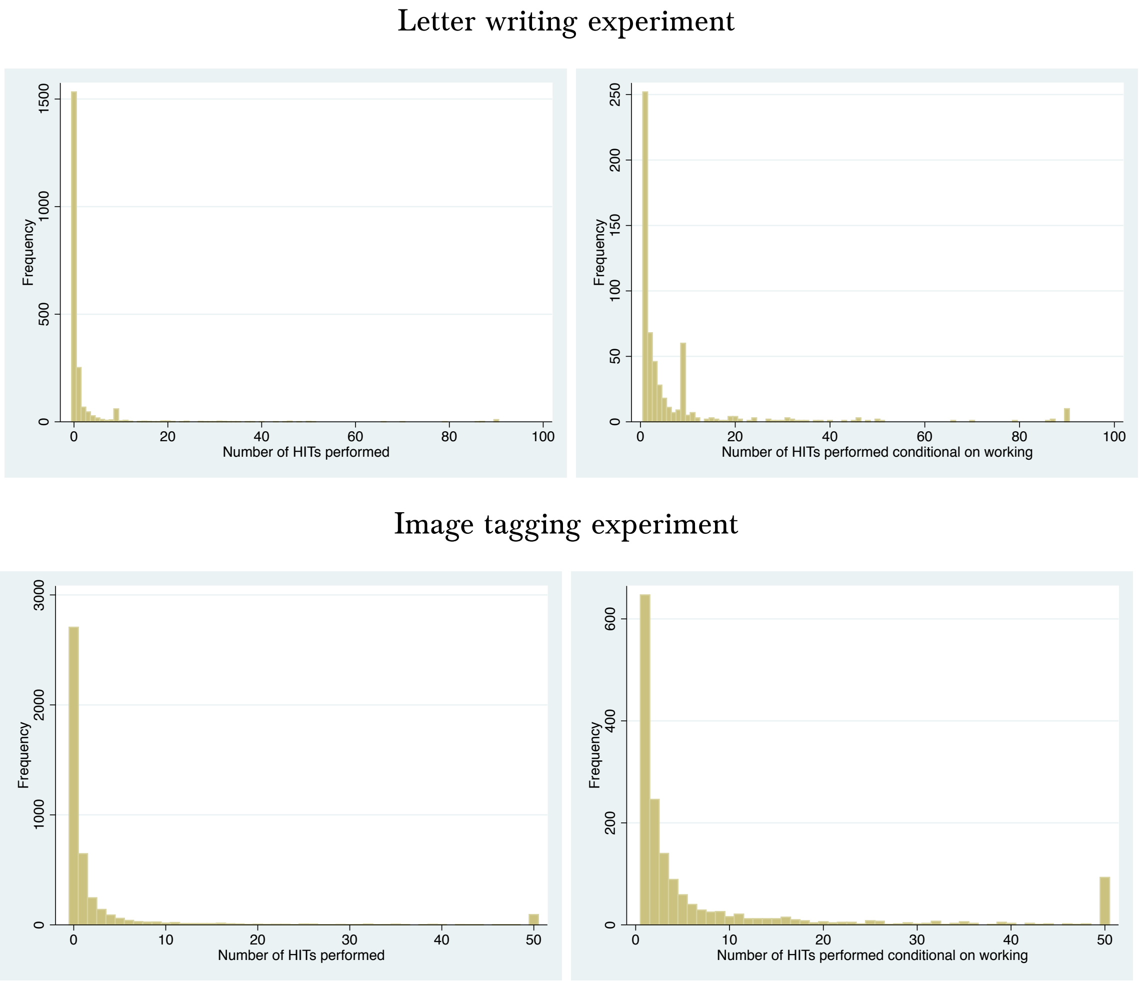

Letter writing experiment

Image tagging experiment

Figure 7: Distribution of work done by experiment

for the "full" sample model there is censoring both at zero and at the maximum number of HITs a worker can perform.[16]

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

[16] Alternative specifications are shown in Appendix Tables B.1 and B.2. 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 B.3 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 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.

percentage points for the letter writing experiment. This is equivalent to 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.[17] 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.[18]

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.[19] 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 cost lead to a reduction of around 6 HITs for both experiments, whereas the lower probability of success lead 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 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 among

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

[18] 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.

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

those who work, may parallel the effect of risk of death on wage in the literature.[20] 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 Ashenfelter (2006, 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.

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 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 shows the results as both absolute changes in wages and percent changes in wages; both evaluated at the mean offered wage.[21] 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 to remain constant.[22] These costs are large relative to the offered wage; the average offered wage is

[20] See, for example, Thaler and Rosen (1976), Biddle and Zarkin (1988), Hamermesh and Wolfe (1990), Viscusi (1993), Viscusi and Aldy (2003), Ashenfelter (2006), and Kniesner, Viscusi, Woock, and Ziliak (2012).

[21] Appendix Table B.4 show the compensating wage differentials for other specifications of the labor supply function. In addition, Appendix Section C shows an alternative approach where we treat observed wages as outcomes and directly estimate the association between job characteristics and wage.

[22] If the probability of working has to remain constant, dividing the point estimate for the job characteristic by

Table 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 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.

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 were required to do). 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 7 show, many workers do relatively few HITs, which increases the cost of taking the test, especially

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.

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.[23]

[23] 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.

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 compensated wages 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.

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.[24]

At the labor market level, workers may, over time, change behavior or decide 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.

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 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 (Farber, 2014). 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.

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

Table 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	Intensive		Extensive	Intensive	
	Worked = 1	HITs Performed		Worked = 1	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 long workers have been on Mechanical Turk, based on prior experiments and the experiments here. Tables 3 and 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 (Toomim, Kriplean, Pörtner, and Landay, 2011). Of the workers in those experiments, only 45 show up among workers who looked at our letter

Table 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 2013 or before	−0.204***	−1.322	−9.166***	−0.185*	−8.887	−5.355
	(0.031)	(1.992)	(1.628)	(0.102)	(6.894)	(5.454)
First observed March/April 2014	−0.226***	2.390	−8.822***	−0.250***	−1.872	−6.107
	(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 × observed 2013				0.038	−1.123	3.209
				(0.047)	(3.415)	(2.707)
Log wage × observed 2014				0.024	−2.146	1.736
				(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 × observed 2013				−0.011	9.058	0.254
				(0.090)	(6.293)	(4.773)
Disagreeableness × observed 2014				−0.007	−2.670	−4.557
				(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 × observed 2013				0.025	0.018	−2.173
				(0.064)	(4.385)	(3.363)
Learning cost × observed 2014				0.013	5.995*	−1.750
				(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 × observed 2013				0.060	4.869	3.098
				(0.065)	(4.178)	(3.371)
Low probability of success × observed 2014				0.128***	0.361	5.790**
				(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.

