Worker Sorting, Work Discipline and Development [*]

David Atkin[†] Antoinette Schoar[‡] Sumit Shinde[§]

October 18, 2022

Abstract

A growing literature explores the impact of home-based versus office-based work. Differences in productivity may arise due to a treatment effect of the office or from workers with different ability sorting into office or home work. If more productive workers find working in the office less costly (a selection effect) or more complementary to their skills (a selection on treatment effect), we expect a positive selection into office work. But if more productive workers have stronger preference for home work or face bigger constraints to sorting into the office, the selection effect could be negative. We conduct a RCT in the data entry sector in India that exogenously allocates workers to the home or office. We find that the productivity of workers randomly assigned to working from home is 18% lower than those in the office. Two thirds of the effect manifests itself from the first day of work with the remainder due to quicker learning by office workers over time. We find negative selection effects for office based work: workers who prefer home-based work are 12% faster and more accurate at baseline. We also find negative selection on treatment: workers who prefer home work are much less productive at home than at the office (27% compared to 13% for worker who prefer the office). These negative selection effects are partially explained by subgroups that might face bigger constraints in selecting into office work, such as people with children or other home care responsibilities as well as poorer households.

Keyword: Worker Productivity, Work-From-Home, Productivity Differences Across Firms JEL Codes: C93, J24, L23, L84, O17, O43

[*]We would like to thank the field research team at LEAD at KREA University, especially the office managers Kothandam D and Christopher Glan. We would like to thank Dahyeon Jeong, Pratibha Joshi, Lipika Kapoor and Mridulya Narasimhan for their excellent research assistance. This project was made possible due Financial support of the ESRC and IGC. We would like to thank all the participants at IPA SME workshop, MIT, UCLA and University of Bonn for their valuable feedback. The experiment received IRB approval from MIT, UCLA and IFMR. All errors are our own.

[†]Department of Economics, MIT, atkin@mit.edu

[‡]MIT Sloan School of Management, aschoar@mit.edu

[§]Department of Economics, UCLA, sumitshinde@ucla.edu

1 Introduction

Several studies have documented that productivity is substantially lower in household enterprises than in larger, typically formal, firms, see, for example, Hsieh and Klenow (2014), La Porta and Shleifer (2014) or McCaig and Pavcnik (2018). Casual observation also seems to suggest a relationship between the rapid expansions in factory and office employment and the high growth rates in the East Asian "miracle" economies. In the popular debate, these findings are often seen as causal evidence that reallocating workers from household enterprises into office- or factory-based employment may play a central role in improving productivity and economic development of a country. However, the observed higher productivity in an office environment might also be driven by selection effects, if more productive workers prefer working in the office or formal firms are able to attract the most talented workers. The debate about the cost and benefits of working from home has recently received heightened interest, even in developed countries, as a result of the Covid-19 pandemic, which forced many employers to shift to working from home (for example see, Barrero et al. (2021), Harrington and Emanuel (2022), and Brynjolfsson et al. (2020)).

In this paper we aim to understand the sources of the productivity differences between home-based and office-based production. Productivity may differ across these two types of work environment for a number of distinct reasons. Most obviously, if production is more efficient or if learning is faster when organized in an office setting, the observed productivity difference between office and home-based work may be driven purely by a treatment effect of office-based production methods (as explored in Bloom et al. (2014), Kaur et al. (2015), and Schoar (2009)).

Alternatively, the productivity difference may be due to sorting of workers with different abilities or preferences. If workers with higher ability or career ambitions have lower cost of working in an office environment or expect greater career benefits, they might select to working in the office (a positive selection effect). In contrast, workers with lower ability might prefer working from home, if it provides a less demanding work environment. In this case, the office would serve as a sorting device for productive workers and could thus be a confounding factor when trying to estimate the benefits of working from the office versus home in non-experimental settings. Furthermore, high-ability workers might have skills that are complements to office work, which increases the treatment effects and might make them more likely to chose office work (a positive selection on treatment effect). On the other hand, however, selection into the office could be constrained by factors that are orthogonal to productivity, or even negatively correlated with it. For example, some highly productive workers might have commitments at home such as child care or elderly care, or use home-based work as a way to bridge unemployment spells while looking for another job. In these cases the selection of workers might be negatively related to productivity, depending on the magnitude of the constraints.[1]

[1]Of course, these workers might not have participated in the labor force at all without the flexibility of home-based work,

While the organizations literature has explored the impacts of the productivity-enhancing practices used in offices and factories, we are not aware of a literature examining this second explanation for such productivity differences; that offices and factories may act as a sorting device.

We set up a randomized control trial in the data entry sector in the city of Chennai, in India. The Indian data-entry sector is an ideal testing environment. First, it is a sector where working from home is quite feasible since workers do not depend directly on the work of others in the organization. Second, it is possible to record productivity and effort in great detail via data entry software on workers' laptops. We were able to establish our own data entry operation so that we can both control work conditions and allocation to home and office work. Third, data entry, and business process outsourcing more generally, is an important and growing sector in India, a country where a large share of production is home or farm based.

Our methodology allows us to separate the treatment effect of the office environment from the selection of high-effort and high-ability employees into these stricter work environments. At the same time, we can also test whether social and cultural constraints affect the ability of workers to sort into different jobs. Our research design is relatively straightforward. Potential data entry workers are recruited through ads in leading newspapers. Qualifying applicants are invited to an entry interview where they complete an initial application as well as some brief data entry tasks to ascertain ability at baseline (measured through data entry tests that record speed and error rates). Applicants are asked at this stage for their (incentivized) preference between office and home work with similar conditions and identical equipment. All applicants are then randomized into either the office or home work treatment for an 8 week data entry job. Minute-by-minute productivity as well as idle times are recorded through the data entry software.[2]

Treatment effects of home work are measured by comparing people randomized into home work to those randomized into the office-based group, independent of their preferences for either work environment. The importance of selection based on initial ability is measured using incentivized choices between the two types of work, which people provided to us in the entry interview. Beyond quantifying the treatment and selection effects of the office versus home, our research design also allows us to answer an additional question. Is there complementarity between high-ability high-effort workers and office-based work? In other words, are the methods used to induce high productivity in the office only effective when the workers are high-ability or high-effort types? If this is the case, the sorting induced by office-based work may be essential in making these workers more productive. To address this question we also explore treatment heterogeneity and selection on treatment, i.e., do those with higher returns to office work disproportionately select into it.

First, we find that the productivity of workers randomly assigned to work from home is 18% lower

an important factor in calculating impacts on aggregate output.

[2] These software features are relatively standard for data entry work and so workers with data entry experience would have seen similar software and metrics at other firms.

than the workers assigned to work from the office. Two thirds of this difference manifests itself immediately, starting from the first day of work. The remainder is a result of slower learning for the home group over the subsequent eight weeks.

These results hold also controlling for baseline speed and characteristics, and hold for the other output measures we collect, for example log speed of typing, accuracy of data entry, or a measure of data entry speed that is adjusted for task difficulty. The treatment effect of working in the office is especially large when workers are assigned to harder tasks. We find some but relatively limited heterogeneity in treatment effects across worker types. Treatment effects are indistinguishable from zero for poorer workers, workers preferring part-time work and women with family care obligations while older female workers, richer workers and married workers exhibit the strongest treatment effects.

Second, contrary to our priors, we find a negative selection effect for office-based work. The workers who state that they prefer a home-based job under incentivized elicitation are 12% faster, not slower, when their data entry ability is measured at baseline as part of the interview process. They also show higher accuracy of data entry and less idle time. What then lies behind the unexpected negative selection effects? Interestingly, our results are not due to selection on treatment: One might have thought that low-ability workers have more to gain from working in the office, if for example, these workers know they have self control issues or need more guidance in the office . Or if high ability workers believe that they will do equally well in both work environments and thus might as well enjoy the convenience of working from home. Our selection on treatment estimates reject this explanation. Specifically, we find evidence of negative selection on treatment. Workers who prefer home-based work are 27% less productive when allocated to working from home compared to working from the office, while this gap is only 13% for workers who prefer office-based work. In other words, the workers who state a preference for working from home have a particularly large negative treatment effect when working at home. Thus, the negative selection of low productivity workers into the office is not because this group sees the largest benefits. Instead, these results suggest that at least some groups of workers are constrained from choosing the work location where they would be most productive.

Finally, we analyze the nature of the constraints and the characteristics associated with more productive workers sorting into home work. For example, norms may prevent educated women or those with home care responsibilities from working outside the house, or working in an office may be a status good for low ability workers even if it does not make them more productive. We test a number of different dimensions of heterogeneity that might explain the negative self-selection. We find some limited support for these explanations, with the strongest effects for low status individuals and those with home pressures, responsibilities and distractions. However, even after including rich controls for baseline characteristics of workers, we still see a substantial negative selection effect. Additionally, we conduct an analysis of heterogeneity in the selection on treatment effects and find selection on treatment is particularly negative among five groups within which heterogeneity in constraints may be particularly acute: workers with family care responsibilities—especially women with such responsibilities—workers with low family income, workers with children, and older workers.

One important caveat is that the size of selection effects maybe somewhat underestimated since, in order to implement this experiment, we had to restrict the sample of workers to those who would in principle be willing to work in either location. Thus applicants with the most extreme preferences were filtered out, since these workers would have attrited from the sample if they were not presented with the location of their choice. That said, our sample still included many with strong preferences: even with this filtering, we observed substantial differential attrition from groups that did not receive the work location of their choice prior to us introducing a big retention bonus. Thus, we were not selecting applicants who were indifferent to work location preferences but also those who had strong preferences but could be incentivized to work in a location of not their choice.

Overall, our results suggest that although there are substantial productivity benefits to working in an office, many workers chose to work from home—particularly those that are high ability and those who would gain the most in terms of productivity from being in the office. Of course, to know whether such choices are optimal from the worker's perspective we need to know more about the nature of the constraints they are making decisions under. For example, these patterns were particularly pronounced among those with care responsibilities at home and those with children. Such findings may be rationalized by heterogeneity in preferences to be able to do chores at home or could be the result of societal or family pressures to stay inside the home. Whatever their source, our results show that constraints on the optimal sorting into office and home based work result in a significant loss in productivity of the work force. These results also suggest that policies which relax the heterogeneity of these constraints, such as providing universal child care, may have important effects on the productivity of the labor force.

2 Related Literature

This paper contributes to several parts of the literature. First, we are motivated by the burgeoning literature that highlights large productivity differences between firms, particularly small household enterprises, and larger formal firms, a pattern that seems particularly prevalent in developing countries. For example, Hsieh and Klenow (2009) and Bartelsman et al. (2013) document large dispersion in total factor productivity across firms within the same industry. Hsieh and Klenow (2014) further highlight the particular prominence of small, informal, low-productivity firms in India and Mexico. A related literature has explored the costs of informality (for example, de Soto (1990), La Porta and Shleifer (2014), Levy (2010), Bruhn (2011), and de Mel et al. (2013)). Most relevant, McCaig and Pavcnik (2018) show substantial labor productivity differences between household enterprises and non-household firms in Vietnam. We aim to shed more light on the origins of these differences and constraints on the optimal sorting of workers into different work environments. In this regard, the paper is related to Hsieh et al. (2019) that investigates the degree of misallocation generated by historic work restrictions on women and blacks in the US.

A second relevant literature is the work in organizational economics that documents the importance of management practices even within large formal workplaces (e.g., Bloom et al. (2013)). For example, Kaur et al. (2015) carry out a range of experimental innovations at a data entry firm in India. They show that some workers have self-control issues and are willing to choose dominated contracts that help them solve these issues by punishing low effort and rewarding high effort. The presence of hard working peers nearby can also mitigate these self-control problems. This set of papers highlights the different channels how better managerial practices can improve worker productivity in the office. However, this literature remains silent about the role of worker sorting in generating productivity gains.

A related strand of the literature considers the development of work structures that accompanied the industrial revolution. These papers argue that some of the expansion of the manufacturing sector at the expense of agriculture, and within manufacturing the movement from the putting out system where manufacturing was done in homes to the factory system, was due to the fact that factory work mitigated worker self control problems that plagued home-based work (see, for example, Clark (1994), Kaur et al. (2010) , Hiller (2011, 2018), Forquesato (2016)). This of course relies on the productivity gains being attributed to factory work itself rather than worker sorting, a hypothesis we test directly.

