AER: Insights 2020, 2(1): 33–46
https://doi.org/10.1257/aeri.20180150

Monopsony in Online Labor Markets[†]

By Arindrajit Dube, Jeff Jacobs, Suresh Naidu, and Siddharth Suri*

Despite the seemingly low switching and search costs of on-demand labor markets like Amazon Mechanical Turk, we find substantial monopsony power, as measured by the elasticity of labor supply facing the requester (employer). We isolate plausibly exogenous variation in rewards using a double machine learning estimator applied to a large dataset of scraped MTurk tasks. We also reanalyze data from five MTurk experiments that randomized payments to obtain corresponding experimental estimates. Both approaches yield uniformly low labor supply elasticities, around 0.1, with little heterogeneity. Our results suggest monopsony might also be present even in putatively "thick" labor markets. (*JEL* C44, J22, J23, J42)

Generations of economics students are taught that the labor market is best described as competitive, with firms facing perfectly horizontal labor supply curves. But a popular alternative view holds that the labor market is characterized by pervasive monopsony, and this view has been bolstered by a recent, fast-growing literature (Naidu, Posner, and Weyl 2018) suggesting that even twenty-first century US labor markets exhibit a substantial degree of market power, possibly due to increased concentration (Benmelech, Bergman, and Kim 2018; Azar, Marinescu, and Steinbaum 2017) or increased use of legal devices such as no-poaching or non-compete agreements (Krueger and Ashenfelter 2018; Starr, Prescott, and Bishara 2017). In this paper, we present direct experimental and quasi-experimental estimates of monopsony in a thick online spot labor market with low putative search frictions. We find considerable market power even here, suggesting that monopsony is *not* limited to thin labor markets, nor markets with high search frictions and/or legal restrictions, and may be far more common than previously thought.

The emergence of online labor platforms represents an idealized environment where frictions are presumably very low. In his review of Manning's 2003 book *Monopsony in Motion*, Peter Kuhn made the following conjecture: "upward-sloping labor supply curves—whether induced by search or other factors—seem unlikely

*Dube: Department of Economics, University of Massachusetts Amherst (email: adube@econs.umass.edu); Jacobs: Department of Political Science, Columbia University (email: jpj2122@columbia.edu); Naidu: SIPA/Department of Economics, Columbia University (email: sn2430@columbia.edu); Suri: Microsoft Research (email: suri@microsoft.com). Amy Finkelstein was the coeditor for this article. This research is derived from a randomized control trial, registered with the AEA (AEARCTR-0001349). We thank Gary Hsieh and Panos Ipeirotis for sharing data as well as Bentley Macleod, Aaron Sojourner, and Glen Weyl for helpful comments.

[†] Go to https://doi.org/10.1257/aeri.20180150 to visit the article page for additional materials and author disclosure statement(s).

to me to be a serious constraint for most firms. This seems even more likely to be the case in the near future, as … information technology has the potential to reduce search frictions" (Kuhn 2004, p. 376). Counter to this conjecture, we find a highly robust and surprisingly high degree of market power even in this large and diverse online spot labor market.

Kingsley, Gray, and Suri (2015) argue employers in online labor markets have significant market power, and show considerable concentration on Amazon Mechanical Turk (MTurk)—a widely-used online labor market—but they stop short of quantifying requester-specific supply elasticities. In this paper, we rigorously estimate the degree of requester market power in MTurk. MTurk is the most popular online micro-task platform, allowing requesters (employers) to post jobs, which workers can complete for pay. In addition to showing that market power can exist even in thick markets for spot labor, understanding monopsony in online labor markets is independently important as they are likely to become much more common.

We provide initial evidence regarding how sensitive the duration of task vacancies are to task rewards, using data from a near-universe of tasks scraped from MTurk. This evidence provides us with an estimate of wage-setting (monopsony) power facing task requesters (Manning 2003, Card et al. 2018). We isolate plausibly exogenous variation in rewards using a double machine learning (Chernozhukov et al. 2018) method, which controls for a highly predictive function of observables generated from the textual and numeric fields associated with each task. This empirical strategy is a labor market analogue to Einav et al. (2015), who match products and sellers using a large sample of listings on eBay to estimate demand elasticities.

We then present results from a number of independent experiments on the sensitivity of workers' acceptance of tasks to the level of pay offered. We analyze data from five previous experiments that randomized wages of MTurk subjects, with the full list of experiments we surveyed given in online Appendix B. While the previous experimenters had randomly varied the wage, none except Dube, Manning, and Naidu (2018) recognized that they had estimated a task-specific labor supply curve, nor noticed that this reflected monopsony power on the MTurk marketplace. We empirically estimate a labor supply elasticity facing requesters on both a "recruitment" margin where workers see a reward and associated task as part of their normal browsing for jobs, and a "retention" margin where workers, having already accepted a task, are given an opportunity to perform additional work for a randomized bonus payment. The experimental recruitment elasticity estimate is obtained from a novel "honeypot" experimental design, where randomly-varied wage offers were made observable only to random subsets of MTurk workers.[1]