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.

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 and 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 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.

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 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).

Table 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 enter the experiment on the same day. The dependent variable takes the value 1 if we observe the worker looking at our offered job again and 0 otherwise, independently of whether any work was done on either day, but conditional on there being at least one day left in the experiment.[25]

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 strong 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.

[25] 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).

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.

Table 6: Previous Observed Job Characteristics' Effects on Return Visits Conditional on Working

	Visit job posting given previous day's characteristics and working[a]				
	2nd Visit	3rd Visit	4th Visit	5th Visit	6th Visit
Log wage	0.067***	0.066**	0.059**	0.005	−0.075
	(0.019)	(0.027)	(0.030)	(0.031)	(0.052)
Disagreeableness	−0.038	−0.032	−0.030	−0.115*	0.015
	(0.038)	(0.048)	(0.052)	(0.064)	(0.077)
Learning cost	0.020	0.091***	−0.006	0.015	0.070
	(0.026)	(0.034)	(0.035)	(0.044)	(0.058)
Low probability of success	−0.000	0.023	0.061*	−0.038	0.068
	(0.026)	(0.033)	(0.035)	(0.044)	(0.056)
Low availability	0.033	0.018	0.132*	−0.128	0.029
	(0.063)	(0.081)	(0.073)	(0.087)	(0.113)
Intercept	0.758***	0.849***	0.956***	0.987***	0.796***
	(0.039)	(0.050)	(0.056)	(0.068)	(0.096)
Observations	1,425	605	316	186	54
Mean of dependent variable	0.655	0.798	0.899	0.914	0.963

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 again and 0 otherwise, conditional on there being at least one day left in the experiment and working on the day in question. Example: If we first observe a worker working on our job on Tuesday and that worker returns on Thursday that would count that as 1 for second day and the job characteristics exposed to would be those observed Tuesday. If no work was done on Thursday, the worker leaves the sample unless the worker shows up and work on a subsequent day and that day is not the last day of the experiment.

Table 6 shows the results if we instead condition on working during the previous visit when estimating the return rate. Job characteristics now show small and statistically insignificant effects on the probability of visiting the job again, while the effect of wage becomes substantially larger and more statistically significant. Hence, we are back in the situation where workers self-select and the effects of job disamenities become muted. The selection evident in both situations is important because it illustrates how longitudinal data can be affected by bias in who we observed changing jobs. In regular labor market data, we only observe people who work, and only those who change job contribute to fixed effects estimations.

6.3 Worker Fixed Effects Results

Table 7 shows fixed effects estimates for both extensive and intensive margin.[26] These results provide us 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 analysis 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.

Table 7: 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.

For the extensive margin, three results stand out. First, the effects of wage and the low

[26] Appendix Tables B.5 and B.6 show additional specifications.

availability condition are substantially stronger in the fixed effects data than in the first day 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.[27] 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 we find results that 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 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.

The effects on labor supply are, however, only one part of the picture. Table 8 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 have 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

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

Table 8: Compensating Wage Differentials for Job Disamenities Based on Estimated Fixed Effects Labor Supply Functions for the Image Tagging Experiment

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

Note. All results are based on Table 7 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.

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 hold.

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 (MBO Partners, 2013).[28]

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

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

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 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.

References

Adams, J. D. (1985): "Permanent Differences in Unemployment and Permanent Wage Differentials," *The Quarterly Journal of Economics*, 100(1), 29–56.

Ashenfelter, O. (2006): "Measuring the Value of a Statistical Life: Problems and Prospects," *The Economic Journal*, 116(510), C10–C23.

Averett, S., H. Bodenhorn, and J. Staisiunas (2005): "Unemployment Risk and Compensating Differentials in New Jersey Manufacturing," *Economic Inquiry*, 43(4), 734–749.

Barron, J. M., M. C. Berger, and D. A. Black (1999): "Do Workers Pay for On-The-Job Training?," *The Journal of Human Resources*, 34(2), 235–252.

Bartik, T. J. (1987a): "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.

——— (1987b): "The Estimation of Demand Parameters in Hedonic Price Models," *Journal of Political Economy*, 95(1), 81–88.

Biddle, J. E., and G. A. Zarkin (1988): "Worker Preference and Market Compensation for Job Risk," *The Review of Economics and Statistics*, 70(4), 660–667.

Blundell, R., T. MaCurdy, and C. Meghir (2007): "Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics," vol. 6, Part A of *Handbook of Econometrics*, chap. 69, pp. 4667 – 4775. Elsevier.

Bonhomme, S., and G. Jolivet (2009): "The Pervasive Absence of Compensating Differentials," *Journal of Applied Econometrics*, 24(5), 763–795.

Brown, C. (1980): "Equalizing Differences in the Labor Market," *The Quarterly Journal of Economics*, 94(1), 113–134.

Buhrmester, M., T. Kwang, and S. Gosling (2011): “Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?,” *Perspectives on Psychological Science*, 6, 3–5.

Collins, L. M., J. J. Dziak, K. C. Kugler, and J. B. Trail (2014): “Factorial experiments: efficient tools for evaluation of intervention components,” *American journal of preventive medicine*, 47(4), 498–504.

Duncan, G. J., and B. Holmlund (1983): “Was Adam Smith Right After All? Another Test of the Theory of Compensating Wage Differentials,” *Journal of Labor Economics*, 1(4), 366–379.

Dustmann, C., and M. E. Rochina-Barrachina (2007): “Selection correction in panel data models: An application to the estimation of females’ wage equations,” *The Econometrics Journal*, 10(2), 263–293.

Eberts, R. W., and J. A. Stone (1985): “Wages, Fringe Benefits, and Working Conditions: An Analysis of Compensating Differentials,” *Southern Economic Journal*, 52(1), 274–280.

Ekeland, I., J. J. Heckman, and L. Nesheim (2002): “Identifying Hedonic Models,” *The American Economic Review*, 92(2), 304–309.

Elliott, R. F., and R. Sandy (1998): “Adam Smith may have been right after all: A new approach to the analysis of compensating differentials,” *Economics Letters*, 59(1), 127 – 131.

Farber, H. S. (2014): “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.

Fisher, R. (1935): *The Design of Experiments.* Macmillan.