Finally, there is a small but growing literature on the productivity effects of working from home. Bloom et al. (2014) find substantial productivity improvements from workers in a large Chinese travel agency who were allowed to work from home compared to those who remained in the office. More recently, Bloom et al. (2022) find that hybrid work-from-home at the same firm reduced worker attrition and had small positive impacts on output. These studies differ in two ways from ours. First, the workers in Bloom et al. (2014) and Bloom et al. (2022) were selected from the subset of workers who were already at the firm (and in Bloom et al. (2014), the subset who additionally volunteered to work from home), thereby shutting down the selection channel at the center of our analysis. Second, and closely related, all of these employees were existing office workers who had previously been working in an office environment and thus might have already absorbed office work norms. Our work is complementary to these studies, since we explicitly set out to analyze the role that sorting plays in driving productivity differences between home-based employment and larger formal work settings. In addition, our population has larger productivity differences since many of them are not intimately familiar with office work and have not already learned many of the productivity enhancing work habits that office work may foster.

3 Theoretical Background

In this section we lay out the theoretical motivation for our treatment design. Productivity in a home versus office environment may differ for at least two distinct reasons, namely *treatment effect* and *worker sorting*. As has been documented in the literature, office-based formal employment might provide both a more productive work environment, more opportunities for learning and stronger incentives, independent of the characteristics of the worker. We call this the *Treatment Effect*, which could be due to better monitoring and routines in the office or because learning occurs faster when surrounded by supervisors and peers (explored in Bloom et al. (2014), Kaur et al. (2015), or Schoar (2009)).

However, these same forces that make office-based work environments more productive may also lead to workers sorting on characteristics such as ability. We term this the *Sorting on Ability* effect. We aim to test whether worker sorting is of first order importance and to identify the forces that shape how employees sort in the labor market. There are several theoretical reasons why office-based work might serve as a sorting device. The first is that office work is likely to be more demanding given fixed schedules, strict norms regarding behaviors in the office, and peer pressures. These demands may be less costly and unpleasant for more productive workers. Alternatively, there may be long term benefits from office work due to greater interactions with supervisors that are attractive to more ambitious types. In either case, the office might attract highly productive workers, while low productivity workers remain in home production. Such sorting patterns will have further repercussions if high-ability or high-effort workers are complements in production, either through peer learning dynamics or some sort of O-ring production function. In this scenario, sorting into office and factory work plays a double role in generating productivity differences across firms as it also leads to further productivity gains through grouping high-effort high-ability types together.

Workers may also differ in their preferences for working in the office environment versus a more flexible home environment and these preferences may be correlated with characteristics that relate to productivity. For example, there may be social and cultural sanctions at play that restrict certain groups from either office- or home-based work. In many conservative societies women are not allowed to work outside the house to protect them from interaction with men. Conversely, men who work at home may be stigmatized. The strength of these social and cultural sanctions may vary with household wealth levels. Relatedly, office work might be a status good, particularly so for workers with low social status. All of these forces might generate correlations with ability (e.g., women might be more productive as has been noted in light manufacturing and garment production, or the stigma of working outside the house might be largest for highly educated women).

A separate sorting mechanism operates through workers having heterogeneous returns to office environments and selecting based on these returns, what we call *Selection on Treatment*. If the productivity enhancing features of the office are complements with ability, the most talented workers might select into the office. For example, high-ability workers may learn relatively more than low-ability workers from being close to supervisors. This channel would lead to a positive correlation between worker ability and the likelihood of selecting into office-based employment.

Alternatively, some workers may self-select into a formal work environment if they desire the delayed payoff that comes from disciplined work but know that they are unable to induce the required effort when working from home. Thus, workers with self-control issues or those who find it difficult to create a productive work environment at home, might choose to work in an office-based environment in order to take advantage of the discipline provided by the work environment. In other words they select to the office in order to benefit from the treatment effect. If high-ability workers are more sophisticated about their self-control problems or more patient, we would see sorting of high ability types on the treatment effect.[3] Alternatively, lower ability workers might, on average, have larger self control problems, more distractions at home, or find it more difficult to find a quiet location. Thus, this group will benefit most from the treatment effect of the office. In this case one might expect lower ability workers to sort into the office disproportionately. But under either scenario, we would expect workers who select to the office to see the largest improvements in productivity when they are in the office relative to working at home.

Based on the above discussion, we evaluate the following three forces in our study:

1. Treatment effects of office based work: Evaluate if, independent of characteristics, office based work induces higher levels or faster increases in worker productivity.

2. Sorting on ability: Evaluate whether more able workers sort into office work, and if so, whether such sorting generates meaningful productivity differences between office- and home-based work.

3. Sorting on treatment effects: Evaluate whether workers select into the job environment where they are most productive.

4 Research Setting

4.1 Context and Implementation

This study implements a randomized control trial in the data entry industry in the South Indian city of Chennai. We selected this sector since it provides a number of unique benefits for our analysis: This type of work is very wide-spread in India, has relatively low skill requirements and people are largely aware about this type of work. In addition, due to the discrete nature of the task, the work can easily be done from home using the same technology as in the office setting. The latter is important to ensure that productivity levels between the two types of settings are not driven by more sophisticated interactions between workers or more efficient technology in the office setting. Finally the ease of measuring output from data entry work (e.g., input per minute, errors, time working etc.) makes it an ideal setting to to answer our research questions. Since this type of performance measurement is used in the data entry industry in general, we are not introducing an artificial monitoring system in our setting.

[3] We can think of office work as providing delayed benefits (i.e., higher wages) with an upfront cost (i.e., more effort now).

We established a data entry operation with the option of both home- and office-based work[4]. The operation was managed by professional data entry supervisors who had previously worked in the data entry industry. We also worked closely with a data entry firm in Chennai to set up the office environment and the support structure for the work, e.g. upfront training, technical help with equipment problems, compensation schemes were all modeled after a typical data entry firm in the city. The workers in the office and at home were provided with identical work assignments and identical laptops to complete the data entry tasks in both work environments[5]. In both work environments, the nature of work was kept as constant as possible. Workers were required to work for 35 hours per week in both locations. The type of work, wage structure, the criterion for not being fired, weekly targets, and managers were also identical. In the office environment, we had up to 25 workers working from 9 am to 5 pm for 5 days a week. In the case of the home environment, workers came into the office every Monday morning to submit the work done and receive new assignments. Home-based workers had access to a telephone helpline to call in with problems. Like office workers, workers in the home environment had to work 35 hours per week, but unlike office workers, home workers had flexibility regarding when to work (both within and across days).

To mimic a real job, all workers were offered a contract for 8 weeks of work. After completion of 8 weeks, workers were provided with references and training certificates and were matched to an employment agency to help find future employment in the industry.

4.2 Recruitment and Sample Selection

To hire workers, entry-level data entry jobs were advertised in the jobs section of the main local newspapers. Two different types of newspaper ads were placed—one type advertising for home-based data entry jobs and another type advertising for office-base data entry jobs[6]. The objective was to reach potential employees aged 18-40 who lived in lower middle class localities and suburbs of the city, which is the target group for these types of jobs. Those interested in the job were invited to an in-person interview at the office location. In these recruitment meetings, applicants had to go through a job interview and typing speed tests. The selection process was set up to elicit baseline characteristics and initial typing speed and accuracy. Furthermore, we asked applicants to choose between home and office-based work to elicit their preferences regarding the work environment. This question was incentivized as applicants were told that they would be more likely to get their first choice than their second but it was not guaranteed.[7] All

[4] Pictures of the office and a few sample home settings can be found in the Appendix figure A.1.

[5] A picture of user interface of the a sample data entry task can be found in Appendix figure A.2.

[6] Sample newspaper ads for home-based and office-based jobs can be found in the appendix figure A.3. We found limited heterogeneity based on the type of ad so our analysis combines the workers attracted by both samples although the appendix A.3 reports results broken out by ad type.

[7] Due to an implementation error by the field team, rather than workers being given their preference with a probability of 0.55 they were given it with a probability of 0.5.

workers participated in a three days of paid training at the office location.

[Table 1 about here.]

The experiment recruited 235 workers in total from an applicant sample of 892 over a period of 15 months beginning in January 2017. Workers were hired in batches since the number of office-based workers at a time was constrained by the office size which could accommodate 25 workers. Workers were rejected for the study if they fell outside our age groups or were unable to confirm that they could work in either work environment if they were not allocated to the work location of their first choice[8].

4.3 Intervention

Once work location preferences were elicited, workers' were randomly allocated a work location with a 50 percent probability of getting the preferred work location. Four groups were formed through this process,

- Preferred home, allocated home
- Preferred home, allocated office
- Preferred office, allocated office
- Preferred office, allocated home

The randomization allows us to estimate the treatment effect of being allocated to home or office independent of worker's preferences. Second the allocation to home or office work conditional on worker's preferences allows us to estimate the selection on treatment effect. Specifically, we can compare the difference in productivity between the office and home (using the random assignment) for the group of people who preferred home work to the same productivity difference for those who preferred office work.

The experiment began hiring workers in March 2016 and lasted until April 2018. The hiring of workers was done in waves. For a period of 8 weeks, the productivity of the employees working in both locations was closely monitored using a data entry software platform designed for the study. After the 8-week employment spell, workers were provided with training certificates and referred to other data entry operations if they wished to continue working as data entry operators.

The data entry tasks that data entry workers had to complete were constructed by us. The objective of the construction was to be able to calibrate the difficulty of the data entry tasks. Each data entry task consisted of four sections with each section focusing on a different type of data entry tasks such as typing type-set text, typing strings of random alpha-numeric characters, etc. The difficult tasks had the same

[8] Roughly 50% of the 892 applicants met these two criteria and were invited to the training but only 280 showed for the training. Of the 280 applicants who received training 45 dropped out prior to beginning of the work. Hence the working sample consists of 235 workers.

types of sections as the easy tasks but the difficulty was increased. For instance, the type-set text was replaced by handwritten text, and strings of random alpha-numeric characters were replaced by strings of random alpha-numeric and special characters which made typing difficult[9]. Workers were assigned easy tasks from week 1 to 3, followed by harder tasks in week 4 to 6, and finally, in last the two weeks, a random mix of both the difficult and easy tasks was assigned.

4.4 Compensation Structure

The compensation structure that was provided to workers was designed to mimic a typical data-entry firm in the market. Both office- and home-based workers faced an identical compensation structure. Additionally, both set of worker were compensated for the monetary travel cost that they incurred.

[Table 2 about here.]

The compensation structure consisted of a fixed component and a performance-based variable component. The fixed component was equal to INR 8500 ($ 128.8[10]) per month which the workers were eligible to receive on completing 35 hours per week and a certain target number of data entry tasks. These data entry task targets increased with each week to accommodate learning. If workers failed to meet either of these targets for three weeks their contracts were terminated. For every data entry task completed beyond the weekly target, workers were compensated an additional INR 65 ($ 1) conditional on the fact that they had 100% accuracy and did not perform any error in data entry. This constituted the performance-based variable component of the compensation. To incentivize accuracy of completed tasks, data entry tasks with highest accuracy were counted towards the weekly target and the remaining lower accuracy tasks which constituted the variable component of the compensation were exponentially penalized for errors. The penalty schedule was based on the difficulty of the task.

[Table 3 about here.]

For error rates between 0-7.5% for easy tasks and 0-15% for hard tasks, a proportionate deduction from INR 65 was done. For error rates between 7.5-10% for easy tasks and 15-20% for hard tasks, 1.5 times of error rate was deducted. Finally, for errors greater than 10% for easy tasks and 20% for hard tasks, 2 times of error rate was penalized. A retention bonus of INR 2000 ($ 30.3)was paid after the completion of week 1.

Incremental modifications were made to the salary structure of workers in the early waves, primarily to address (initially severe) issues with selective attrition, specifically that workers allocated their non-preferred work type would quit before starting work.[11] Ultimately, a retention bonus at the end of the

[9] Examples of data entry tasks for both difficulty levels can be found in the appendix A.5

[10] We use the average exchange rate between Indian Rupee and United States Dollar during the period of experiment which is INR 66 $\approx$ $ 1.