Together, these very different pieces of evidence provide a remarkably consistent estimate of the labor supply elasticity facing MTurk requesters, indicating the robustness of our results. The three experiments with a "honeypot" design suggest a *recruitment* elasticity between 0.05 and 0.11. Similarly, *retention* probabilities do not increase very much as a function of reward posted, with implied retention elasticities in the 0.1 to 0.5 range for the two experiments using that design. The precision-weighted average experimental requester's labor supply elasticity is 0.14, and

[1] In search-based models of dynamic monopsony, the labor supply to a firm includes both recruitment and retention margins.

in particular the pooled recruitment elasticity is 0.06, remarkably close to the corresponding 0.096 estimate produced by our preferred double ML specification. The estimates are uniformly small across subsamples, with little heterogeneity by reward amount. This close agreement suggests that the constant elasticity specification, commonly used in the literature, may not be a bad approximation in this context. As a further contribution, our paper provides an independent—and favorable—assessment of the double ML estimator against an experimental benchmark.

I. Monopsony in a Task Market

Monopsony is characterized by two features: wage-setting power and inability to wage-discriminate. MTurk, with its task-posting structure, did not offer many margins for wage-discrimination until very recently (after our sample period). In our sample period, requesters could only restrict the set of eligible workers based on prior acceptance rates (the rates at which previous requesters had deemed their work satisfactory) or location (e.g., by country or by US state).

Monopsony power may arise due to a small number of employers on the platform, from search frictions in locating higher paying tasks, or from idiosyncratic preferences over task characteristics. Prior work has shown that all three of these reasons are at play in MTurk. First, about 10 percent of all requesters post approximately 98–99 percent of all tasks to the AMT platform implying substantial market concentration (Kingsley, Gray, and Suri 2015; Ipeirotis 2010). Second, workers often resort to communicating via off-platform online forums to reduce search costs (Gray et al. 2016). Third, there is evidence for task specialization among workers (Yin et al. 2016).

In online Appendix A, we present a simple model of the MTurk market where employers set wages and wait for tasks to be filled. Each job is seen by a constant fraction λ of workers, who have a distribution of reservation wages (derived from a random utility or rational inattention model) given by $F(w) \propto w^{\eta}$. We show that the labor supply elasticity, η, can be recovered from a regression of log duration on log reward, as well as directly from experimental estimates.

II. Observational Evidence on Recruitment Elasticity from MTurk

A. *Data and Empirical Strategy*

For our observational analysis, we use two primary sources of scraped MTurk data. The first dataset was obtained from Ipeirotis (2010), and covers the January 2014 to February 2016 period. The data consists of over 400,000 scraped HIT batches from the MTurk Tracker web API. This scraper downloaded the newest 200 HIT batches posted to MTurk every six minutes, then the status page for each discovered HIT batch was checked every minute until the page reported that all HITs in the batch had been accepted.

Beginning in May 2016 we launched our own scraper, which took snapshots of all HIT batches on MTurk every 30 minutes, and later increased to every 10 minutes beginning in March 2017. This scraping strategy may miss batches that are posted and filled too quickly for the scraper to detect (i.e., duration less than 30 or

10 minutes). This scraping strategy yielded over 300,000 HIT batches, but stopped working on August 22, 2017, and we have been unable to collect more data since then. We show results separately for these two datasets, and find broadly similar results. Further details on the data are in online Appendix C, including densities of the log duration separately by dataset.

We use the time it takes for a posted batch to disappear as a measure of the probability of acceptance, and regress the duration of the task posting on the observed reward to obtain a "recruitment" estimate of η. As we show in the model in online Appendix A, this is valid under the assumption that the rate at which a job is observed by workers is independent of the wage. We take advantage of the vast amount of available online crowdsourcing data to estimate η, using high-dimensional regression adjustment to control for possibly confounding task characteristics. Duration of a HIT batch is an imperfect proxy for the actual time until a worker takes the job, as batches differ in the number of tasks they offer, and whether workers can do many (e.g., image tagging) or just one (e.g., surveys). Further, batches can be terminated by the requester, for example when they see that it is being filled too slowly. The complementary and quite similar experimental estimates we show below are reassuring that we are in fact measuring the labor supply elasticity with the duration elasticity.

The resulting linear specification is estimated on observations of HIT batches, denoted h, and is given by

$$\ln(duration_h) = -\eta \ln(reward_h) + \nu_h + \epsilon_h, \tag{1}$$

where ν is a nuisance parameter that is correlated with both rewards and durations, and ϵ is an error term that is conditionally independent of durations, so $E[\epsilon \mid \nu] = 0$. An unbiased estimate of η requires that we correctly control for ν, the determinants of duration that are correlated with rewards, and in particular, labor demand. The virtue of the experimental estimates in the third section is that randomization ensures that ν is independent of $\ln(reward)$. With observational data, we must rely on a sufficiently rich set of observables to control for ν, and it is impossible to be completely confident that all possible sources of omitted variable bias have been eliminated. However, the large and high-dimensional nature of the observational MTurk data lets us push the limits of observational analysis. We use two different approaches for the observational analysis, namely fixed effects regression and double machine learning.