Goddeeris, J. H. (1988): “Compensating Differentials and Self-Selection: An Application to Lawyers,” *Journal of Political Economy*, 96(2), 411–428.

Hamermesh, D. S., and J. R. Wolfe (1990): “Compensating Wage Differentials and the Duration of Wage Loss,” *Journal of Labor Economics*, 8(1), S175–S197.

Hartog, J., and W. P. Vijverberg (2007): "On compensation for risk aversion and skewness affection in wages," *Labour Economics*, 14(6), 938 – 956, Education and RiskEducation and Risk S.I.

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

Hwang, H.-s., W. R. Reed, and C. Hubbard (1992): "Compensating Wage Differentials and Unobserved Productivity," *Journal of Political Economy*, 100(4), 835–858.

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

——— (2010): "Demographics of Mechanical Turk," *New York University Working Paper.*

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

Kniesner, T. J., and J. D. Leeth (2010): "Hedonic Wage Equilibrium: Theory, Evidence and Policy," Discussion Paper 5076, IZA, Bonn, Germany.

Kniesner, T. J., W. K. Viscusi, C. Woock, and J. P. Ziliak (2005): "How Unobservable Productivity Biases the Value of a Statistical Life," Working Paper 11659, National Bureau of Economic Research.

Kniesner, T. J., W. K. Viscusi, C. Woock, and J. P. Ziliak (2012): "The value of a statistical life: Evidence from panel data," *Review of Economics and Statistics*, 94(1), 74–87.

Kostiuk, P. F. (1990): "Compensating Differentials for Shift Work," *Journal of Political Economy*, 98(5), 1054–1075.

Li, E. H. (1986): "Compensating Differentials for Cyclical and Noncyclical Unemployment: The Interaction between Investors' and Employees' Risk Aversion," *Journal of Labor Economics*, 4(2), 277–300.

Marder, W. D., and D. E. Hough (1983): "Medical Residency as Investment in Human Capital," *The Journal of Human Resources*, 18(1), 49–64.

MBO Partners (2013): "The State of Independence in America - Third Annual Independent Workforce Report," Discussion paper.

Moretti, E. (2000): "Do Wages Compensate for Risk of Unemployment? Parametric and Semiparametric Evidence from Seasonal Jobs," *Journal of Risk and Uncertainty*, 20(1), 45–66.

Pörtner, C. C., and N. Hassairi (2015): "Labor Supply Elasticities in a Low Friction Labor Market," Working paper, Seattle University, Seattle, WA.

Powell, D., and H. Shan (2012): "Income Taxes, Compensating Differentials, and Occupational Choice: How Taxes Distort the Wage-Amenity Decision," *American Economic Journal: Economic Policy*, 4(1), 224–47.

Roback, J. (1982): "Wages, Rents, and the Quality of Life," *Journal of Political Economy*, 90(6), pp. 1257–1278.

Rosen, S. (1972): "Learning and Experience in the Labor Market," *The Journal of Human Resources*, 7(3), 326–342.

——— (1974): "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82(1), 34–55.

——— (1981): "The Economics of Superstars," *The American Economic Review*, 71(5), 845–858.

——— (1986): "The theory of equalizing differences," vol. 1 of *Handbook of Labor Economics*, chap. 12, pp. 641–692. Elsevier.

Shapiro, C., and J. E. Stiglitz (1985): “Equilibrium Unemployment as a Worker Discipline Device: Reply,” *American Economic Review*, 75(4), 892–93.

Smith, A. (1776): *An Inquiry into the Nature and Causes of the Wealth of Nations.* W. Strahan and T. Cadell.

Thaler, R., and S. Rosen (1976): “The Value of Saving a Life: Evidence from the Labor Market,” in *Household Production and Consumption*, ed. by N. E. Terleckyj, pp. 265–302. NBER.

Toomim, M., T. Kriplean, C. C. Pörtner, and J. A. Landay (2011): “Utility of Human-Computer Interactions: Toward a Science of Preference Measurement,” in *Proceedings of the 2011 annual conference on Human factors in computing systems*, pp. 2275–2284. ACM.

Villanueva, E. (2007): “Estimating Compensating Wage Differentials Using Voluntary Job Changes: Evidence from Germany,” *Industrial and Labor Relations Review*, 60(4), 544–561.

Viscusi, W., and J. Aldy (2003): “The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World,” *Journal of Risk and Uncertainty*, 27(1), 5–76.

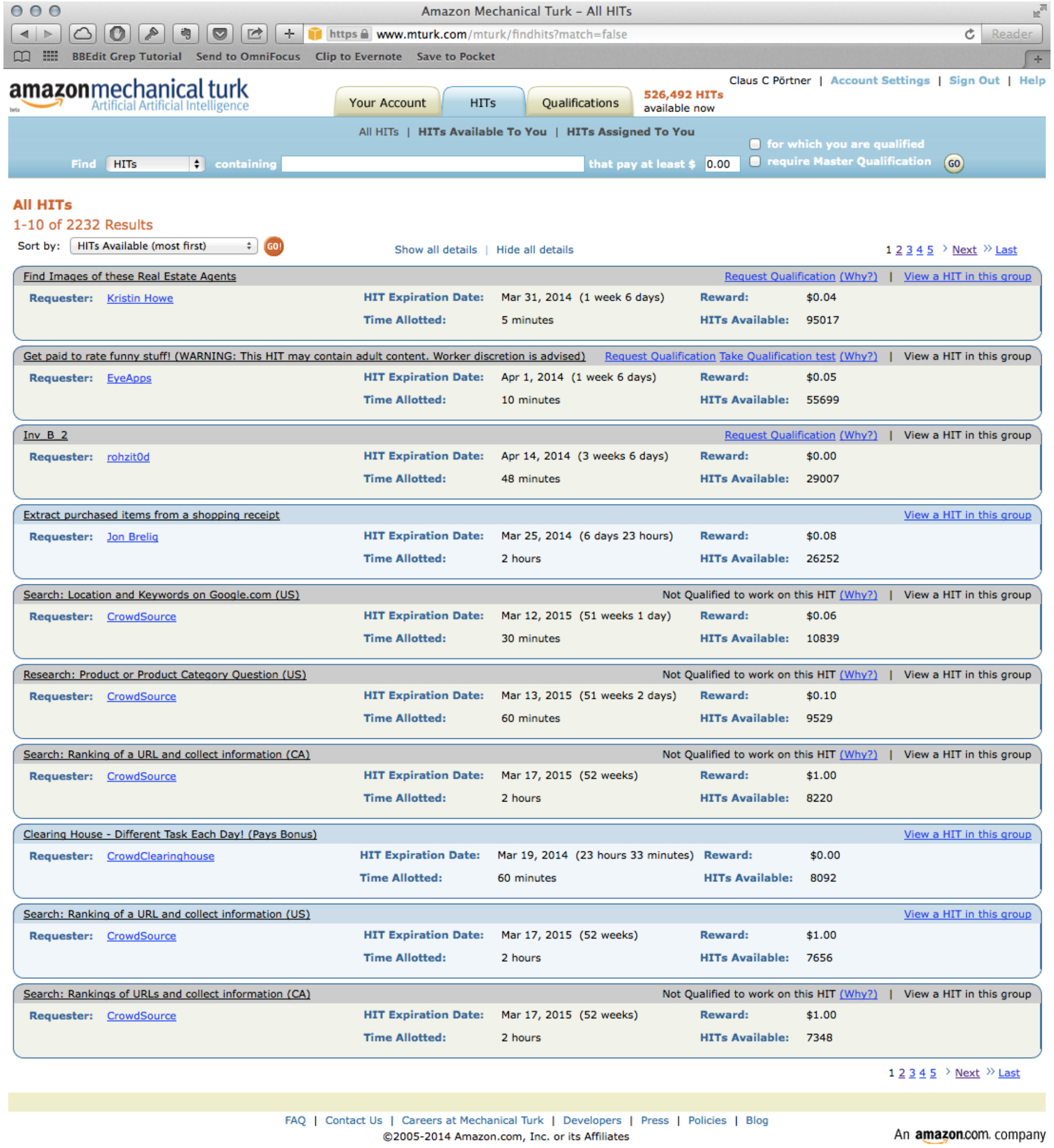
Viscusi, W. K. (1993): “The Value of Risks to Life and Health,” *Journal of Economic Literature*, 31(4), 1912–1946.