[11] Appendix A.4 presents a complete list of these modifications.

first week of work solved these issues with no statistical difference in attrition between those allocated their preferred work location and those not (see Section 4.6 for these results and for further analyses and discussion of attrition). Our analysis presented in the paper mainly focuses on the sample after this retention bonus was put in place with results for the earlier (high attrition) waves relegated to the appendix table A.3.

4.5 Outcome Measures

As part of the hiring process, we administered a short survey collecting information on demographics, education, data entry and other work experience, employment status, job search, work preferences, and family care and other time commitments. Further, upon selecting the workers, during training, a baseline survey collected information in greater depth and covered additional domains such as household characteristics and income, social and economic status and computer literacy. Along with the baseline survey, selected workers had to take an aptitude test, a personality test, a risk preference test, and a time preference test. To gauge the baseline ability of applicants and eventually workers, several speed tests were used before random allocation to home/office. As mentioned earlier, all applicants were required to do a typing speed test along with the job interview. Furthermore, during training workers were required to complete a cash-incentivized and non-incentivized typing speed.

A variety of data entry job outcomes were collected over the 8 week work period. The data entry job outcomes that were collected over time were the same as the ones the data entry companies collect for administrative purposes and so they did not pose additional burdens on the workers. We hired developers to create proprietary data entry software which kept detailed logs of data entry tasks, keystrokes typed, accuracy and time spent working or idle for each worker. The measure of accuracy is defined to be the proportion of correct entries to total entries. The main productivity measure that we use is net typing speed which is defined as correct entries typed per minute. An attendance record for both home and office workers was maintained with which attrition was measured.

4.6 Attrition

In the early waves, we experienced substantial attrition in the first few days of work and the attrition was heterogeneous across groups. The left panel of figure 1 highlights the attrition problem by plotting the proportion of workers left in each intervention group across 8 weeks of employment. The group that was experiencing the most attrition was workers who preferred to work from home but were randomly allocated to work from the office. In the first 3 waves, in total 49 workers were part of this intervention group but only 16% of the workers began the work whereas 41 workers quit the job the moment they got to know they were allocated to work from the office. Additionally, the attrition rate post allocation and before the beginning of work was fairly high around 30% for workers who preferred working in the office irrespective of their work location assignment.

[Figure 1 about here.]

To address attrition and incentivize workers to stay longer a retention bonus of INR 2000 ($ 30.3) was introduced which was paid upon the completion of the first week of work. This amount was approximately equivalent to the average weekly earnings of workers. Additionally, selection criteria were tweaked as well[12]. These changes were able to mitigate the attrition problem for all intervention groups as it is evident in the right panel of figure 2.

[Table 4 about here.]

To extend this analysis further, in table 4 we explore whether the number of days worked by 280 workers, who were selected from applicants in the post-retention bonus waves, can be explained by any of 3 attributes- type of advertisement, preference about work location and assignment of work location. Across all columns, the regressions control for wave fixed effect and cluster the standard error using the variable wave. Column 1 in the table 4 considers all the 3 attributes simultaneously. We find that workers who responded to home-based work ad on average work for 3.29 fewer days compared to the ones who responded to office-based work. Though this effect is not statistically significant. Further, workers who preferred home and workers who were assigned home on an average ended working 0.58 and 0.71 additional days respectively. None of these effects are statistically significantly different than zero. Similarly, the subsequent three column, (2)-(4) consider each attribute individually. In summary, none of the 3 attributes seem to cause a significant difference in the number of days workers worked in waves with a retention bonus.

Of the 280 workers in waves with retention bonus, 45 quit during the training and prior to the beginning of the work. Hence we have working data for 235 workers. Going forward we will focus on workers from these later waves, although we report the main results for all waves in Appendix table A.3.

5 Results

5.1 Baseline Characteristics

We first present a balance table that compares baseline characteristics for the groups of workers assigned to home and work locations.

Columns (2)-(4) of table 5 compare the 124 workers who were randomly assigned to work from home to the 111 workers who were randomly assigned to work in the office. The two groups are reasonably but not perfectly balanced. Critically, we see no difference in initial worker performance, either measured by

[12] Also reduced job duration from 12 to 8 weeks to reduce the time between waves when computers were unused for attrited workers.

the speed test during their initial interview or the two speed tests administered as part of training including once that was incentivized through cash payments based on performance.[13] We also find no differences in proportion of workers who preferred to work from home in the randomly assigned groups. Across both assigned work locations, 37% of workers preferred to work from home. There are three (out of 22) characteristics where we do find significant differences. Of the workers who were assigned to work from home, 58% are women whereas only 43% are women in the assigned-office group (significant at the 2% level). The home group has 6% fewer workers with family care responsibilities and 7% more workers who have used a computer before. These differences are significant at the 7% and 3% level, respectively.

[Table 5 about here.]

The last 3 columns of Table 5 compare 87 workers who preferred to work from home to the 148 workers who preferred to work from office. Not surprisingly, these groups considerably differ from each other, since these preferences are not randomized but are in part explained by people's characteristics as discussed above. Demographically, workers preferring home are older by 1.8 years, have a higher proportion of married individuals by 16%, and have higher proportion of individuals with family responsibilities by 6%. First two of these differences are significant at 1% level and the last one is significant at 5% level. In line with home-preferring workers being older they also seem to have more years of work experience and higher number of office jobs held previously. Both these differences are significant at 1% level. Finally, 5% fewer workers preferred to do a full time job when workers who preferred to work from home were compared to workers who preferred to work from office. This difference was significant at 10% level. We explore differences in the speed tests across these two groups when analyzing selection on ability.

5.2 Treatment Effects

To estimate the impact of working from home on worker performance, we run the following specification:

$$\text{Worker Performance}_{i,t} = \alpha \text{Alloc\_home}_i + \gamma X_{i,t} + \epsilon_{i,t} \tag{1}$$

Where Alloc_home_i is a binary variable that takes value equal to one if the worker was randomly assigned to work from home; $X_{i,t}$ captures three different fixed effects that we control for, namely, Wave fixed effects capturing the batch the time the worker was hired (either wave 3.5 or wave 4), Week fixed effects capturing the week of employment the outcomes is measured at (e.g., week 1–8)[14] and Section fixed effects

[13] Each worker attempted 3 typing speed tests in our data entry software. The first test was conducted during the recruitment along with an interview. This was hour long typing test that a novice with no introduction to data entry could complete. The next two tests were shorter 25 minute long tests which the workers attempted after couple of hours of couching during the first day of training. Both these test had identical format except the second one incentivised workers by paying them for the total correct characters they type.

[14] We use week fixed effects instead of finer day fixed effects because home workers had the freedom to work any day during the week. Thus, it does not make sense to compare work done on a particular day of the week across home and office workers when their daily schedule differs.

capturing the difficulty of the data entry task being performed (easy or hard).

Although our unit of observation is a particular data entry task, a survey that is entered into the software, we reweight the regressions so that each workers has an equal weight and standard errors are clustered at the level of the individual.

[Table 6 about here.]

We explore a range of outcomes. Our primary measure of worker performance is log(Net Speed) where Net Speed is defined as correct entries typed per minute. We find that the employees allocated to work from home experience a drop of 18% in net speed (Table 6 col (1)). This effect is statistically significant at the 1% level. The next two performance measures that we consider are Gross speed and Accuracy which are defined as total entries typed per minute and the ratio of correct entries typed to total entries typed, respectively[15]. Employees working from home see a drop of 12% in the Gross speed (Table 6 col (2)) and 2.48 % in accuracy (Table 6 col (2)). Thus, the drop in net speed is mostly attributable to the drop in gross speed although accuracy also plays its part.

The magnitude of the treatment effect is larger when we use alternative measures of worker performance. In column (7) of Table 6, we explore whether the treatment effect changes with the difficulty of the underlying data entry task by limiting the sample of data-entry tasks only to hard tasks (which would require workers to concentrate harder and expend higher cogitative effort). We find that participants assigned to working from home display 30% lower net speed on hard tasks. To incentivize workers to make fewer errors, in keeping with industry practice, we imposed an exponentially increasing penalty for remuneration. Thus employees were paid on basis of remunerated speed and not net speed. We find that the magnitude of the treatment effect is as high as 24% when measured by remunerated speed that punishes errors more heavily than net speed (Table 6 col (8)).

One key benefit of working from home is the flexibility that it affords workers concerning their time. We consider three outcomes pertaining to time use. First, we explore how time worked entering data per week changes across home and office. Irrespective of work locations, all employees were mandated to work 35 hours per week. Though we had an imperfect monitoring system for these 35 hours, the data entry software kept an accurate record of time spent while doing data entry. We find that employees across both locations spent 33.7 hours actively entering data while 1.3 hours were spent on ancillary tasks pertaining to data entry with no significant difference in this behavior across both work locations (Table 6 col (4)). Second, though the total time spent entering data does not differ across work locations, when the time is spent differs substantially. Individuals working in the office completed 97% of their work during office hours (i.e., between 9 am to 6 pm from Monday to Friday) (Table 6 col (5)). The remaining 3% of the work that occurred during non-office hours was due to office employees compensating for public

[15]Net speed, Gross speed and Accuracy are related to each other by following expression: Net speed $=$ Gross speed $*$ Accuracy.

and personal holidays. On the other hand, only 46% of the work done by home employees was done during office hours indicating that workers used the flexibility afforded to them by the working from home option. Finally, along with choosing when to work, working from home provides workers greater autonomy around breaks taken during working hours and perhaps helps workers deal with moderate distractions. The software was programmed to measure intervals of time if no action was performed by the worker using either the mouse or the keyboard while entering data. The ratio of the total time spent in such intervals to the total time spent entering data is defined as idle time. We consider idle time to be a measure of breaks taken while typing. Employees working from the office spend 14.6% of their time idle and this rises by 2.46% for those working from home, indicating that home workers took more breaks (Table 6 col (6)).

[Table 7 about here.]

Next, we run several robustness checks to explore the sensitivity of the treatment effect to different specifications. First, we control for workers' baseline speed during the cash-incentivized speed test. This control should increase precision and control for bias if, despite randomization, the initial level of performance differences are driving the productivity drop. The 18% drop in net speed persists with a marginal decrease in the standard error of the coefficient (Table 7 col (2)). Further, we expand our sample of observations to include performances of workers from pre-retention bonus waves as well. Column (3) of Table 7 shows that the treatment effect remains at -18%, unchanged. The average treatment effect that we estimate, re-weights each observation of the data entry task such that each employee has equal weight in the estimation of equation 1. Thus an individual observation of a data entry task of a worker performing a higher number of tasks would get a lower weight compared to an individual observation of a data entry task of a worker who has performed fewer tasks because of slower typing speed or early attrition. Instead, we use task-based weighting in column (4) with each survey task carrying equal weight. The table shows that the 18% drop in productivity from working from home is almost unchanged, falling only slightly to 16%. Finally, as we discussed earlier, the two groups formed by randomly assigning work locations are reasonably but not perfectly balanced (columns (2)-(4), Table 5). In column (5), we control for three characteristics across which the two groups are not balanced. These three characteristics are gender, family care responsibilities and prior computer usage. Controlling for these characteristics, we find that the treatment effect marginally increases to 19%.

5.2.1 Cumulative Learning

Workers in both home and office locations experience an increase in productivity over the period of employment. This can be seen in the left panel of the figure 2 which plots the average speed of workers in both locations over the 8 week employment period. The large drop in speeds across both work locations in week 4 is due to the fact that the workers were assigned easy tasks from week 1 to 3, followed by hard tasks in week 4 to 6, and finally, in the last two weeks, a random mix of both difficulty tasks was assigned. To address the change in difficulty of the data entry tasks, the center panel of figure 2 separately plots the average speed for each difficulty level for each work location across the 8 weeks. Finally, the right panel of figure 2 plots cumulative learning across both difficulty types for both locations along with weeks since that level of difficulty was introduced. Cumulative learning is defined as the percentage increase in net speed compared to the baseline speed for that particular difficulty type. It is evident from all panels of figure 2 (but especially from the right panel) that the learning, in both locations and for both difficulty types, is high in earlier weeks and then plateaus in later ones.