B. *Fixed Effects Regression*

In our first strategy, we control for requester and time fixed effects along with fixed effects for deciles of the time allotted by the requester and the number of HITs in the batch. Time allotted is the maximum time the requester allows a worker to finish the task, and can be taken as a very rough proxy for how long the task takes to finish. Controlling for these fixed effects is an attempt to control for task and requester characteristics within a given time period and ideally isolates exogenous variation in labor demand. Formally, we assume that $\nu_h = \rho_r + \tau_t + \delta_d + \delta_N$. This assumption says that the unobserved relative HIT batch attractiveness is captured by the identity of the employer ρ_r, the day the task is first posted τ_t, the decile of the number of minutes allotted for the task δ_d, and the decile of the number of HITs in the batch δ_N. We can then estimate a standard fixed effects regression:

$$\ln(duration_h) = -\eta \ln(reward_h) + \rho_r + \tau_t + \delta_d + \delta_N + \epsilon_h. \tag{2}$$

Results.—In Table 1 we present basic OLS results and fixed effects regressions. Column 1 shows the simple bivariate regression of log duration on log reward. Unsurprisingly this regression is inconclusive, likely because of extensive omitted variables that are correlated with task attractiveness and the intensity of requester demand, both of which would be correlated with both the reward posted as well as the time until the HIT is filled. Column 2 implements the fixed-effects specification, controlling for deciles of time allotted for the task as well as fixed effects for requester and the date posted described above. The coefficient on log reward is −0.06, but it is imprecise and statistically indistinguishable from 0.

C. *Double Machine Learning*

As our second approach, we implement a "double machine learning" (double ML) estimator recently developed by Chernozhukov et al. (2018), which in our case uses an ensemble machine learning approach to model the unobserved ν.

In particular, we suppose that ν in equation (1) is equal to $g_0(Z)$, an unknown function of a high-dimensional vector of observable variables Z. We further suppose that variation in rewards is generated by another function of Z so that $\ln(rewards) = m_0(Z) + \mu$. Combining these two equations we get

$$\ln(duration) = -\eta \ln(reward) + g_0(Z) + \epsilon, \quad E\left[\epsilon \,|\, Z, \ln(reward)\right] = 0, \tag{3}$$

$$\ln(reward) = m_0(Z) + \mu, \quad E[\mu \,|\, Z] = 0. \tag{4}$$

The benefit of the procedure proposed by Chernozhukov et al. (2018) stems from the fact that it allows us to utilize any number of state-of-the-art machine learning methods, such as neural nets or random forests, to obtain estimates of the conditional expectation functions $\hat{l}_0(Z) = E\left[\ln\left(\widehat{duration}\right)|Z\right]$ and $\widehat{m}_0(Z) = E\left[\ln\left(\widehat{rewards}\right)|Z\right]$ which are then "partialled out" to obtain our desired estimator $\check{\eta}$. Note that l_0 is different from g_0 because it is not conditional on $\ln(reward)$. Specifically, from our machine learning-estimated $\hat{l}_0(Z)$ and $\widehat{m}_0(Z)$ we can compute the residuals from (3) and (4) as $\widehat{\mu} = \ln(reward) - \widehat{m}_0(Z)$ and $\hat{\xi} = \ln(duration) - \hat{l}_0(Z)$, respectively, and use these residuals to compute the final estimator as

$$\check{\eta}^0 = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i^{\,2}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i \hat{\xi}_i. \tag{5}$$

The bias from overfitting will not asymptotically go to 0 if the same data is used to estimate $l_0(Z)$ and $m_0(Z)$ and η. However, if a different sample is used to estimate

Table 1—Duration Elasticities from Observational MTurk Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log reward	0.186	−0.0600					
	(0.0947)	(0.0585)					
log reward-ML res.			−0.0958	−0.0787	−0.198	−0.181	−0.0299
			(0.00558)	(0.00651)	(0.0281)	(0.0161)	(0.00402)
Observations	644,873	629,756	644,873	629,756	93,775	292,746	258,352
Clusters	41,167	26,050	41,167	26,050	6,962	18,340	24,923
Type	OLS	FE	ML	ML–FE	ML	ML	ML
Data	Pooled	Pooled	Pooled	Pooled	2017	2016–2017	2014–2016

Notes: This table presents η estimates using data scraped from MTurk. Units are HIT batches. Column 1 presents the unadjusted coefficient from a bivariate regression of log duration on log reward. Column 2 estimates the specification in equation (2). Column 3 presents estimates from an OLS regression of the residualized log duration on the residualized log reward, as in equation (5) averaged across the two sample splits. Column 4 adds the fixed effects in column 2 as further controls to column 3. Columns 5–7 present the double ML estimate from different scraped subsamples. Standard errors are clustered at the requester level.

$l_0(Z)$ and $m_0(Z)$ and $\check{\eta}$ is averaged over multiple folds, then the estimator is consistent and unbiased.