Weiss, Y. (1986): “The determination of life cycle earnings: A survey,” vol. 1 of *Handbook of Labor Economics*, chap. 11, pp. 603 – 640. Elsevier.

Wu, C., and M. Hamada (2011): *Experiments: Planning, Analysis, and Optimization*, Wiley Series in Probability and Statistics. Wiley.

A Additional Figures

Figure A.1: Listing of jobs on Mechanical Turk


Amazon Mechanical Turk – All HITs
https www.mturk.com/mturk/findhits?match=false
Reader
BBEdit Grep Tutorial
Send to OmniFocus
Clip to Evernote
Save to Pocket
amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
Claus C Pörtner | Account Settings | Sign Out | Help
All HITs | HITs Available To You | HITs Assigned To You
Find
containing
that pay at least $
GO
GO!
Show all details | Hide all details
1 2 3 4 5 › Next » Last
FAQ | Contact Us | Careers at Mechanical Turk | Developers | Press | Policies | Blog
An amazon.com. company

526,492 HITs available now

Find HITs containing that pay at least $ 0.00

□ for which you are qualified

□ require Master Qualification

All HITs

1-10 of 2232 Results

Sort by: HITs Available (most first)

Find Images of these Real Estate Agents — Request Qualification (Why?) | View a HIT in this group

Requester: Kristin Howe — HIT Expiration Date: Mar 31, 2014 (1 week 6 days) — Reward: $0.04
Time Allotted: 5 minutes — HITs Available: 95017

Get paid to rate funny stuff! (WARNING: This HIT may contain adult content. Worker discretion is advised) — Request Qualification Take Qualification test (Why?) | View a HIT in this group

Requester: EyeApps — HIT Expiration Date: Apr 1, 2014 (1 week 6 days) — Reward: $0.05
Time Allotted: 10 minutes — HITs Available: 55699

Inv_B_2 — Request Qualification (Why?) | View a HIT in this group

Requester: rohzit0d — HIT Expiration Date: Apr 14, 2014 (3 weeks 6 days) — Reward: $0.00
Time Allotted: 48 minutes — HITs Available: 29007

Extract purchased items from a shopping receipt — View a HIT in this group

Requester: Jon Brelig — HIT Expiration Date: Mar 25, 2014 (6 days 23 hours) — Reward: $0.08
Time Allotted: 2 hours — HITs Available: 26252

Search: Location and Keywords on Google.com (US) — Not Qualified to work on this HIT (Why?) | View a HIT in this group

Requester: CrowdSource — HIT Expiration Date: Mar 12, 2015 (51 weeks 1 day) — Reward: $0.06
Time Allotted: 30 minutes — HITs Available: 10839

Research: Product or Product Category Question (US) — Not Qualified to work on this HIT (Why?) | View a HIT in this group

Requester: CrowdSource — HIT Expiration Date: Mar 13, 2015 (51 weeks 2 days) — Reward: $0.10
Time Allotted: 60 minutes — HITs Available: 9529

Search: Ranking of a URL and collect information (CA) — Not Qualified to work on this HIT (Why?) | View a HIT in this group

Requester: CrowdSource — HIT Expiration Date: Mar 17, 2015 (52 weeks) — Reward: $1.00
Time Allotted: 2 hours — HITs Available: 8220

Clearing House - Different Task Each Day! (Pays Bonus) — View a HIT in this group

Requester: CrowdClearinghouse — HIT Expiration Date: Mar 19, 2014 (23 hours 33 minutes) — Reward: $0.00
Time Allotted: 60 minutes — HITs Available: 8092

Search: Ranking of a URL and collect information (US) — View a HIT in this group

Requester: CrowdSource — HIT Expiration Date: Mar 17, 2015 (52 weeks) — Reward: $1.00
Time Allotted: 2 hours — HITs Available: 7656

Search: Rankings of URLs and collect information (CA) — Not Qualified to work on this HIT (Why?) | View a HIT in this group

Requester: CrowdSource — HIT Expiration Date: Mar 17, 2015 (52 weeks) — Reward: $1.00
Time Allotted: 2 hours — HITs Available: 7348

©2005-2014 Amazon.com, Inc. or its Affiliates

B Alternative Specifications

Table B.1: Effects of Job Characteristics on Extensive and Intensive Margins for Image Tagging Experiment