[Figure 2 about here.]

Despite learning over 8 weeks of employment, working from home remains less productive throughout the period of employment. It raises an interesting question—how much of this productivity difference is acquired due to differential learning? and how much of it existed at baseline? To answer this question, we use the same specification as equation 1 where we consider three different measures of worker performance for all tasks completed over 8-week of employment. These finding are reported in columns (1) - (3) of Table 8. The first measure of worker performance is the log of net speed, results for which are reported in column (1). The second measure is the log of net speed at the baseline which is defined as the average net speed for the initial four surveys completed for particular difficulty level by each data entry operator[16](column (2)). Finally, the third measure is cumulative learning which is defined as the percentage increase in net speed compared to the average net speed of the initial four surveys completed (baseline speed) for that particular difficulty type (column (3)). We find that of the 20% lower typing speed that home workers experience in comparison to office workers, 13% exists from the first day whereas 7% is due to differential accumulated learning[17].

[Table 8 about here.]

Deeper inspection shows that almost all the differential learning happens in regard to the difficult data-entry task. Focusing solely on easy surveys completed by workers, the home workers are 10% slower, most of which manifests itself immediately with negligible learning effects (col (4)-(6), table 8). Similarly, a sole focus on the completion of hard surveys illustrates that the productivity difference between home and office based workers is 19%. Furthermore, office workers have 14% higher cumulative learning thus leading to an overall productivity difference of 33% (col (7)-(9), table 8).

[16]It takes approximately 8 hours to complete 4 surveys.

[17]The negative 20% treatment effect reported in column (1) of the Table 8 is slightly different than the treatment effect reported in column (1) of the Table 6, because regressions in the Table 8 are weighted such that the treatment effect can be precisely decomposed in baseline difference on day 1 and the subsequent acquired difference due to learning.

5.2.2 Daily and Weekly Work Patterns

As mentioned earlier, home based workers use the flexibility afforded to them by their work location. The smallest share of work was done on Mondays as home based workers were required to visit the office to upload the data entry tasks completed in the prior week and to receive a new set of tasks to complete in the following week (Figure 3 top-left panel). Interestingly, in a typical week, for home-based work the proportion of work done steadily rises as Monday approaches, the highest proportion of work being done on Sunday and followed by Saturday. The proportion of work done past official working hours (i.e. after 6 pm) is roughly similar to the proportion of work done during official working hours (Figure 3 top-right panel). However, there is a steady decline as we approach midnight. Although a sizeable proportion of work is done during night time from midnight to 8 am, this proportion is much lower than the proportions from 9 am until midnight.

[Figure 3 about here.]

The productivity of office based workers steadily rises over the week and always remains higher than the productivity of home based workers for all days (Figure 3 bottom-left panel). Home based workers are approximately equally productive all days except on Mondays and Sundays when they experience a more than 5% drop in productivity. On a typical day, the productivity of office based workers is highest once they get settled by 10 am and steadily declines over the work day (Figure 3 bottom-right panel). But it always remains higher than that of home based workers (with an exception of productivity at 8 am where the amount of work done is almost negligible). On other hand, the productivity of home based workers roughly remains constant with a considerable drop being observed during 2-4 am.

5.3 Selection on Initial Ability

Next we turn to the question of how workers sort into office versus home work on the basis of their innate ability. We run two specifications to investigate if higher ability workers select into office work. The first specification regresses initial worker performance on stated preferences for home work:

$$\text{Initial Worker Performance}_{i,n} = \beta \text{Pref\_home}_i + \gamma \text{Test}_n + \epsilon_{i,n} \tag{2}$$

where Initial Worker Performance$_{i,n}$ is log of net speed from different speed tests that were conducted prior to beginning the job; Pref_home$_i$ is a binary variable representing the work location choice of the employee, and takes value equal to one if the worker preferred to work from home and is equal to zero otherwise; Test$_n$ is fixed effect for which of the three different speeds tests that were administered to employees.

[Table 9 about here.]

We present the results of estimating of equation 2 in columns (1)-(3) of Table 9. First, we consider the sample of all applicants who showed up for walk-in interviews and did a speed test. This amounts to a total of 884 applicants. As the coefficient on Pref_home in column (1) indicates, contrary to the hypothesis that there may be positive selection on ability into office work, applicants preferring home-based work are 15% faster in whatever location they are allocated to compared to applicants preferring office-based work. This effect is significant at 1 % significance level. In column (2), we restrict our sample to only include the 234 workers who moved forward to training.[18] When the restricted sample is considered, the selection effect persists although it is diminished to 10% (significant at the 5 % level). Our preferred selection specification is presented in column (3) where we include speeds from the three different speed tests conducted prior to the start of work (including the cash-incentivized test). As two of these tests were part of the training, we focus only on the sample of workers who progressed from the interview stage and started training. We find workers preferring home are 12% faster than workers preferring office (col (3) Table 9). This effect is significant at the 1% significance level. In sum, whether we look at the full sample of job applicants or those ultimately selected for work, we see that more productive worker at baseline are more likely to prefer working from home.

We next investigate whether the same selection effect is present in performance of employees over the subsequent two months of employment. To do so, we run the specification given by equation 3:

$$\text{Worker Performance}_{i,t} = \alpha \text{Pref\_home}_i + \gamma X_{i,t} + \epsilon_{i,t} \tag{3}$$

where $X_{i,t}$ captures the wave, week and section fixed effects. The measure of worker performance is log of net speed. Results for this specification are reported in columns (4) and (5) in Table 9. Though slightly smaller in magnitude than the initial difference, we again find those who prefer home work perform better in whatever location they were assigned to, with 8.4% higher speed (statistically significant at 10% level), see Table 9 column (4). Finally, in column (5) of table 9 we explore what happens to this selection effect once we control for the work location they were allocated to. Since the allocation of work location is randomized and so should be uncorrelated with preferences, it is reassuring that controlling for work location barely changes the magnitude and significance of the selection effect.

[Table 10 about here.]

Apart from self-reported work location preferences during walk-in interviews, an additional measure of selection that we recorded was the type of advertisements applicants responded to (recall that some weeks we posted adds highlighting work from home opportunities, other weeks office opportunities). In Table 10, we run an identical set of specifications as the preceding table 9 except we control for the selec-

[18] The actual number of workers in post-retention bonus waves is 235. For one worker we are missing walk-in speed test results.

tion effect introduced by the type of newspaper ad workers responded to. On adding the type of newspaper ad control, the selection effect based on self-reported work location preference in the 884 applicants sample declines to 12% from 15% in Table 9. The selection effect is still significant at the 1% significance level. Next, in columns (2) we run the same specification but restrict the sample to only workers who attended training and in columns (3) we further include the two additional baseline speed measures for each worker. The selection effect based on self-reported work location preferences remains unchanged in both the magnitude as well as statistical significance. Finally, columns (4) and (5) of Table 10 repeat the exercise for the the full employment period with a little change from the earlier results. In summary, the selection driven by self-reported work location preferences remains similar when controlling for the type of newspaper ad the applicant responded to.

The variation generated by the types of newspaper ad gives us another dimension of selection to explore, rather than just a confound to our previous selection analysis. In the applicant sample, we find that applicants responding to home-based work newspaper ads are faster than the applicants responding to office-based work ads. The direction of selection driven by the type of ad responded to is the same as the selection driven by self-reported work location preference. Column (1) of table 10 illustrates the selection in responding to home ads is 7.6% and is significant at 1%.[19].

[Table 11 about here.]

Finally, we explore how the selection on work location preference effect manifests itself in other productivity measures. Overall, just like net speed, both applicants' and workers' samples reflect a positive selection on home-based work in gross speed, accuracy, and proportion of idle time as well. Table 11 reports these results. Columns (1) and (2) show that gross speed is higher for those preferring home work by 14% and 9.5% in the applicant and worker samples, respectively. Applicants and workers chosing home are also more accurate than their counterparts choosing office (columns (5) and (6), Table 11) but these differences are not statistically significant. Columns (7) and (8) show that applicants and workers preferring home have -1.96% and -1.57% lower idle time indicating that home-preferring applicants and workers take fewer breaks while typing (Table 11 columns (7) and (8)).

5.4 Selection on Treatment Effect

Given that we find negative selection into office work based on initial ability, one obvious explanation would be that the low ability workers benefit more than high ability workers from being in the office

[19] We do not further consider the selection driven by advertisement type in part because our filtering process to select workers from the applicant sample tampers the selection by ad type. This is evident in columns (2) and (3) of the workers' sample where the selection driven by ad type operates in the opposite direction. The applicants from home-based work adverts are 7.5% faster than applicants from office-based adverts but the selected workers from home-based work adverts are 4.3 to 5.6 % slower compared to the workers selected from office-based work advert. This negative selection effect towards home based work ad grows even stronger in worker performance during 8 weeks of employment.

and so they are more likely to choose to work in the office. These higher returns might come from facing more distractions at home or more need to learn and get help from others around them. This selection on treatment would mean that the workers who prefer office-based work should see greater improvements in their productivity if they are allocated the office, compared to workers who prefer home-based work. To discover whether this is true, we explore heterogeneity in treatment effects by preference for office- or home-based work:

$$\text{Worker Performance}_{i,t} = \alpha \text{Alloc\_home}_i + \delta \text{Pref\_home}_i + \lambda \text{Pref\_home}_i * \text{Alloc\_home}_i + \gamma X_{i,t} + \epsilon_{i,t} \quad (4)$$

Where the coefficient λ on the variable $\text{Pref\_home}_i * \text{Alloc\_home}_i$ captures the selection on treatment effect; Alloc_home_i and Pref_home_i are binary variables taking value equal to 1 when worker was randomly assigned to work at home and when workers self-reported preference is to work from home, respectively; $X_{i,t}$ capture week, section and wave fixed effects. The four $\text{Worker Performance}_{i,t}$ measures that we consider are log of net speed, log of gross speed, accuracy, and proportion of idle time.

[Table 12 about here.]

Results of the specification given by equation 4 are presented in Table 12. In column (1) controlling for self-reported location preferences, we see that workers who prefer to work in the office but are allocated to home-based work exhibit 14% lower net speed compared to the office-based workers. However, for workers who prefer to work from home and are allocated to work from home, i.e. they were allocated to their preferred allocation, the same difference increases by a further 12% though this additional increase in difference is not statistically significant. But in column (2), once we control for baseline net speed, the selection on treatment becomes statistically significant at 10%, and the magnitude of difference in net speed between home- and office-preferring workers increases to 14%. Similar negative selection on treatment can be observed in the case of gross speed with the difference being 8.1% without baseline gross speed control and 13% with baseline gross speed control in columns (3) & (4) of the Table 12, respectively. Finally, home-based workers who preferred home, compared to the home-based workers who preferred office, exhibit an accuracy lower by 2.17% and a higher proportion of idle time by 1.88% (though non of these effects are statistically significant (Table 12 col (5) & (7)). Controlling for baseline accuracy and baseline proportion of idle time in columns (6) and (8), the selection on treatment results do not change.

In sum, people who prefer home based work experience a larger increase in productivity when they work from the office compared to people who are allocated to home work, even though they preferred to work at home. We find a parallel, but crucially smaller effect, for people who prefer office work. Thus selection on treatment is negative and does not explain the negative selection on baseline ability explored in the preceding section. Our results do not support the hypothesis that lower ability workers or those with self-control problems choose the office as they stand to gain the most from doing so. Instead the results again point to an explanation where some workers might be constrained from choosing the most efficient work location.

5.5 Heterogeneity

The prior analysis shows that despite a strong positive treatment effect of the office, high ability workers are not more but less likely to select into the office. Furthermore, high ability workers are the ones who benefit most from being in office. In this section, we explore what might be the origin of the negative selection effect and the negative selection on treatment effects.