The intuition behind this estimator is similar to the classic partial regression formula. In equation (1) the partial regression formula implies that η could be recovered from a regression of $E[\ln(duration)\,|\,\nu]$ on $E[\ln(reward)\,|\,\nu]$. The double ML estimator uses machine learning to form proxies for ν that predict both conditional expectations very well, implying that the resulting residuals have "partialled out" a very flexible function of all covariates that capture as much of the variation as possible without overfitting.

Double Machine Learning Features.—Double machine learning allows us to leverage a large number of covariates for identifying causal effects, using whichever prediction algorithm has highest goodness-of-fit (see Appendix Table 6 for R^2) in held-out data. We construct a large set of both textual and non-textual covariates as inputs to the double ML procedure. We generate four distinct types of textual features from each HIT group's description, title, and list of keywords: n-grams, topic distributions, Doc2Vec embeddings, and hand-engineered features. The details can be found in online Appendix D. Additionally, we use non-textual features from the HIT including information about the batch size, time allotted for each HIT in the group by the requester, time remaining before expiration of the HIT group, required qualifications (e.g., worker acceptance rate required to be above x percent), the volume of HIT groups posted by the requester across the marketplace, and so on (the full set of features is described in online Appendix D.3).

To satisfy the sample-splitting requirement of the double ML estimator, the full set of HIT groups is split into two equally-sized subsets, A and B. Each subset is further split into training and validation sets, with 80 percent of the observations in A going into A_{train} and 20 percent into A_{val}, and similarly for B_{train} and B_{val}. The machine learning then proceeds in two "stages."

In the first stage, the n-gram features are computed for A_{train} and B_{train}, and two series of learning algorithms are run, the first with A_{train} as training data and A_{val} as test data, the second with B_{train} as training data and B_{val} as test data. For each series and each dataset (Ipeirotis 2010 and our own scraped data) the algorithm which achieves the highest total validation score (here the sum of validation scores for reward prediction and duration prediction) is selected as the "final" algorithm to be used for the remainder of the procedure. In each case we ran, scikit-learn's RandomForestRegressor achieved the highest score, and so is the machine learning method underlying all of the double ML results.[2] The random forest regression constructs a series of decision trees, each of which is built based on a random subset of all available features, and takes the mean prediction over all of these trees to be the estimate. For more on random forest regression, see Breiman (2001, section 11).

To begin the second stage of the procedure, we select the 100 textual features which best predicted the reward values in the first stage, along with the 100 which best predicted the duration values, and set these as the first 200 columns of our second-stage feature matrix. The additional text and numeric features, described in online Appendix D, are then appended to the matrix. The "final" algorithm discovered in stage one is then run twice, the first time with the HIT groups in A used as training data and groups in B used as test data, and the second with the training and test sets reversed. These two predictions are then subtracted from the true values, and these residuals are combined as specified in equation (5) to produce the final estimate $\check{\eta}^0$ (along with its standard error) for each dataset.

Results.—In Table 1 we present the double ML regressions (with and without fixed effects) alongside the basic OLS results and fixed effects regressions. Columns 3 through 7 show the results from the double ML estimator. Column 3 shows the bivariate OLS regression of residualized durations on residualized rewards, and here the coefficient on residualized rewards is a strongly significant -0.096. Figure 1 shows the corresponding binned scatterplot, which shows the binned residuals falling quite close to the linear fit implied by a constant elasticity. Moving from the 25th to the 75th percentile of the rewards distribution (from \$0.05 to \$0.60) would result in a 24 percent decrease in duration, a reduction in the time to completion of roughly 13 hours, over a mean duration of 55 hours.

Column 4 in Table 1 adds the fixed effects from column 2 to the ML specification, and obtains a quite similar estimate of -0.079, suggesting that the double ML procedure is effectively purging the effects of observable variables omitted from column 1 (as a large change in the coefficient would suggest that there were other unobserved variables confounding the regression). Columns 5–7 show the double ML specifications for the different scraped samples. While there is some heterogeneity, the implied elasticities are uniformly small.

III. Experimental Evidence on Labor Supply Elasticity Facing Requesters on MTurk

The observational evidence is quite suggestive of a requester's recruitment elasticity, η, being low, but even in the double ML estimates concerns about omitted variable bias may linger. It is possible that not all task-relevant characteristics have been adequately controlled for, despite the high predictive power of our conditional

[2] RandomForestRegressor consistently achieved the highest score out of {AdaBoostRegressor, BaggingRegressor, ExtraTreesRegressor, GradientBoostingRegressor, RandomForestRegressor, and SVR (SupportVectorRegressor)}.