	Extensive		Intensive		
	Worked = 1, not = 0		Number of HITs Performed		
	LPM	Logit	OLS	Censored	
Sample	Full[a]	Full[a]	Worked[b]	Worked[c]	Full[d]
Wage	0.233***	1.044***	11.936***	12.633***	15.117***
	(0.052)	(0.232)	(2.326)	(2.458)	(2.297)
Disagreeableness	−0.115***	−0.512***	−3.914***	−4.020***	−6.179***
	(0.022)	(0.097)	(0.974)	(1.029)	(0.962)
Learning cost	−0.162***	−0.713***	−0.578	−0.633	−6.155***
	(0.015)	(0.066)	(0.665)	(0.703)	(0.660)
Low probability of success	−0.077***	−0.343***	−1.059	−1.101	−3.405***
	(0.015)	(0.066)	(0.656)	(0.693)	(0.654)
Low availability	−0.016	−0.050	−2.525	−2.587	−1.705
	(0.036)	(0.160)	(1.614)	(1.705)	(1.595)
Intercept	0.487***	−0.041	6.740***	6.864***	−4.376***
	(0.022)	(0.098)	(0.951)	(1.004)	(0.977)
Log wage	0.048***	0.220***	2.669***	2.820***	3.277***
	(0.011)	(0.049)	(0.487)	(0.514)	(0.484)
Disagreeableness	−0.115***	−0.514***	−3.952***	−4.058***	−6.215***
	(0.022)	(0.097)	(0.973)	(1.028)	(0.961)
Learning cost	−0.161***	−0.711***	−0.593	−0.649	−6.133***
	(0.015)	(0.066)	(0.664)	(0.702)	(0.660)
Low probability of success	−0.077***	−0.344***	−1.050	−1.090	−3.413***
	(0.015)	(0.066)	(0.655)	(0.693)	(0.653)
Low availability	−0.027	−0.100	−3.070*	−3.162*	−2.421
	(0.036)	(0.160)	(1.610)	(1.701)	(1.593)
Intercept	0.623***	0.573***	13.983***	14.523***	4.651***
	(0.023)	(0.104)	(0.980)	(1.036)	(1.010)
Observations	4,311	4,311	1,605	1,605	4,311
Mean of dependent variable	0.372	0.372	7.6	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.

[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.

Table B.2: Effects of Job Characteristics on Extensive and Intensive Margins for Letter Writing Experiment

	Extensive		Intensive		
	Worked = 1, not = 0		Number of HITs Performed		
	LPM	Logit	OLS	Censored	
Sample	Full[a]	Full[a]	Worked[b]	Worked[c]	Full[d]
Wage	0.171***	0.889***	11.260***	13.390***	14.854***
	(0.033)	(0.175)	(2.292)	(2.524)	(2.186)
Disagreeableness	−0.075***	−0.388***	0.820	0.666	−3.751***
	(0.019)	(0.100)	(1.251)	(1.376)	(1.229)
Learning cost	−0.110***	−0.570***	0.046	0.428	−5.831***
	(0.019)	(0.100)	(1.257)	(1.385)	(1.234)
Low probability of success	−0.054***	−0.279***	1.451	1.637	−2.139*
	(0.019)	(0.100)	(1.246)	(1.373)	(1.224)
Low availability	−0.014	−0.073	−7.749***	−5.421***	−3.416***
	(0.019)	(0.100)	(1.247)	(1.378)	(1.225)
Intercept	0.311***	−0.830***	3.450*	2.177	−15.973***
	(0.029)	(0.148)	(1.907)	(2.093)	(1.942)
Log wage	0.073***	0.400***	4.186***	4.962***	6.175***
	(0.014)	(0.076)	(0.981)	(1.074)	(0.934)
Disagreeableness	−0.074***	−0.385***	0.827	0.673	−3.717***
	(0.019)	(0.100)	(1.257)	(1.384)	(1.232)
Learning cost	−0.110***	−0.567***	0.116	0.502	−5.789***
	(0.019)	(0.100)	(1.263)	(1.393)	(1.237)
Low probability of success	−0.054***	−0.277***	1.440	1.628	−2.127*
	(0.019)	(0.100)	(1.253)	(1.380)	(1.228)
Low availability	−0.014	−0.072	−7.728***	−5.406***	−3.415***
	(0.019)	(0.100)	(1.253)	(1.386)	(1.228)
Intercept	0.462***	−0.030	12.949***	13.466***	−3.012*
	(0.025)	(0.124)	(1.478)	(1.623)	(1.547)
Observations	2,111	2,111	578	578	2,111
Mean of dependent variable	0.274	0.274	7.6	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.

[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 578 workers, 68 were right-censored.

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

Table B.3: Effects of Job Characteristics on Extensive and Intensive Margins for Image Tagging Experiment Excluding All Workers Assigned to Low Availability Condition

	Extensive		Intensive		
	Worked = 1, not = 0		Number of HITs Performed		
	LPM	Logit	OLS	Censored	
Sample	Full[a]	Full[a]	Worked[b]	Worked[c]	Full[d]
Wage	0.233***	1.044***	11.936***	12.637***	15.143***
	(0.052)	(0.232)	(2.332)	(2.466)	(2.304)
Disagreeableness	−0.115***	−0.512***	−3.914***	−4.021***	−6.190***
	(0.022)	(0.097)	(0.976)	(1.032)	(0.965)
Learning cost	−0.162***	−0.713***	−0.578	−0.634	−6.170***
	(0.015)	(0.066)	(0.667)	(0.705)	(0.663)
Low probability of success	−0.077***	−0.343***	−1.059	−1.101	−3.412***
	(0.015)	(0.066)	(0.658)	(0.695)	(0.656)
Intercept	0.487***	−0.041	6.740***	6.865***	−4.404***
	(0.022)	(0.098)	(0.953)	(1.007)	(0.981)
Log wage	0.048***	0.220***	2.669***	2.821***	3.283***
	(0.011)	(0.049)	(0.488)	(0.516)	(0.485)
Disagreeableness	−0.115***	−0.514***	−3.952***	−4.059***	−6.227***
	(0.022)	(0.097)	(0.975)	(1.031)	(0.964)
Learning cost	−0.161***	−0.711***	−0.593	−0.649	−6.149***
	(0.015)	(0.066)	(0.666)	(0.704)	(0.662)
Low probability of success	−0.077***	−0.344***	−1.050	−1.091	−3.420***
	(0.015)	(0.066)	(0.657)	(0.695)	(0.655)
Intercept	0.623***	0.573***	13.983***	14.525***	4.638***
	(0.023)	(0.104)	(0.982)	(1.039)	(1.013)
Observations	4,096	4,096	1,526	1,526	4,096
Mean of dependent variable	0.373	0.373	7.61	7.61	2.84

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.

[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 1,526 workers who worked on the first day they were observed, 88 were right-censored observations.

[d] Of the 4,096 observations, 2,570 were left-censored observations, 1,438 uncensored observations, and 88 right-censored observations.