5.5.1 Heterogeneity in Selection on Initial Ability

We explore the following set of potential hypotheses that could lead to negative selection on ability:

1. High ability workers tend to live further away and so incur higher time and effort costs commuting to the office.
2. The office serves as a disciplining device for low self-control/ low-productivity workers.
3. Office work is a status good for lower ability workers.
4. High-ability workers combine the job with searching for better job opportunities which is easier with a more flexible schedule.
5. Distractions at home are larger for low ability workers.
6. High ability women face greater social sanctions or pressures to work inside the home.

In order to explore if any of the above hypotheses explains the selection towards home work, we include proxies for these underlying characteristics in our selection regression. The first hypothesis (high-ability workers have higher commute times) is explored using the distance between home and office[20] reflecting the idea that commute costs would be directly proportional to the travel distance. The second hypothesis (office serves as discipline for low self-control/ low-productivity types) is investigated using two proxies– first, the response to the statement "I never leave things to the last minute" on a personality test that asked workers to rank statements expressing various positive attributes, and second, the time discount parameter estimated using a questionnaire created by Andreoni et al. (2015). The third hypothesis (office work is a status good for low-ability workers) is explored using a variable for the number of previous office jobs the worker has done and the total monthly income of all the members of the worker's household. The fourth hypothesis (high-ability workers choose homework to search for better job opportunities) is

[20] Workers were compensated for incurred monetary travel costs.

captured by two variables which are worker's answer in the entry survey to the questions whether they prefer to work full-time or part-time, and whether they have time commitments such as job search and study. The fifth hypothesis explores whether low-ability workers experience greater pressures at home. For this, we consider four variables, namely, if the worker has family care commitments, if the worker is married, if the worker has kids, and the worker's age. We anticipate that being old, married, and having kids will lead to greater home responsibilities. The final hypothesis that we consider is that high-ability women experience greater pressures at home due to social norms and are attracted to home due to their preference for the flexibility that home work offers. To explore the final hypothesis, we consider five variables. These five variables are if the worker is female, if the worker is female who has family care responsibilities, if the worker is a married female, if the worker is a female with kids, and finally, the age of female worker.

If some of these hypotheses explain workers selection decisions, and individual characteristics are good proxies for these, they should attenuate the coefficient on the preference for home work. Specifically, we run the following specification

$$\text{Initial Worker Performance}_{i,n} = \beta \text{Pref\_home}_i + \sum_h \gamma_{1,h} \text{Characteristic}_{i,h} + \gamma_2 \text{Test}_n + \epsilon_{i,n} \quad (5)$$

Where Pref_home_i takes a value equal to one if the worker prefers to work from home; $\{\text{Characteristic}_{i,h}\}_{h=1}^{H}$ denotes the set of characteristics that proxy for the hypothesis; Initial Worker $\text{Performance}_{i,n}$ is the log of speed in the three initial speed tests done by workers; Test_n are fixed effects that control for each of the speed tests.

The findings of this specification are reported in Table 13 which consists of three panels. Panel A reports the coefficient and the standard error of the estimated selection effect from the baseline specification in columns (3) and (4), respectively. These results replicate Table 9 column (3) and the magnitude of the selection effect is 11.6% (significant at 1 % level). Panels B explores how this selection effects is attenuated when we add controls proxying the hypotheses above, and Panel 3 explores the attenuation when all controls are added simultaneously.

[Table 13 about here.]

First, we discuss Panel B of Table 13. Column (1) lists the hypothesis. Column (5) lists the characteristics that are used as proxies and column (6) reports the coefficient estimates for the selection effect, Pref_Home, conditional on these proxies. Rows (1), (4), (7), (10), (15) and (21), in columns (5)-(7), represent the estimated selection effect from specifications that simultaneously control for all characteristic proxies that we use for a particular hypothesis. As the number of characteristics representing a particular hypothesis varies, we also conduct a principal component analysis using each hypothesis's complete set of characteristics and use the first component that accounts for the most variance to control for the hypothesis. The results of this comparison exercise are presented in rows (1), (4), (7), (10), (15), and (21) of columns (2)-(4).

We now report the summary of results from the heterogeneity analysis of our different hypotheses: The individual characteristic for which we see the biggest attenuation in the selection effect when controlled for is the number of previous office jobs the worker has done. As seen in row (5) of column (6), the selection effect attenuates to 9.1%, which is an attenuation of 2.5%.

Crucially, controlling for the complete set of characteristics representing the two hypotheses that high ability workers face greater distractions at home and that office work is a status good leads to the highest attenuation in the selection effect. First, accounting for the characteristics representing distractions at home attenuates the selection effect by 3.2% resulting in an estimate of 8.4% shown in row (15) of column (6). The significance drops to 5% significance level. Controlling for family income and previous office job experience, which proxy for status that workers associate with working in our office, attenuates the selection effect by 2.7%. Row (7) of column (6) shows that the coefficient of Pref_Home drops to 8.9% and is significant at 1% level. As mentioned earlier, to fairly compare different hypotheses represented by varying numbers of characteristics, we control using the first principal component of the set of characteristics representing a particular hypothesis. Controlling for the first principal components shows that for both hypotheses, the selection effect coincidentally attenuates to 10.9 % (row (7) and row (15) of column (3)). This is the highest attenuation in the selection effect while controlling for the first principal component across all hypotheses except one[21].

In the bottom panel of Table 13, panel C, we present the selection effect when we control for all the hypotheses simultaneously. When we use the complete set of characteristics for all hypotheses, we find that the selection attenuates by 6.5% to 5.1% and is significant at 10% significance level (columns (6) & (7)). Thus, to summarize, these hypotheses can help explain part of the selection effect but far from its entirety. Among the different hypotheses, that high ability workers face greater distractions at home, and that office work is a status good for low ability workers have the greatest explanatory power. This conclusion is supported by both, using the complete set of characteristics proxying an hypothesis as control and, a fairer comparison, using the first principal components of the complete set of characteristics as control.

5.5.2 Heterogeneity in Treatment Effect

It is also valuable to study heterogeneity in treatment effects along these sames dimensions of heterogeneity. We divide our sample into two subgroups across all characteristics. For instance, using characteristic such as age we bisect our sample into young and old, and using characteristic such as family income we bisect our sample into low and high family income. We then interact our treatment effects with a dummy

[21] The highest attenuation of selection effect takes place when the hypothesis that office serves as disciplining device for low self-control/ low-productivity workers is considered. Controlling for the first principal component of set characteristics representing this hypothesis shows that the selection effect attenuates to 10.1% but is still significant at 1% level (row (4) of columns (2)-(4)).

for membership of one of the subgroups:

$$\text{Worker Performance}_{i,t} = \alpha \text{Alloc\_home}_i + \alpha' \text{Alloc\_home}_i * \text{sub\_group}_i + \theta \text{sub\_group}_i + \gamma X_{i,t} + \epsilon_{i,t} \tag{6}$$

Where the measure of Worker Performance that we consider is the net typing speed of workers; Alloc_home$_i$ is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home; $X_{i,t}$ captures three different fixed effects that we control for, namely, Wave fixed effect captures the variation in hiring batches of the workers, Week fixed effect captures the variation in the duration of the employment and Section fixed effect captures the variation in data entry task being performed. The coefficient of Alloc_home interacted with the sub_group provides an estimate of treatment heterogeneity by subgroup. The result of this estimation is presented in Table 14.

[Table 14 about here.]

Panel A of the Table 14 in column (3) presents the baseline average treatment effect of an 18% lower productivity due to home assignment. This effect is significant at 1% significance level and was presented earlier in Table 6 in column (1). Panel B of Table 14 presents results of the new specifications given by equation 6 which help us relate the Average treatment effects (ATE) estimated using equation 1 to Conditional average treatment effects(CATEs) based on subgroups $Z \in \{0,1\}$. Specifically,

$$\text{ATE} = \text{CATE}(Z=0) + Z * \underbrace{[\text{CATE}(Z=1) - \text{CATE}(Z=0)]}_{=\Delta\text{CATE(Z)}}$$

For the all the characteristics the CATE$(Z=0)$ and ΔCATE(Z) are presented in column (3) and (4) of Panel B in Table 14. The headers of the columns are named CATE and ΔCATE(Z) which are representing coefficients of variables Alloc_Home and Alloc_Home $*$ subgroup in equation 6. If heterogeneity in treatment effect exists between two subgroups then the coefficient of the variable Alloc_Home$*$subgroup would be statistically different than zero. For instance, considering gender as a characteristic, we investigate whether the allocation effect of working at home (which leads to an 18% lower productivity) is different for men and women? These results are represented in first two rows in Female Constaints section of the table of columns (1) and (2) of panel B. According to which men experience a 16% drop in typing speed but women experience an additional drop of 7% though this additional drop is statistically insignificant.

We find limited evidence for heterogeneity in the treatment effect. There are three characteristics where we find the coefficient of ΔCATE(Z) to be significantly different from zero at 10% significance level. This means that the two subgroups created, by splitting the sample along the dimension of these characteristics, have statistically distinct treatment effect. These characteristics are the monthly family income of the worker's household, female workers with family care responsibilities, and older female workers. Interestingly female workers with family care responsibilities is the only subgroup that exhibits higher productivity while working at home compared to office. These workers are 11% faster at home although this treatment effect is not statistically significant. Similarly, subgroups such as workers with family care responsibilities, workers preferring to work part-time and workers with low family income experience the smallest treatment effect at -5, -8 and -10 %, respectively. These effects are statistically indistinguishable from zero. On the contrary, subgroups such as older female workers, workers with high family income and married workers exhibit the strongest treatment effect at -38, -28 and -25 %, respectively. All these effects are significant at 5% significance level.

5.5.3 Heterogeneity in Selection on Treatment

To shed light on the negative selection on treatment we now explore whether our findings come from comparisons between groups that might have different constraints when choosing an optimal work locations. We run the following specification to see if accounting for characteristics or constraints explains the selection on treatment,

$$\begin{aligned}\text{Worker Performance}_{i,t} &= \tau \text{Alloc\_home}_i + \delta \text{Pref\_home}_i + \lambda \text{Alloc\_home}_i * \text{Pref\_home}_i + \\ &\sum_h \gamma_{1,h} \text{Alloc\_home}_i * \text{Characteristic}_{i,h} + \sum_h \gamma_{2,h} \text{Characteristic}_{i,h} + \gamma X_{i,t} + \epsilon_{i,t}\end{aligned} \tag{7}$$

Where the coefficient λ on the variable $\text{Pref\_home}_i * \text{Alloc\_home}_i$ captures the selection on treatment effect; Alloc_home_i and Pref_home_i are binary variables taking value equal to 1 when worker was randomly assigned to work at home and when workers self-reported preference is to work from home, respectively; $X_{i,t}$ capture week, section and wave fixed effects;$\{\text{Characteristic}_{i,h}\}_{h=1}^{H}$ denotes the set of characteristics that proxy for the hypothesis. The $\text{Worker Performance}_{i,t}$ measures that we consider is log of net speed.

[Table 15 about here.]

The estimated selection on treatment effect for various specifications are displayed in Table 15 in column (3). The first row indicates the baseline coefficient of $\text{Alloc\_home}_i * \text{Pref\_home}_i$ which is -14% and significant at 10% significance level. This estimate was earlier presented in Table 12 in column (2). We again proxy for the same characteristics and hypotheses as in the prior heterogeneity analysis. When we control for our proxies of constraints and characteristics the negative selection on treatment persists. In the last row, we control for all characteristics across all hypotheses and find that the magnitude of the selection effect increases to -18% and the coefficient is significant at a 5% significance level.

Finally, we explore the heterogeneity in selection on treatment effect across various subgroups. In an effort to do so we run the following specification,

$$\begin{aligned} \text{Worker Performance}_{i,t} = \tau \text{Alloc\_home}_i + \delta \text{Pref\_home}_i + \lambda \text{Pref\_home}_i * \text{Alloc\_home}_i + \\ \tau' \text{Alloc\_home}_i * \text{sub\_group}_i + \delta' \text{Pref\_home}_i * \text{sub\_group}_i + \\ \lambda' \text{Pref\_home}_i * \text{Alloc\_home}_i * \text{sub\_group}_i + \theta \text{sub\_group}_i + \gamma X_{i,t} + \epsilon_{i,t} \end{aligned} \quad (8)$$

Where the coefficient λ on the variable $\text{Pref\_home}_i * \text{Alloc\_home}_i$ captures the selection on treatment effect; Alloc_home_i and Pref_home_i are binary variables taking value equal to 1 when worker was randomly assigned to work at home and when workers self-reported preference is to work from home, respectively; $X_{i,t}$ capture week, section and wave fixed effects; sub_group_i is binary variable that is created to bisect the sample along dimension of the particular characteristic. The Worker $\text{Performance}_{i,t}$ measures that we consider is log of net speed. The coefficient of $\text{Pref\_home}_i * \text{Alloc\_home}_i$ and $\text{Pref\_home}_i * \text{Alloc\_home}_i * \text{sub\_group}_i$ are of interest to us and we report those in Table 16 as CATE and ΔCATE.