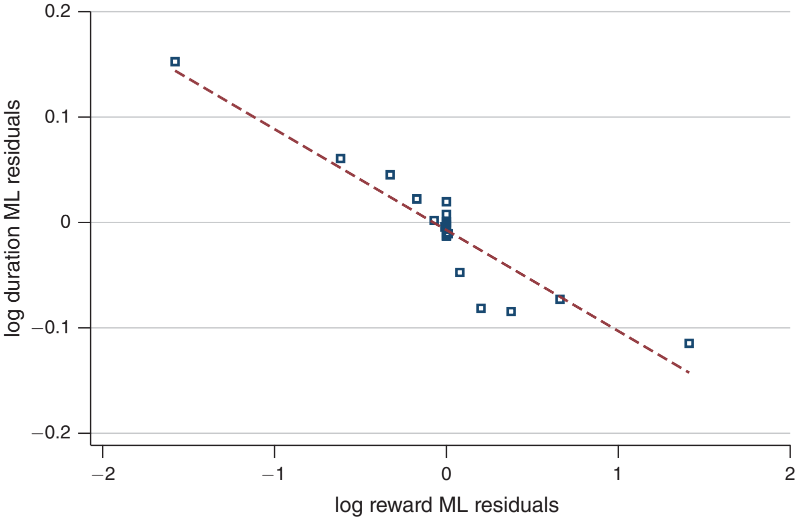

Figure 1

Notes: Binned scatterplot (20 ventiles) for double ML residuals of log duration and log rewards, with $N = 644{,}873$. Residuals are calculated as difference between observed value and predicted value from a random forest trained on a held-out sample, as described in Section IIC.

expectation functions above. If we have experimental (random) variation in rewards, we can estimate the following regression at the worker level i:

$$\Pr(Accept_i) = \alpha + \beta reward_i + \epsilon_i, \tag{6}$$

yielding an estimate of η recovered by $\eta = \beta \times \frac{E[reward]}{\Pr(Accept)}$ with the expectation taken over the population of workers in the sample. We can compare this estimate of η to the double ML estimate from the observational data above to bolster our confidence because if both estimates yield similar results then the double ML estimator is indeed adequately controlling for unobserved variation, and the experimental estimates are externally valid. Next we report experimental estimates of η from the retention margin, and then proceed to estimate η from the recruitment margin—which is most directly comparable to the double ML estimates.

A. *Experimental Retention Elasticities*

Horton, Rand, and Zeckhauser (2011) and Dube, Manning, and Naidu (2018) both run variants of the following experiment. A simple uniformly priced (say, \$0.10) HIT is posted. Subjects give demographic information and perform a simple task (e.g., tagging an image). The subjects are then asked if they would like to perform a given number of additional identical tasks for a randomized bonus wage. The change in the probability of acceptance as a function of the wage gives the responsiveness of requester's labor supply to random wage posting, with low values suggesting a great deal of market power. This is a "retention" estimate of η as workers have already been drawn into a HIT i when asked whether they wish to continue.

Table 2—Offer Acceptance and Offered Rewards from Retention Experiments

	(1)	(2)	(3)	(4)
Panel A. Horton et al. (2011) probability of accepting offer				
Reward	0.127	0.140	0.0861	0.0973
	(0.0219)	(0.0241)	(0.0292)	(0.0333)
Observations	328	307	125	107
η	0.234	0.241	0.192	0.202
SE	0.0334	0.0364	0.0594	0.0664
Average reward	11.60	11.63	11.37	11.50
Sophisticated	No	No	Yes	Yes
Controls	No	Yes	No	Yes
Panel B. Dube et al. (2017) probability of accepting offer				
Reward	0.0267	0.0486	0.0764	0.0782
	(0.0171)	(0.0202)	(0.0348)	(0.0329)
Controls	No	Yes	No	Yes
Observations	5184	5017	1702	1618
η	0.052	0.077	0.118	0.114
SE	0.0333	0.0322	0.0534	0.0479
Average reward	9	9	9	9
Sophisticated	No	No	Yes	Yes

Notes: Coefficients from equation (6) from "retention" experiments, and calculated elasticities, assessed at the specification sample mean. Units are individual workers. Robust standard errors in parentheses.

Experiment 1 was conducted by Horton, Rand, and Zeckhauser (2011), and was among the earliest attempts to estimate economic parameters from MTurk. The authors aimed to elicit the labor-supply elasticity of online workers to the market, but this design does not elicit the market labor supply, but rather the requester's labor supply (i.e., the supply to the experimenter/requester for the particular task). The task in this experiment was transcribing Tagalog translations of paragraphs from Adam Smith's *The Theory of Moral Sentiments*.

Experiment 2 was conducted by Dube, Manning, and Naidu (2018) in 2016, deliberately emulating the design of the Horton, Rand, and Zeckhauser (2011) study with the aim of testing for left-digit bias in the requester's labor supply of online workers. Hence the rewards are substantially lower, between \$0.05 and \$0.15, but the sample sizes are correspondingly larger. The task here was tagging sheets of the 1850 US census slave schedules for the presence of marks in the fugitive slave columns.