Table B.4: Compensating Wage Differentials for Job Disamenities Based on Estimated Labor Supply Functions

	Image Tagging[a]				Letter Writing[b]			
	Wage		Log(Wage)		Wage		Log(Wage)	
	$	%	$	%	$	%	$	%
Extensive — LPM								
Disagreeableness	0.49	179	0.66	240	0.44	78	0.56	101
Learning cost	0.70	258	0.92	335	0.64	117	0.83	151
Low probability of success	0.33	120	0.44	160	0.32	57	0.41	74
Low availability	0.07	25	0.15	56	0.08	15	0.11	19
Extensive — Logit								
Disagreeableness	0.49	178	0.64	234	0.44	79	0.53	96
Learning cost	0.68	248	0.88	323	0.65	117	0.78	142
Low probability of success	0.33	119	0.43	156	0.31	57	0.38	69
Low availability	0.05	17	0.13	45	0.08	15	0.10	18
Intensive — "Worked"								
Disagreeableness	0.33	119	0.41	148	−0.07	−13	−0.11	−20
Learning cost	0.05	18	0.06	22	−0.00	−1	−0.02	−3
Low probability of success	0.09	32	0.11	39	−0.13	−23	−0.19	−34
Low availability	0.21	77	0.32	115	0.69	125	1.02	185
Intensive — "Full"								
Disagreeableness	0.41	149	0.52	190	0.25	46	0.33	60
Learning cost	0.41	148	0.51	187	0.39	71	0.52	94
Low probability of success	0.23	82	0.29	104	0.14	26	0.19	34
Low availability	0.11	41	0.20	74	0.23	42	0.30	55

Note. All results are based on Table B.1 for the image tagging experiment and Table B.2 for the letter writing experiment. See those tables for significance levels. For specifications with linear wage the necessary increase in wage to compensate for worse job disamenities is $-\frac{\beta_c}{\beta_w} \times \Delta c$, whereas for specifications with log wage it 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, if required.

[b] Evaluated at the mean offered wage, $0.55, if required.

Table B.5: Effects of Job Characteristics on Extensive and Intensive Margins—Worker Fixed Effects

	Extensive		Intensive	
	Worked = 1, not = 0		Number of HITs Performed	
	Linear	Logit	Linear	linear
Sample	Full[a]	Full[b]	Worked[c]	Full
Wage	0.567***	4.620***	33.828***	23.115***
	(0.043)	(0.379)	(2.506)	(1.252)
Disagreeableness	−0.131***	−1.023***	−6.486***	−3.894***
	(0.018)	(0.151)	(1.012)	(0.527)
Learning cost	−0.095***	−0.750***	−0.196	−0.846**
	(0.012)	(0.102)	(0.688)	(0.357)
Low probability of success	−0.056***	−0.441***	−1.747**	−1.096***
	(0.012)	(0.102)	(0.703)	(0.360)
Low availability	−0.109***	−0.880***	−11.586***	−5.934***
	(0.027)	(0.209)	(1.543)	(0.787)
Log wage	0.113***	0.885***	7.365***	4.562***
	(0.009)	(0.076)	(0.552)	(0.259)
Disagreeableness	−0.130***	−1.002***	−6.487***	−3.867***
	(0.018)	(0.150)	(1.014)	(0.529)
Learning cost	−0.095***	−0.758***	−0.192	−0.854**
	(0.012)	(0.102)	(0.689)	(0.359)
Low probability of success	−0.056***	−0.433***	−1.758**	−1.099***
	(0.012)	(0.102)	(0.704)	(0.362)
Low availability	−0.135***	−1.062***	−13.025***	−7.021***
	(0.027)	(0.209)	(1.538)	(0.789)
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

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 1,605, whereas it is 1,830 for the fixed effects estimations on intensive margin.

Table B.6: Compensating Wage Differentials for Job Disamenities Based on Estimated Fixed Effects Labor Supply Functions

	Image Tagging[a]			
	Wage		Log(Wage)	
	\$	%	\$	%
Extensive — LPM				
Disagreeableness	0.23	84	0.32	115
Learning cost	0.17	60	0.23	84
Low probability of success	0.10	36	0.14	50
Low availability	0.19	70	0.33	119
Extensive — Logit				
Disagreeableness	0.22	81	0.31	113
Learning cost	0.16	59	0.24	86
Low probability of success	0.10	35	0.13	49
Low availability	0.19	69	0.33	120
Intensive — "Worked"				
Disagreeableness	0.19	70	0.24	88
Learning cost	0.01	2	0.01	3
Low probability of success	0.05	19	0.07	24
Low availability	0.34	125	0.49	177
Intensive — "Full"				
Disagreeableness	0.17	61	0.23	85
Learning cost	0.04	13	0.05	19
Low probability of success	0.05	17	0.07	24
Low availability	0.26	94	0.42	154

Note. All results are based on Table 7 for the image tagging experiment. See those tables for significance levels. For specifications with linear wage the necessary increase in wage to compensate for worse job disamenities is $-\frac{\beta_c}{\beta_w} \times \Delta c$, whereas for specifications with log wage it 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, if required.

C Direct Estimation of Compensating Wages

An alternative method to infer compensating wages is regressing job characteristics against observed wages. Although this appears similar to the standard approach to estimating compensating wage differentials, there are important differences. First, we assign wages randomly to arriving workers, so in that sense the wages are exogenous. From the worker's perspective, however, it is a choice whether to accept the wage or not and that makes *observed* wages for those who work endogenous. For all workers who decided not to work we treat their observed wage as missing, just like it would be in a regular labor market data.

Second, in "normal" labor market data we should observe only one wage per job disamenities combination: that of the marginal worker and firm.[29] Our setup is clearly different. The workers were unaware that other wages were available for their specific combination of job disamenities, and we offered the same wage distribution across all combinations of job disamenities. The wages that we observe are therefore not market clearing wages in the standard sense. We can, however, still use the observed wages to calculate how much the average workers value avoiding job disamenities and show how selection affects this estimate.

Using workers who completed at least one HIT on their first visit to an experiment, we first estimate

$$w_i = \alpha + \mathbf{c}_i \beta_1 + \epsilon_i. \tag{14}$$

In the presence of self-selection, using only workers who work should yield estimates of job characteristics' effects on wage that are smaller and less statistically significant than the true effects for the entire population of workers.