[Table 16 about here.]

Though negative selection on treatment is present in almost all subgroups, it is magnified in certain groups. These subgroups are female workers with family care responsibility, workers with family care responsibility, workers with kids, older workers, and workers with low family income. Compared to the baseline selection on treatment effect of -14% (table 16, panel A, column (3)), these subgroups respectively exhibit selection on treatment effect of -92, -83, -55, -38 and -34 %. All these effects are significant at a 1% significance level.

As evident from table 15, controlling for various characteristics associated with constraints does not explain the negative selection on treatment. Thus, on average, these subgroups don't appear more constrained but the subsequent exploration in table 16, indicates that within certain subgroups there exists greater heterogeneity in degree to which the constraints are binding. Such heterogeneity may came from widely varying socioeconomic conditions within particular subgroups, for example expectations regarding childcare or the acceptability of work outside the home. Such heterogeneity deserve further investigation in future work.

6 Conclusions

We conducted a randomized control trial in the data entry sector in Chennai, India that exogenously allocated workers to home-based or office-based work, while holding all other dimensions of the work constant. We first find a large positive and significant treatment effect of working from the office. The productivity of workers randomly assigned to working from the office is 18% higher than those working from home, independent of people's preferences for where they want to work. Two thirds of the effect manifests itself from the first day of work with the remainder due to quicker learning by office workers over the subsequent weeks. However, we find negative selection effects for office based workers. Those who prefer home-based work are 12% faster and more accurate at baseline. We also find negative selection on treatment: workers who prefer the home have larger negative productivity effects when allocated to home. The negative selection effects are stronger within subgroups that typically face bigger constraints in selecting to office work, such as workers with children and with other home care responsibilities as well as poorer households.

Our results show that understanding the self-selection of workers into different work locations is of first order importance when evaluating the merits of policies that aim to alter the allocation of workers to different work environments. If office work has positive productivity effects, constraining some parts of the population from allocating to their most productive work environment could hold back the success of these workers and further widen inequality between, for example, women and men or lower and higher class groups. This misallocation also leads to distortions in the productivity of the labor force overall. To the extent that these constraints are due to social or family pressures, policies that explicitly reduce or flatten such constraints, such as widening access to child care, could improve the allocation of workers to jobs. Of course, some of these choices might be the result of cultural or personal preferences. For example in some societies women themselves might feel that women should not work outside the home independent of her work productivity. Under these circumstances even policies that increase the treatment effect in the office or increased childcare access might not have a large effect on female labor force participation.

Our results are also important when evaluating industrial policies that are not directly aimed at changing constraints to working from home, but change the availability of different jobs in the economy. Take for example, India's Small Scale Reservation policy where certain products were reserved for manufacture by very small firms. If larger-scale factory environments substantially raise productivity, such policies detrimentally affect growth. However, if these small units allow some workers to participate in the economy who would otherwise not be able to take up a job, these policies can be viewed in a more favorable light.

References

Andreoni, James, Michael A Kuhn, and Charles Sprenger, "Measuring time preferences: A comparison of experimental methods," *Journal of Economic Behavior & Organization*, 2015, *116*, 451–464.

Barrero, Jose Maria, Nicholas Bloom, and Steven J Davis, "Why Working from Home Will Stick," Working Paper 28731, National Bureau of Economic Research April 2021.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta, "Cross-Country Differences in Productivity: The Role of Allocation and Selection," *American Economic Review*, February 2013, *103* (1), 305–34.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does management matter? Evidence from India," *The Quarterly journal of economics*, 2013, *128* (1), 1–51.

_ **, James Liang, John Roberts, and Zhichun Jenny Ying**, " Does Working from Home Work? Evidence from a Chinese Experiment *," *The Quarterly Journal of Economics*, 11 2014, *130* (1), 165–218.

_ **, Ruobing Han, and James Liang**, "How Hybrid Working From Home Works Out," Working Paper 30292, National Bureau of Economic Research July 2022.

Bruhn, Miriam, "License to Sell: The Effect of Business Registration Reform on Entrepreneurial Activity in Mexico," *The Review of Economics and Statistics*, 02 2011, *93* (1), 382–386.

Brynjolfsson, Erik, John J Horton, Adam Ozimek, Daniel Rock, Garima Sharma, and Hong-Yi TuYe, "COVID-19 and Remote Work: An Early Look at US Data," Working Paper 27344, National Bureau of Economic Research June 2020.

Clark, Gregory, "Factory Discipline," *The Journal of Economic History*, 1994, *54* (1), 128–163.

de Mel, Suresh, David McKenzie, and Christopher Woodruff, "The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka," *American Economic Journal: Applied Economics*, April 2013, *5* (2), 122–50.

de Soto, Hernando, *The other path: The invisible revolution in the third world*, New York Harper Row Publisher, 1990.

Forquesato, Pedro, "Social norms of work ethic and incentives in organizations," *Journal of Economic Behavior Organization*, 2016, *128*, 231–250.

Harrington, Emma and Natalia Emanuel, "'Working' Remotely? Selection, Treatment, and Market Provision of Remote Work (JMP)," *Working Paper*, 2022.

Hiller, Victor, "Work organization, preferences dynamics and the industrialization process," *European Economic Review*, 2011, *55* (7), 1007–1025.

— , "Self-control and the rise and fall of factory discipline," *Journal of Development Economics*, 2018, *133*, 187–200.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 2009, *124* (4), 1403–1448.

— **and** — , "The life cycle of plants in India and Mexico," *The Quarterly Journal of Economics*, 2014, *129* (3), 1035–1084.

— **, Erik Hurst, Charles I Jones, and Peter J Klenow**, "The allocation of talent and us economic growth," *Econometrica*, 2019, *87* (5), 1439–1474.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-control and the development of work arrangements," *American Economic Review*, 2010, *100* (2), 624–28.

— **,** — **, and** — , "Self-control at work," *Journal of Political Economy*, 2015, *123* (6), 1227–1277.

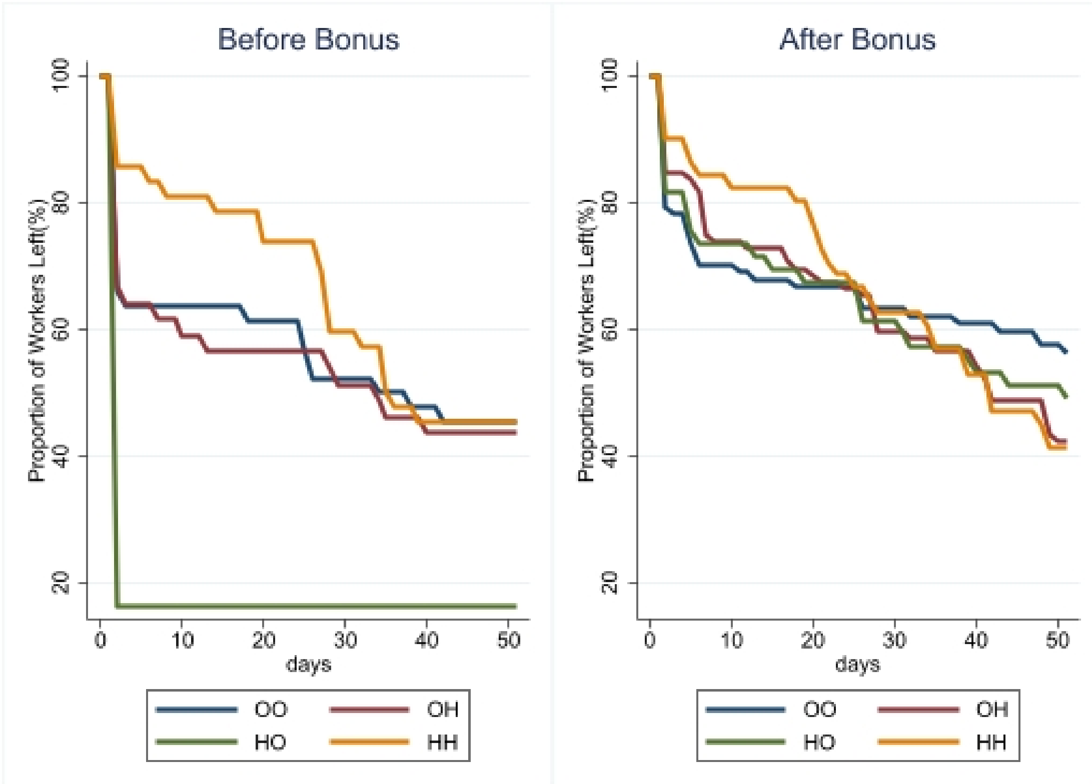
Levy, Santiago, *Good intentions, bad outcomes: Social policy, informality, and economic growth in Mexico*, Brookings Institution Press, 2010.

McCaig, Brian and Nina Pavcnik, "Export markets and labor allocation in a low-income country," *American Economic Review*, 2018, *108* (7), 1899–1941.

Porta, Rafael La and Andrei Shleifer, "Informality and development," *Journal of economic perspectives*, 2014, *28* (3), 109–26.

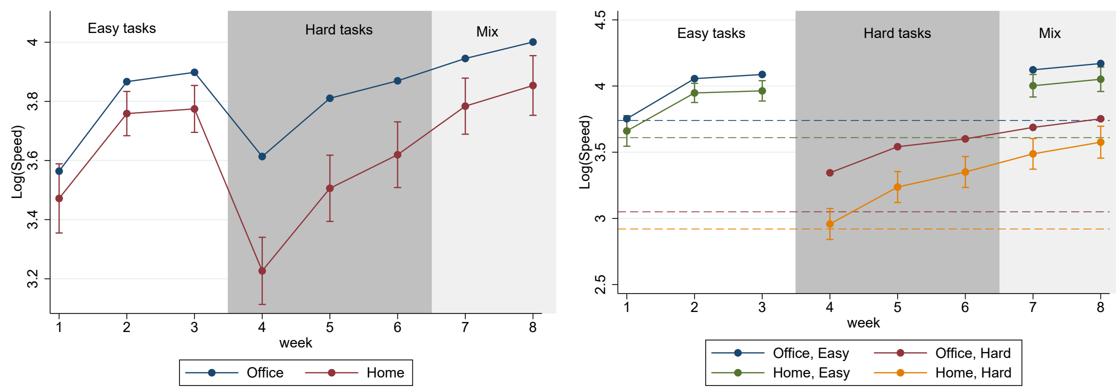
Schoar, Antoinette, "Supervisory Skills Training in the Cambodian Garment Industry: A Randomized Impact Evaluation," Technical Report, IFC Advisory Services in East Asia and the Pacific, World Bank Group. 2009.

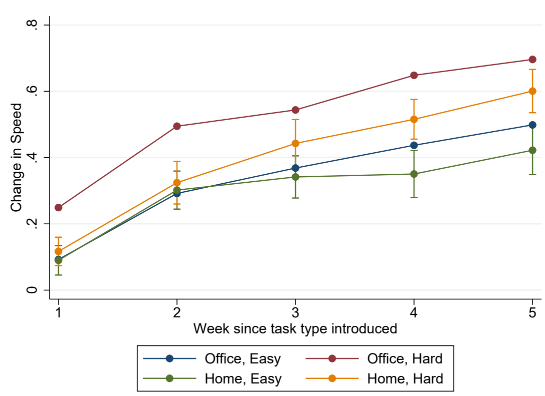
Figure 1: Attrition before and after retention bonus



Notes: This figure plots the proportion of workers left over the number of days of employment in each worker group. This is done separately for before and after the introduction of the retention bonus in the left and right panel, respectively. The 4 worker group are denoted by OO, HO, OH and HH. OO represents the worker who preferred office and were assigned office. HO represents the workers who preferred home but were assigned office. OH represents the workers who preferred office but were assigned home. Finally, HH represents workers who preferred home and were assigned home.