We show results for both the full sample and sophisticates (defined as working more than 10 hours on MTurk and primarily for money). The resulting requester's labor supply elasticities are shown in columns 1–4 of Table 2. The implied η from the Horton, Rand, and Zeckhauser (2011) estimates are quite low, between 0.19 and 0.25, while implied η from the Dube, Manning, and Naidu (2018) estimates are even lower, always below 0.12. Besides differences in the tasks, one likely reason for the very slight difference is the different support of the reward variation (Dube, Manning, and Naidu 2018 randomize between \$0.05 and \$0.15, while Horton, Rand, and Zeckhauser 2011 randomize between \$0.10 and \$0.25), and the composition of workers and requesters likely changed considerably between 2011 and 2016. Despite these differences, the estimates are similarly small.

B. *Experimental Recruitment Elasticities*

Engineering an experiment to test the recruitment elasticity is much more challenging than estimating the retention elasticity. We take advantage of three pieces of prior work, Ho et al. (2015); Hsieh and Kocielnik (2016); and Yin, Gray, and Suri (2018), that presented tasks with varying pay rates to random subsets of the MTurk population such that workers assigned one pay rate could not see the tasks available to other workers who had a different pay rate. We stress that none of the papers actually estimated a labor supply elasticity using this random variation in pay.

All of these experiments use a two-phase "honeypot" design. In the first phase a generic HIT is posted at a fixed pay rate. In this simple task, workers are asked a couple of survey questions including whether they would like to be notified of future work opportunities. The IDs of the workers who said yes are then randomized into treatment conditions. During the second phase of the experiment HITs corresponding to the different treatment conditions are launched with identical tasks but varying rewards. This design uses a relatively obscure piece of the MTurk API that lets a requester make a HIT group visible to only a subset of workers. Thus each HIT group can only be seen by and accepted by those treated, and it appears as a regular HIT group in the MTurk interface for them. This design, which first appeared in Ho et al. (2015, section 5) and was later refined in Yin, Gray, and Suri (2018), replicates the search environment workers are facing before having said yes to the task.

In the first experiment (Ho et al. 2015), 800 people were recruited via a $0.05 "honeypot" HIT, and then randomly split into four treatment groups of 200 workers each. The control group (68.5 percent accept rate) earned $0.50 to complete the HIT, one treatment (control for our purposes) had an additional surprise $1 bonus, of whom 64.5 percent accepted, another treatment had an additional performance-based bonus, and a fourth treatment had a base rate of $1.50, of whom 70.5 percent accepted. We drop the group that was given a performance-based bonus incentive and focus on the base payment, ignoring the unexpected bonus payment, to isolate the recruitment elasticity.

In the second experiment (Yin, Gray, and Suri 2018), 1,800 workers recruited using the same "honey pot" protocol were randomly split into three treatment groups, with rewards for the additional task of $0.03, $0.04, and $0.05, respectively. For the task itself, users were asked to categorize an Amazon.com review as positive or negative. Of the 600 in each group, 357 in the $0.03 group accepted, 351 in the $0.04 group accepted, and 371 in the $0.05 group accepted.

In the third experiment (Hsieh and Kocielnik 2016), 927 workers were recruited via a similar design, with the task being to brainstorm the "number of uses of a brick" (a measure of creative thinking) and were given one of seven random rewards: $0.00, $0.05, $0.25, 1 percent chance of $5, 1 percent chance of $25, and $0.25 and $0.50 donation to charity. We drop the lottery and charity treatments and examine only the variation in rewards ($0.00, $0.05, or $0.25), which leaves us with 338 observations. Of these, 131 were in the $0.00 reward group (68 accepted), 89 were in the $0.05 group (52 accepted), and 118 were in the $0.25 group (82 accepted). We made a synthetic dataset based on these numbers in communication with the authors, as the replication data was unavailable.

Table 3—Recruitment Elasticities from Three Experiments

	(1)	(2)	(3)	(4)
Reward	0.00186	0.0451	0.0287	0.00744
	(0.00188)	(0.0587)	(0.0104)	(0.00385)
Observations	600	1,800	338	2,738
η	0.0497	0.0724	0.115	0.0610
SE	0.0503	0.0944	0.0417	0.0290
Average reward	83.33	4	10.04	22.13
Experiment	Spot diff.	Classify reviews	Brainstorming	Pooled

Notes: Coefficients from equation (6) estimated from "recruitment" experiments, and calculated elasticities, assessed at the experimental sample mean. Units are individual workers. The pooled specification includes experiment fixed effects, and is weighted by the inverse of the standard deviation of rewards within each experiment. Robust standard errors in parentheses.

Neither of the first two experiments asked demographic characteristics, and replication data for the third is unavailable, so there is limited capacity to control for observables. However, the randomized assignment of the reward mitigates any role for covariates besides improving precision. Table 3 shows the simple OLS regression results using the same logit specification as equation (6), separately by experiment, and then pooled. The pooled regression controls for experiment fixed effects and weights by the inverse of the standard deviation of rewards within each experiment.

While the first two experiments have insignificant elasticities, in the third experiment we obtain a statistically significant, but still small elasticity, despite a smaller sample size, possibly due to the more attractive nature of the ex post task relative to the other two. Even when all experiments are pooled, the point estimates are remarkably similar despite the very different wage levels at which the experiments were run, and close to the very small estimates obtained from the double ML procedure above. The implied recruitment elasticity from the pooled three experiments is 0.06 and is distinguishable from 0 at 5 percent significance.