The self-selection problem arises because we do not know workers' reservation wages. We know only that workers who work have reservation wages below or equal to their observed wages and that workers who choose not to work have reservation wages that are higher than the

[29] If a higher wage was available somewhere else for the same combination of job disamenities all workers would move to that company instead. Similarly, if workers could be hired at a cheaper wage, companies would offer a lower wage. These pressures lead to only one observed wage per job disamenities combination.

offered wage. To model the self-selection we estimate the effects of job characteristics on wage using a Heckman two-step selection model.[30] We present two versions. The first relies solely on functional form to identify the effects of job characteristics on wage, which is equivalent to prior research that had either no or weak exclusion restrictions. The second relies on the experiments' randomization of wages. The randomization ensures that the wage offered to a worker is orthogonal to worker characteristics and preferences. Hence, the offered wage can serve as an exclusion restriction when estimating the first step in the Heckman model.

Table C.7: Effects of Job Characteristics on Wage for Image Tagging Experiment

	Pay per HIT ($)			Log of Pay per HIT ($)		
	OLS	Heckman		OLS	Heckman	
Sample	Worked[a]	Full[b]		Worked[a]	Full[b]	
Disagreeableness	0.006	−0.031	0.480**	0.042	0.028	2.195**
	(0.010)	(0.122)	(0.201)	(0.050)	(0.582)	(0.916)
Learning cost	−0.000	−0.053	0.680***	0.005	−0.014	3.102***
	(0.007)	(0.171)	(0.234)	(0.034)	(0.821)	(1.064)
Low probability of success	−0.002	−0.027	0.323**	−0.014	−0.023	1.465**
	(0.007)	(0.081)	(0.136)	(0.034)	(0.388)	(0.620)
Low availability	−0.034**	−0.039	0.031	0.051	0.049	0.350
	(0.017)	(0.025)	(0.220)	(0.083)	(0.108)	(1.000)
Intercept	0.284***	0.164	1.833***	−1.442***	−1.486	5.604**
	(0.007)	(0.390)	(0.501)	(0.035)	(1.872)	(2.280)
Identification		Model[c]	Wage[d]		Model[c]	Wage[d]
Inverse Mill's ratio		0.173	−2.239***		0.063	−10.185***
		(0.562)	(0.709)		(2.694)	(3.226)
Observations	1,605	4,311	4,311	1,605	4,311	4,311

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.

[a] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment.

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

[c] Two-step Heckman model where identification comes from non-linearity of first stage, which predicts whether a worker performs at least one HIT.

[d] Two-step Heckman model where identification comes from offered wage, i.e. offered wage is included as additional regression in the first stage that predicts whether a worker performs at least one HIT.

An alternative way to establish the cost of job disamenities is to regress job characteristics on wages. Tables C.7 and C.8 show estimated effects of job characteristics on wage and log wage for the image tagging and the letter writing experiments. We first estimate effects of job characteristics on observed wage per HIT for workers who completed one or more HITs. Con-

[30] For examples of Heckman model estimate of compensating wage differentials, see Kostiuk (1990) and Moretti (2000).

Table C.8: Effects of Job Characteristics on Wage for Letter Writing Experiment

	Pay per HIT ($)			Log of Pay per HIT ($)		
	OLS	Heckman		OLS	Heckman	
Sample	Worked[a]	Full[b]		Worked[a]	Full[b]	
Disagreeableness	0.002	0.180	0.427**	0.003	0.585	0.959**
	(0.023)	(0.969)	(0.195)	(0.054)	(3.157)	(0.438)
Learning cost	−0.009	0.263	0.634***	−0.040	0.846	1.404***
	(0.023)	(1.473)	(0.237)	(0.054)	(4.802)	(0.533)
Low probability of success	−0.019	0.111	0.306*	−0.049	0.375	0.680*
	(0.023)	(0.709)	(0.179)	(0.053)	(2.312)	(0.402)
Low availability	0.021	0.054	0.084	0.051	0.159	0.194
	(0.023)	(0.192)	(0.157)	(0.053)	(0.625)	(0.352)
Intercept	0.599***	1.596	2.938***	−0.657***	2.593	4.596***
	(0.024)	(5.385)	(0.677)	(0.057)	(17.554)	(1.521)
Identification		Model[c]	Wage[d]		Model[c]	Wage[d]
Inverse Mill's ratio		−1.068	−2.540***		−3.481	−5.704***
		(5.769)	(0.706)		(18.806)	(1.586)
Observations	578	2,111	2,111	578	2,111	2,111

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%.
[a] Sample consists of workers who worked, i.e. completed at least one HIT, on the first day the worker was observed during the experiment.
[b] Sample consists of all workers on the first day they are observed during the experiment, whether they worked or not.
[c] Two-step Heckman model where identification comes from non-linearity of first stage, which predicts whether a worker performs at least one HIT.
[d] Two-step Heckman model where identification comes from offered wage, i.e. offered wage is included as additional regression in the first stage that predicts whether a worker performs at least one HIT.

sistent with the prior literature and our labor supply results, the effects of job characteristics on observed pay are small and not statistically significant for both experiment. The one statistically significant effect is low availability for the image tagging experiment and that effect is negative rather than the expected positive.

Second, we show results for Heckman selection models where there is no identifying variables, i.e. identification is solely through the non-linearity of the first stage. For the image tagging experiment the estimated effects of job characteristics on wage generally become more negative, while the estimates for the letter writing experiment all become positive. For neither experiment do any of the job characteristics have a statistically significant effect on observed wage. Furthermore, identifying off the non-linearity in the model works poorly. The inverse Mill's ratio is small and far from being statistically significant for both experiments.

The final results in Tables C.7 and C.8 are for Heckman selection models, where the model is identified through the randomly offered wage. With the exception of low availability, all characteristics now have substantial and statistically significant positive effects on wage. In the image tagging experiment, going from zero to five disagreeable pictures per HIT increases the wage per HIT by almost 50 cents, or a 220% increase if using log of wage as the dependent variable, which at the average offered wage is equal to 90 cents. Similarly, disagreeableness in the letter writing experiment increases the wage by just over 40 cents, or 96% from the log specification, which at the average offered wage is equal to 1.1 dollars. The additional pay required if we asked the workers to take a test before they could work is 68 cent for the image tagging experiment and 63 cent for the letter writing experiment or 310 and 140% higher using log of wage as the outcome. Finally, having a lower probability of success requires 32 cent for the image tagging experiment and 31 cent for the letter writing experiment. In percent the corresponding numbers are 150% and 70%. Lowering availability does not have a statistically significant impact on wages.