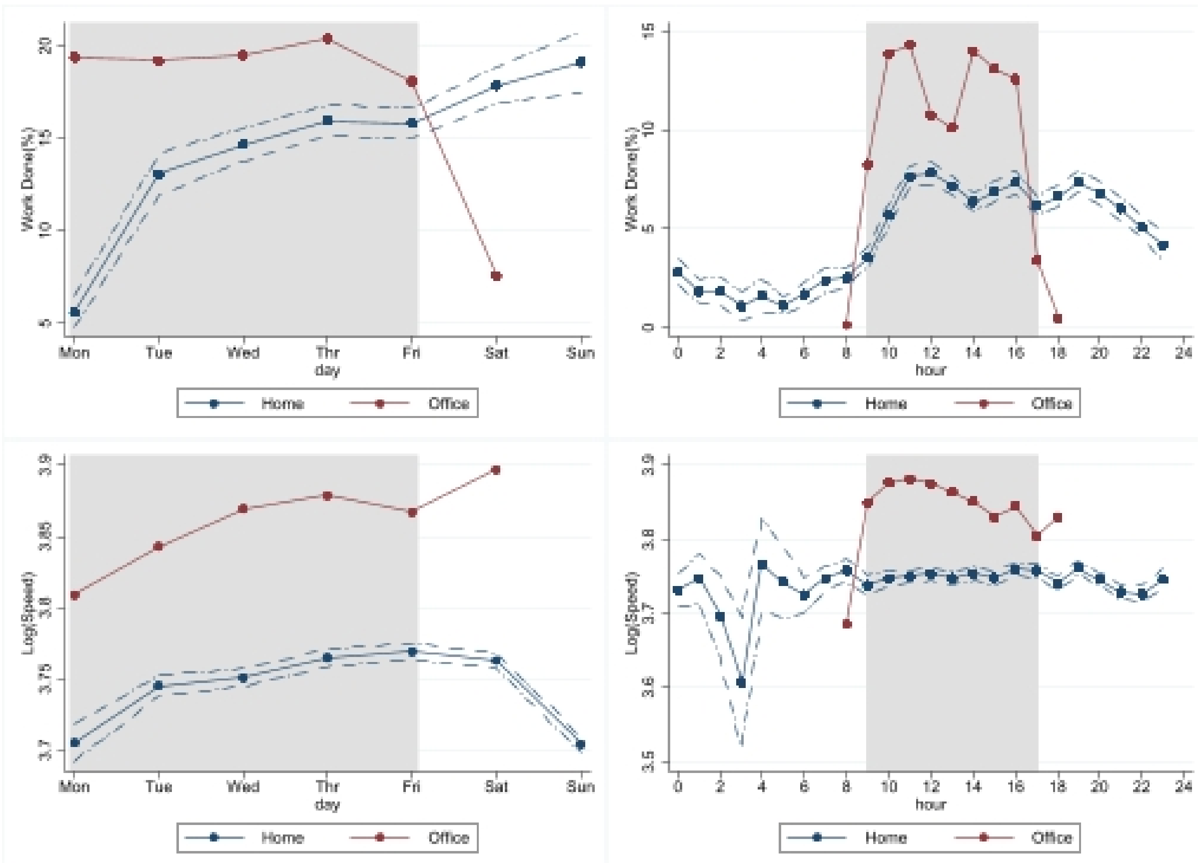
Figure 2: Speed and Learning over time





Notes: These figures plot speed and learning over 8 weeks for home- and office-based workers. The top-left panel plots the average speed for home- and office-based workers over 8 weeks. The top-right panel plots average speed separated by task difficulty for home- and office-based workers. The dashed line represent the average speed on first day for each difficulty-work location. The bottom-left panel plots average cumulative learning for home- and office-based workers for each difficulty. Cumulative learning is defined as the percentage increase in net speed compared to the baseline speed for that particular difficulty type. To make cumulative learning across both difficulties and work locations comparable, it is plotted over weeks since introduction of the particular difficulty.

Figure 3: Daily and Weekly Distribution of Work and Typing Speed



Notes: This figure plots distribution of work done and typing speed by work location over a typical day and a typical week. The top-left panel plots the distribution of work-done over a week across both the work locations. The top-right panel plots the distribution of work-done over a day across both the work locations. The bottom-left panel plots the average log(speed) for both the work locations over a typical week. Similarly, the bottom-right panel plots the average log(speed) for the both locations over a typical day.

Table 1: Worker Timeline

Recruitment	
Week -1	Applicants with relevant characteristics are invited to in person screening via newspaper ads
Day 0	On-site evaluation of applicants, recorded home versus office preference, initial typing speed test
3 Day Training	
Day 1-3	Training and orientation, do incentivized and non-incentivized typing speed tests
Day 3	Workers are allocated to office or home work
Work Assignment	
Week 1-3	Easy data-entry tasks
Week 4-6	Difficult data-entry tasks
Week 7, 8	Both difficulty data-entry tasks
Week 8	Job ends

Table 2: Compensation Structure

(1) Week	(2) Fixed component: Tasks Target	(3) Fixed component: Amount Paid INR ($)	(4) Performance-based variable component INR/task ($ / task)	(5) Retention Bonus INR ($)
1	18	2125 (32.2)	65 (1)	2000 (30.3)
2	20	2125	65	0
3	24	2125	65	0
4	24	2125	65	0
5-8	26	2125	65	0

Notes: This table explains the compensation structure of workers in both the work locations. Each row indicates the compensation structure for a particular week. The weeks are displayed in column (1). Columns (2) and (3), display the fixed component of the compensation structure. Upon completing a task target that is listed in the column (2), workers were payed a fixed amount listed in column (3). Column (4) lists the performance based pay which payed a piece rate per task completed beyond the weekly task target. Finally, column (5) displays the retention bonus that was paid. Figures in parenthesis are amounts in dollars at the exchange rate of INR 66 $\approx$ $ 1.

Table 3: Compensation Penalty for Errors

Penalty	Easy Task	Hard Task
	Error rate between (%)	
1X	0 - 7.5	0 - 15
1.5X	7.5-10	15-20
2X	10+	20+

Notes: This table explains the penalty schedule imposed for various levels of error rates.

Table 4: Attrition- Dependence of days worked on Ad type, location preference and location allocation

VARIABLES	(1) daysworked	(2) daysworked	(3) daysworked	(4) daysworked
ad_home	-3.29	-3.21		
	(6.45)	(6.37)		
pref_home	0.58		0.14	
	(0.38)		(0.67)	
alloc_home	0.71			0.72
	(4.62)			(4.64)
Constant	37.6***	38.2**	36.7***	36.3**
	(0.44)	(2.92)	(0.24)	(2.39)
Observations	280	280	280	280
R-squared	0.011	0.010	0.006	0.006
Wave FE	Yes	Yes	Yes	Yes

Notes: This table presents the result number of days worked regressed on type of ad worker responded to, their preference of work location and their assigned work location. The dependent variable is number of days worker which is same across all regressions. variable ad_home is a dummy variable taking value equal to one when the worker responded to a home-based work ad otherwise it takes value equal to zero. variable pref_home is a dummy variable taking value equal to one when the worker requested to work form home and is zero otherwise. Variable alloc_home takes value equal to one when the work is randomly assigned to work from home and is equal to zero is the worker is randomly assigned to work from office. Standard errors (in parentheses) are clustered at wave level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all specifications the unit of observations is a workers.

Table 5: Baseline Characteristics

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Assigned:			**Preferred:**		
	Home	Office	P-Value	Home	Office	P-Value
N	124	111		87	148	
Preferred home work (==1)	0.37	0.37	0.98			
Speed Tests						
Walk-in Speed	26.9	26.0	0.51	27.7	25.7	0.15
Cash Incentive Speed	33.1	33.4	0.78	35.9	31.7	0.00
No Incentive Speed	29.8	29.6	0.85	32.2	28.3	0.00
Demographic						
Female (==1)	0.58	0.43	0.02	0.49	0.52	0.70
Age (years)	24.7	25.3	0.38	26.1	24.3	0.00
Married (==1)	0.20	0.22	0.78	0.31	0.15	0.00
# of Kids	0.21	0.20	0.87	0.25	0.18	0.29
Has family care responsibility (==1)	0.04	0.10	0.07	0.11	0.04	0.03
Monthly family income (INR)	21,149	19,104	0.36	20,684	19,889	0.73
Commute Distance (km)	13.0	12.5	0.68	12.3	13.0	0.55
Education						
Education (years)	15.4	15.6	0.36	15.4	15.5	0.57
Used Computer before (==1)	0.98	0.91	0.03	0.95	0.94	0.63
Typing course- self reported (==1)	0.44	0.38	0.31	0.45	0.39	0.40
Typing course- showed certication (==1)	0.21	0.11	0.03	0.14	0.18	0.45
Work						
Work Exp (Years)	2.1	2.6	0.24	3.3	1.8	0.00
No. of Previous Office Jobs	1.1	1.2	0.40	1.4	0.9	0.00
Unemployment duration (months)	3.0	3.1	0.41	3.0	3.1	0.54
Miscellaneous						
Least concerned with Last Minute Effort (Ranking 1-6)	3.1	3.0	0.70	2.8	3.2	0.11
Estimated Time Discount Rate	0.98	0.95	0.19	0.99	0.95	0.05
Prefers Full-time Job (Yes)	0.94	0.98	0.13	0.93	0.98	0.06
At Home Commitments -Aspiration (Yes)	0.32	0.35	0.64	0.32	0.34	0.72

Notes: This table contains baseline comparison of workers randomly assigned to work at home and in office in columns (2)-(4) and baseline comparison of workers who requested to work in home and office in columns (5) - (7). columns (2) and (3) display the mean values of characteristics of workers who were assigned to work in home and office, respectively. Column (4) displays the P-value for the differences between means. Next, columns (5) and (6) display the mean values of characteristics of workers who requested to work from home and office, respectively, and column (7) shows the p-value for the differences between these means.

Table 6: Allocation Effect—Main Table

VARIABLES	(1) Net Speed	(2) Gross Speed	(3) Accuracy (in %)	(4) Time worked	(5) Time spent during Off hours	(6) Idle (in %)	(7) Net Speed Hard Tasks	(8) Net Speed High Penalty
alloc_home	-0.18***	-0.12***	-2.48**	-0.042	-0.51***	2.46***	-0.30***	-0.24***
	(0.050)	(0.034)	(1.14)	(0.23)	(0.018)	(0.84)	(0.066)	(0.078)
Constant	3.67***	3.80***	86.6***	33.7***	0.97***	14.6***	3.47***	3.45***
	(0.058)	(0.041)	(1.43)	(0.18)	(0.032)	(0.83)	(0.047)	(0.10)
Section+Week+Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	138,646	138,646	138,646	1,128	1,451	138,646	72,625	138,646
R-squared	0.194	0.205	0.266	0.014	0.805	0.072	0.108	0.233

Notes: This table contains estimates of effect of allocating workers to home based work environment for various performance measures. The regression specification is given by equation 1. In column (1), we have the primary measure which is the log of net speed. Net speed is defined as number of accurate characters typed per minute. In column (2) and (3) the dependant variable is log of gross speed and accuracy. Gross speed is defined as number of total characters typed per minute and accuracy is defined as ratio of accurate characters typed to total characters typed in percentage terms. The dependent variable in column (4) and (5) are the time spent entering data (in hours) and the proportion of work during the office hours (i.e., between 9 am to 6 pm from Monday to Friday) respectively. The software was programmed to measure intervals of time if no action was performed by employee using either the mouse or the keyboard while entering data. Ratio of the total time spent in such intervals to the total time spent entering data is defined as Idle time which is the dependent variable in column (6). In column (7) the dependant variable is the same as column (1) , log of net speed and in column (8) the dependant variable is the number of remunerated characters typed per minute. Remunerated characters typed is defined as total characters typed minus exponentially increasing penalty for incorrectly typed characters. alloc_home is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home and zero if assigned to work in office. All regressions account for variation arising from type of survey section being attempted, week of employment and hiring batch of workers using section, week and wave fixed effects respectively. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. Except column (4) and (5), for all other specifications the unit of observations is survey task attempted. In column (4) and (5) the unit of observation is week with column (4) further excluding week of observations when workers decided to drop out. Despite observations being at survey tasks or weeks, all regressions are re-weighted to give equal weight to each worker.