C. *Comparison of Estimates*

Figure 2 shows the double ML estimates obtained from pooling the two samples, split by quintiles of the reward distribution, together with the estimates from each of the experiments. The graph also plots the precision-weighted mean elasticity of the experimental estimates (weighted by the inverse of the variance of the estimated elasticities) of 0.14. The double ML estimates are all very close to this line, despite being estimated using very different sources of variation in the rewards. The consistency of the estimates is remarkable, and generally implies a low labor supply elasticity facing requesters on MTurk, with some estimates unable to rule out 0 with 95 percent confidence. Moreover, the labor supply elasticity is largely independent of the reward.

We can use our estimates to infer the distribution of MTurk surplus between workers and requesters, following the formula in online Appendix A that accounts for the dynamics of the requester's problem. The general formula is different from the standard static monopsony problem because a task refused in a given period can be filled in the future, thereby reducing the costs of offering "too low" a wage.

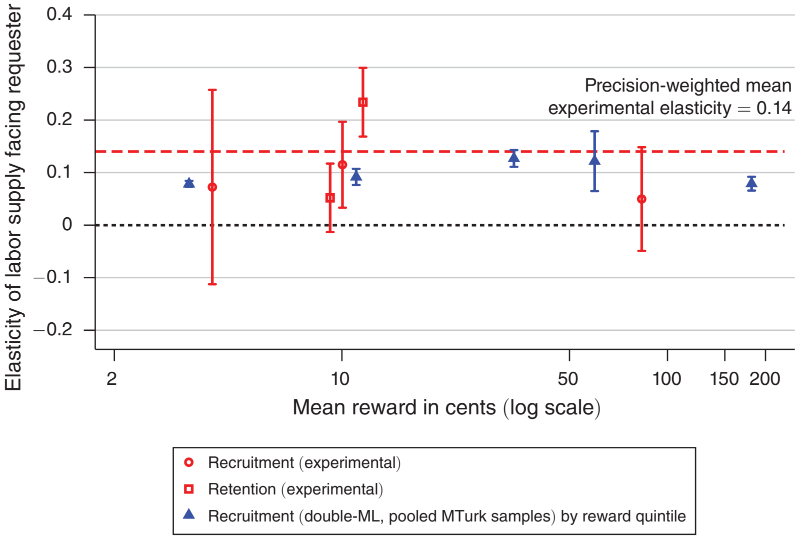

Figure 2

Note: Baseline estimates from both "recruitment" and "retention" experimental designs (column 1 of Table 2 and columns 1–3 of Table 3), as well as double ML recruitment elasticities from observational data estimated by quintile of the reward distribution (*N* for each quintile is between 83,195 and 175,000).

However, when employers are sufficiently impatient (because the task is time-sensitive), the markdown falls to the static Lerner rule. Even these static markdowns are quite large, with workers paid less than 13 percent of their productivity. Despite considerable differences in the institutional environment and type of work, these are close to the markdowns implied by firm labor supply elasticities estimated for nurses by Staiger, Spetz, and Phibbs (2010), among the lowest in the literature.

Are employers using their market power? To check this rigorously, we would need variation in the extent of market power facing requesters, and our observational analysis suggests that elasticities are generally constant. We examine heterogeneity in the double ML elasticities by task type, using the categorization developed by Gadiraju, Kawase, and Dietze (2014). While there are only six categories and the elasticities do not vary very much across categories, online Appendix Figure C.3 shows that tasks with a higher elasticity do have higher reward per minute of time allotted, suggesting that employers are using their monopsony power. The calibrated model explaining round number bunching in Dube, Manning, and Naidu (2018) provides additional evidence on employer optimization on MTurk. Consistent with greater competition in familiar tasks, we also find that more frequent types of tasks have slightly higher elasticities.

IV. Discussion and Conclusion

The findings in this paper provide strong evidence that even in a thick labor market where search frictions may appear to be low, there is considerable monopsony power. As discussed in the introduction, this finding is consistent with the growing body of observational evidence from offline labor markets suggesting monopsony might be at play in those markets as well. Overall, these results call into question the idea that monopsony power is relevant only in unusual cases like company towns or in the presence of legal restrictions on worker mobility.

The source of the monopsony power on MTurk likely lies in the information and market environment presented to workers and requesters, together with the absence of bargaining or many margins of wage discrimination. In particular, the tastes different workers have for a given task may be quite dispersed and not easily discerned by requesters, which induces requesters posting a wage to trade-off the probability of acceptance against a lower wage. Further, this may be exacerbated by the information environment facing workers, which makes searching for alternative jobs difficult. Jobs are highly heterogeneous in time required, entertainment value ("fun") to the worker, and the reliability of the requester in approving payments (Benson, Sojourner, and Umyarov 2017). There is no single dimensional index of job quality that can be used to order HIT groups while searching: workers cannot sort HIT batches by the *real* wage.