Comparing the compensated differentials between the labor supply results and the wage estimations show that the wage estimations are closest to the extensive margin results. All wage estimations results are only a few cents below the extensive margin results and the percentage changes are close as well. Furthermore, disagreeableness is not statistically significant in the standard wage results for image tagging experiment, although it had a statistically significant effect on intensive margin labor supply even without controlling for selection. This suggests that what we capture using the "standard" way of estimating compensating wage differentials tell us more about the extensive margin than what happens at the intensive margin.

Finally, we estimate the effects of job characteristics on observed wages and the effects of job characteristics and wage on effort supplied using fixed effects models. For workers who perform at least one HIT we estimate

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

Table C.9 shows the results and the main conclusion is that, if we use only observed wages over

Table C.9: Effects of Job Characteristics on Observed Wage Worker Fixed Effects Results

	Observed Wage if Working	
	Wage	Log wage
Disagreeableness	0.000	0.001
	(0.010)	(0.047)
Learning cost	0.000	0.001
	(0.007)	(0.032)
Low probability of success	−0.006	−0.026
	(0.007)	(0.033)
Low availability	−0.063***	−0.096
	(0.016)	(0.072)
Observations	3,330	3,330
Number of workers	1,830	1,830

Note. Standard errors in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%. Sample consists of all workers on the days that they have worked, i.e. done one or more HITs, on the image tagging experiment.

time, the fixed effects results show very small effects of job characteristics on wage. The results are in line with what we found using the first day of the experiment, confirming that selection is difficult to overcome, even with fixed effects.

D Selection through Survey Response

After the conclusion of the image experiment, we sent a survey to all workers who came in contact with that experiment, whether they worked or not. The main purpose was to obtain workers' feedback on the job. In addition, the survey asked basic demographic information, employment status, annual household income, and where the worker was when looking at our job. For most questions we allowed a "Prefer not to answer" option. We solicited survey participation through emails sent using Mechanical Turk's system. We paid between \$0.50 and \$2.00 per survey, with \$2.00 the maximum offered payment if a worker did not respond to the initial email. A downside of using Mechanical Turk's system is that some emails probably ended up in workers' spam folders. Despite this we achieved a 47% response rate (2,021 workers responded). The problem with a survey like this is that response is clearly not random. For example, workers who worked were slightly more likely to fill out the survey. Of the workers who never did any HITs 45.7% responded (1,133 workers), while 48.5% of workers who completed one of more HITs over the six days responded (888 workers).

Tables D.10 and D.11 reproduce the main tables using only first day data from survey participants. In each table the top panel shows results with only the experimental conditions as explanatory variables, while the bottom panel includes worker characteristics as additional variables.[31] Two patterns stand out. First, workers who responded to the survey care more about pay and less about job characteristics. The effects of wage on the probability of working are larger than in the main results and the effects of disagreeableness, learning cost, and low probability of success are all smaller for the survey sample than the full sample. Lower statistically significance could be explained by the smaller sample size, but the differences in point estimates must come from differences in the workers. Second, there is very little difference in results whether or not we control for worker characteristics. These results may not be surprising given that we are paying workers to fill out the survey, and those workers who respond to that incentives

[31] In the interest of space we do not present the effects of worker characteristics. The results are available upon request.

Table D.10: Effects of Job Characteristics and Log of Wage on Extensive and Intensive Margins for Image Tagging Experiment using Survey Respondents

	Extensive	Intensive	
	Worked = 1	Number of HITs Performed	
	LPM	Censored	
Sample	Full[a]	Worked[c]	Full[d]
Log wage	0.066***	3.402***	4.506***
	(0.016)	(0.836)	(0.796)
Disagreeableness	−0.092***	−4.137**	−5.788***
	(0.032)	(1.620)	(1.547)
Learning cost	−0.140***	−0.623	−5.871***
	(0.022)	(1.110)	(1.061)
Low probability of success	−0.055**	−0.850	−2.642**
	(0.022)	(1.096)	(1.054)
Low availability	0.044	−2.625	0.357
	(0.056)	(2.656)	(2.610)
Intercept	0.624***	16.357***	5.309***
	(0.035)	(1.670)	(1.648)
Worker characteristics	No	No	No
Log wage	0.065***	3.267***	4.632***
	(0.016)	(0.829)	(0.792)
Disagreeableness	−0.096***	−3.542**	−5.608***
	(0.032)	(1.611)	(1.527)
Learning cost	−0.136***	−0.881	−5.861***
	(0.022)	(1.082)	(1.047)
Low probability of success	−0.062***	−0.544	−2.948***
	(0.022)	(1.087)	(1.044)
Low availability	0.025	−2.468	−0.696
	(0.057)	(2.632)	(2.603)
Intercept	1.000***	30.316**	22.375*
	(0.298)	(12.035)	(12.956)
Worker Characteristics	Yes	Yes	Yes
Observations	2,007	769	2,007

Notes. Standard errors in parentheses; * sign. at 10%; ** sign. at 5%; *** sign. at 1%. Other variables not shown include dummies for sex, age groups, state of residence, income groups, education groups, employment status, and where the worker normally is when working on Mechanical Turk. All samples are conditional on responding to the survey.

[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.

[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 769 workers who worked on the first day they were observed, 54 were right-censored observations.

[d] Of the 2,007 observations, 1,238 were left-censored observations, 715 uncensored observations, and 54 right-censored observations.

are almost by definition likely to care more about money. It does, however, underscore that we should treat survey using Mechanical Turk or other experimental platforms with a substantial amount of skepticism, unless adequate controls for selection are incorporated.

Table D.11: Compensating Wage Differentials for Job Disamenities Based on Estimated Labor Supply Functions for Survey Participants Only on Image Tagging Experiment

	\$	%
Extensive — LPM		
Disagreeableness	0.41	148
Learning cost	0.58	209
Low probability of success	0.26	95
Low availability	−0.11	−38
Intensive — “Worked”		
Disagreeableness	0.30	109
Learning cost	0.08	28
Low probability of success	0.05	20
Low availability	0.21	77
Intensive — “Full”		
Disagreeableness	0.33	121
Learning cost	0.35	127
Low probability of success	0.18	64
Low availability	0.04	15

Note. All results are based on Table D.10 using results controlling for worker characteristics. 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.

Table D.11 show the calculated wage differentials for the four job disamenities. For those job disamenities that have statistically significant effects on labor supply workers in the survey sample require 40 to 50% less increase in pay to accept the worse job disamenities than what we found in the full sample. This reinforces that self-selection, whether from choosing to work or a given job or not or from answering surveys, can substantially affect the estimates for how much compensation is required to entice people to accept worse job disamenities.