As online platforms for data work have increased in prevalence, efforts to mitigate the effects of market power have emerged. For example, workers created their own mechanisms for sharing information about good and bad requesters and HITs via online discussion forum (Gray et al. 2016). Tools like Turkopticon (Irani and Silberman 2013) reduce the information asymmetry by supplying workers with reputation information on requesters. Platforms such as Upwork allow workers to bargain on the wages for a task. Furthermore, some platforms are designed from the ground up to be "worker-friendly" such as Stanford's Dynamo. Also, scientific funders such as Russell Sage have instituted minimum wages for crowdsourced work. The high value data services have as inputs into artificial intelligence has led some to call for "data labor unions" to collectively bargain over high-quality labeled data (Arrieta-Ibarra et al. 2018). Our results suggest that these sentiments and policies may have an economic justification.

REFERENCES

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez Hernández, Jaron Lanier, and E. Glen Weyl. 2018. "Should We Treat Data as Labor? Moving Beyond 'Free.'" *AEA Papers and Proceedings* 108: 38–42.

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2017. "Labor Market Concentration." National Bureau of Economic Research Working Paper 24147.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim. 2018. "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" National Bureau of Economic Research Working Paper 24307.

Benson, Alan, Aaron Sojourner, and Akhmed Umyarov. 2017. "The Value of Employer Reputation in the Absence of Contract Enforcement: A Randomized Experiment." Unpublished.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Card, David, Ana Rute Cardoso, Jörg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36 (S1): S13–S70.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–C68.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in Online Labor Markets: Dataset." *American Economic Review: Insights.* https://doi.org/10.1257/aeri.20180150.

Dube, Arindrajit, Alan Manning, and Suresh Naidu. 2018. "Monopsony and Employer Mis-optimization Explain Why Wages Bunch at Round Numbers." National Bureau of Economic Research Working Paper 24991.

Einav, Liran, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan. 2015. “Assessing Sale Strategies in Online Markets Using Matched Listings.” *American Economic Journal: Microeconomics* 7 (2): 215–47.

Gadiraju, Ujwal, Ricardo Kawase, and Stefan Dietze. 2014. “A Taxonomy of Microtasks on the Web.” In *HT ’14 Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 218–23. New York: ACM.

Gray, Mary L., Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. “The Crowd Is a Collaborative Network.” In *CSCW ’16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 134–47. New York: ACM.

Ho, Chien-Ju, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. “Incentivizing High Quality Crowdwork.” In *WWW ’15 Proceedings of the 24th International Conference on World Wide Web*, 419–29. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. “The Online Laboratory: Conducting Experiments in a Real Labor Market.” *Experimental Economics* 14 (3): 399–425.

Horton, John Joseph, and Lydia B. Chilton. 2010. “The Labor Economics of Paid Crowdsourcing.” In *EC ’10 Proceedings of the 11th ACM Conference on Electronic Commerce*, 209–18. New York: ACM.

Hsieh, Gary, and Rafał Kocielnik. 2016. “You Get Who You Pay For: The Impact of Incentives on Participation Bias.” In *CSCW ’16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 823–35. New York: ACM.

Ipeirotis, Panagiotis G. 2010. “Analyzing the Amazon Mechanical Turk Marketplace.” *XRDS: Crossroads* 17 (2): 16–21.

Irani, Lilly C., and M. Six Silberman. 2013. “Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk.” In *CHI ’13 Proceedings of the SIGCHI Conference on Human Factors in Computing System*s, 611–20. New York: ACM.

Kingsley, Sara Constance, Mary L. Gray, and Siddharth Suri. 2015. “Accounting for Market Frictions and Power Asymmetries in Online Labor Markets.” *Policy & Internet* 7 (4): 383–400.

Krueger, Alan B., and Orley Ashenfelter. 2018. “Theory and Evidence on Employer Collusion in the Franchise Sector.” National Bureau of Economic Research Working Paper 24831.

Kuhn, Peter. 2004. “Is Monopsony the Right Way to Model Labor Markets? A Review of Alan Manning’s *Monopsony in Motion*.” *International Journal of the Economics of Business* 11 (3): 369–78.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets.* Princeton: Princeton University Press.

Naidu, Suresh, Eric A. Posner, and E. Glen Weyl. 2018. “Antitrust Remedies for Labor Market Power.” *Harvard Law Review* 132: 536–601.

Staiger, Douglas O., Joanne Spetz, and Ciaran S. Phibbs. 2010. “Is There Monopsony in the Labor Market? Evidence from a Natural Experiment.” *Journal of Labor Economics* 28 (2): 211–36.

Starr, Evan, J. J. Prescott, and Norman Bishara. 2017. “Noncompetes in the US Labor Force.” Unpublished.

Yin, Ming, Mary L. Gray, and Siddharth Suri. 2018. “Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork.” In *CHI ‘18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York: ACM.

Yin, Ming, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. “The Communication Network within the Crowd.” In *WWW ’16 Proceedings of the 25th International Conference on the World Wide Web*, 1293–1303. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.