Table 7: Allocation- Robustness checks

VARIABLES	(1) Net Speed Main	(2) Net Speed	(3) Net Speed All Waves	(4) Net Speed Task-based weighting	(5) Net Speed
alloc_home	-0.18***	-0.18***	-0.18***	-0.16***	-0.19***
	(0.050)	(0.043)	(0.043)	(0.054)	(0.049)
ln_n_speed_cash		0.74***			
		(0.14)			
Characteristics Control					Yes
Section+Week+Wave FE	Yes	Yes	Yes	Yes	Yes
Constant	3.67***	1.01**	3.65***	3.88***	3.64***
	(0.058)	(0.50)	(0.082)	(0.040)	(0.095)
Observations	138,646	138,646	213,859	138,646	138,646
R-squared	0.194	0.266	0.189	0.197	0.196

Notes: This table contains robustness checks performed on effect of allocating workers to home based work environment for various performance measures. In all the columns, the dependent variable is the primary productivity measure of the log of net speed. Net speed is defined as number of accurate characters typed per minute. alloc_home is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home and zero if assigned to work in office. ln_n_speed_cash is the log of net speed during a speed test. This speed test was conducted during training of workers prior to work and workers were incentivized to type faster using cash reward. All regressions account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted. Except column (4), all regressions are re-weighted to give equally weight for all the workers. In column (3), observation for all hiring waves are included. In column (4), workers attempting greater number of surveys tasks get proportionately greater weight assigned to their observations. In column (5), we control for three characteristics across which the two groups are not balanced. These three characteristics are gender, family care responsibilities and prior computer usage.

Table 8: Cumulative Learning

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Surveys		Both Surveys			Easy			Hard	
VARIABLES	Net Speed	Baseline	Learning	Net Speed	Baseline	Learning	Net Speed	Baseline	Learning
alloc_home	-0.20***	-0.13**	-0.070***	-0.10**	-0.086*	-0.013	-0.33***	-0.19***	-0.14***
	(0.049)	(0.052)	(0.023)	(0.042)	(0.051)	(0.023)	(0.066)	(0.064)	(0.035)
Constant	3.58***	3.29***	0.28***	3.55***	3.27***	0.26***	3.54***	3.40***	0.13***
	(0.044)	(0.042)	(0.021)	(0.042)	(0.042)	(0.021)	(0.048)	(0.042)	(0.026)
Observations	131,923	131,923	131,923	62,251	62,251	62,251	69,672	69,672	69,672
R-squared	0.271	0.383	0.118	0.217	0.273	0.109	0.102	0.118	0.113
Section+Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table contains estimates of differential learning that takes placed based on work environment. In columns (1), (4) & (7), we have the primary measure which is the log of net speed. Net speed is defined as number of accurate characters typed per minute. In columns (2), (5) & (8) the dependant variable is Baseline Net Speed which is defined as average net speed for initial four surveys completed for particular difficulty level by each DEO. In columns (3), (6) & (9), the dependant variable is Cumulative learning which is defined as percentage increase in net speed compared to net speed of initial four surveys completed for particular difficulty level. Columns (1)-(3) consider both, easy and hard survey tasks. Whereas columns (4)-(6) and (7)-(9) consider only easy and only hard surveys, respectively. alloc_off is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from office and zero if assigned to work in home. All regressions account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted. Despite observations being at survey tasks or weeks, all regressions are re-weighted to give equally weight for all the workers.

Table 9: Selection on Initial Ability—Main Table

Sample	(1) Applicants 1 test	(2) Workers 1 test	(3) Workers 3 test	(4) Workers Work data	(5) Workers Work data
pref_home	0.15***	0.10**	0.12***	0.084*	0.083*
	(0.025)	(0.049)	(0.033)	(0.050)	(0.048)
alloc_home					-0.18***
					(0.049)
Constant	3.08***	3.13***	3.22***	3.55***	3.64***
	(0.023)	(0.037)	(0.032)	(0.058)	(0.063)
Speed Test FE			Yes		
Section+Week+Wave FE				Yes	Yes
Observations	884	234	704	138,646	138,646
R-squared	0.089	0.040	0.148	0.181	0.197

Notes: This table contains estimates of effect of workers selecting home based work environment. In all the columns, the dependent variable is the primary productivity measure of the log of net speed. Net speed is defined as number of accurate characters typed per minute. pref_home is a binary variable representing workers choice of work location taking value one if the choice is home based work and zero if the choice is office based work. alloc_home is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home and zero if assigned to work in office. Speed Test FE are fixed effect that account for variation that occurs in productivity due to 3 different types of typing speed test performed by workers prior to beginning the work. Column (1) uses data from speed test attempted by all applicant who applied for the data entry jobs. Column (2) filters the sample of applicants to include only workers who were selected. Column (3) adds observations from two additional tests performed by hired workers. The regression specification for columns (1) to (3) is given by equation 2. Regressions (4) and (5) consider data of survey tasks completed over next two months of employment and account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. The same regressions are re-weighted to give equally weight for all the workers.The regression specification for columns (4) and (5) is given by equation 3. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted.

Table 10: Selection on Initial Ability- controlling ad type

VARIABLES	(1) Pre-Filter 1 test	(2) Post-filter 1 test	(3) Post-filter 3 test	(4) Post-filter Work data	(5) Post-filter Work data
pref_home	0.12***	0.11**	0.12***	0.10**	0.10**
	(0.028)	(0.050)	(0.033)	(0.049)	(0.047)
alloc_home					-0.18***
					(0.049)
ad_home	0.076***	-0.056	-0.043	-0.16***	-0.15***
	(0.028)	(0.050)	(0.036)	(0.051)	(0.049)
Constant	3.06***	3.15***	3.23***	3.62***	3.71***
	(0.025)	(0.041)	(0.033)	(0.061)	(0.065)
Speed Test FE			Yes		
Section+Week+Wave FE				Yes	Yes
Observations	884	234	704	138,646	138,646
R-squared	0.097	0.045	0.152	0.191	0.206

This table contains estimates of effect of workers selecting home based work environment when accounted for type of advertisement workers responded to. In all the columns, the dependent variable is the primary productivity measure of the log of net speed. Net speed is defined as number of accurate characters typed per minute. pref_home is a binary variable representing workers choice of work location taking value one if the choice is home based work and zero if the choice is office based work. alloc_home is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home and zero if assigned to work in office. ad_home is a binary variable taking value one if the worker responded to employment advertising home based jobs and zero if responded to office based jobs. Speed Test FE are fixed effect that account for variation that occurs in productivity due to 3 different types of typing speed test performed by workers prior to beginning the work. Column (1) uses data from speed test attempted by all applicant who applied for the data entry jobs. Column (2) filters the sample of applicants to include only workers who were selected. Column (3) adds observations from two additional tests performed by hired workers. Regressions (4) and (5) consider data of survey tasks completed over next two months of employment and accounts for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. The same regressions are re-weighted to give equally weight for all the workers. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted.

Table 11: Selection on Initial Ability- Other outcome measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	Applicants	Workers	Applicants	Workers	Applicants	Workers	Applicants	Workers
VARIABLES	net speed	net speed	gross speed	gross speed	accuracy	accuracy	idle time	idle time
pref_home	0.15***	0.10**	0.14***	0.095*	0.71	0.49	-1.96***	-1.57*
	(0.025)	(0.049)	(0.028)	(0.052)	(0.87)	(1.67)	(0.42)	(0.82)
Constant	3.08***	3.13***	3.61***	3.62***	60.6***	62.3***	14.5***	13.9***
	(0.023)	(0.037)	(0.026)	(0.039)	(0.83)	(1.24)	(0.40)	(0.61)
Observations	884	234	884	234	884	234	884	234
R-squared	0.089	0.040	0.037	0.018	0.045	0.025	0.045	0.044

Notes: This table contains estimates of effect of workers selecting home based work environment. The regression specification for all columns is given by equation 2. In columns (1) & (2), the dependent variable is the primary productivity measure of the log of net speed. Net speed is defined as number of accurate characters typed per minute. In columns (3) and (4) the dependant variable is log of gross speed. Gross speed is defined as number of total characters typed per minute. In columns (5) and (6) the dependant variable is accuracy. Accuracy is defined as ratio of accurate characters typed to total characters typed in percentage terms. The software was programmed to measure intervals of time if no action was performed by employee using either the mouse or the keyboard while entering data. Ratio of the total time spent in such intervals to the total time spent entering data is defined as Idle time which is the dependent variable in columns (7) and (8). Columns (1), (3), (5) and (7) uses data from speed tests attempted by all applicant for the data entry jobs. Columns (2), (4), (6) and (8) filters the sample of applicants to include only workers who were selected. pref_home is a binary variable representing workers choice of work location taking value one if the choice is home based work and zero if the choice is office based work. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted.

Table 12: Selection on Treatment- Main Table

VARIABLES	(1) Log(net speed)	(2)	(3) Log(gross speed)	(4)	(5) Accuracy	(6)	(7) Idle time	(8)
pref_home	0.14**	0.067	0.062	0.028	3.83***	3.91***	-0.32	-0.34
	(0.064)	(0.058)	(0.049)	(0.032)	(1.37)	(1.37)	(1.04)	(1.04)
alloc_home	-0.14**	-0.13**	-0.086*	-0.055*	-1.67	-1.60	1.76*	1.74*
	(0.066)	(0.055)	(0.044)	(0.030)	(1.53)	(1.55)	(1.05)	(1.04)
c.pref_home#c.alloc_home	-0.12	-0.14*	-0.081	-0.13***	-2.17	-2.26	1.88	1.99
	(0.094)	(0.082)	(0.067)	(0.049)	(2.11)	(2.12)	(1.68)	(1.68)
Baseline_wrker_performance		0.75***		0.57***		-0.033		0.042
		(0.14)		(0.066)		(0.043)		(0.11)
Constant	3.62***	0.96*	3.78***	1.52***	85.1***	87.3***	14.7***	14.6***
	(0.067)	(0.49)	(0.047)	(0.26)	(1.60)	(3.14)	(0.90)	(1.05)
Section+Week+Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	138,646	138,646	138,646	138,646	138,646	138,646	138,646	138,646
R-squared	0.198	0.268	0.208	0.357	0.270	0.270	0.074	0.075

Notes: This table contains estimates of effect of workers selecting home based work environment in anticipation of its effect on their productivity. The regression specification for all columns is given by equation 4.In columns (1) & (2), the dependent variable is the primary productivity measure of the log of net speed. Net speed is defined as number of accurate characters typed per minute. In columns (3) and (4) the dependant variable is log of gross speed. Gross speed is defined as number of total characters typed per minute. In columns (5) and (6) the dependant variable is accuracy. Accuracy is defined as ratio of accurate characters typed to total characters typed in percentage terms. The software was programmed to measure intervals of time if no action was performed by employee using either the mouse or the keyboard while entering data. Ratio of the total time spent in such intervals to the total time spent entering data is defined as Idle time which is the dependent variable in columns (7) and (8). pref_home is a binary variable representing workers choice of work location taking value one if the choice is home based work and zero if the choice is office based work. alloc_home is a binary variable representing the treatment and takes value equal to one if the worker was randomly assigned to work from home and zero if assigned to work in office. c.pref_home#c.alloc_home is a binary variable which takes value one if worker chose to work at home and was assigned to work at home. It takes value of zero otherwise. ln_n_speed_cash, ln_cash_gross_speed, accuracy_cash and cash_idle are log of net speed, log of gross speed, accuracy and proportion of idle time during a speed test. This speed test was conducted during training of workers prior to work and workers were incentivized to type faster using cash reward. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted. All the regressions are re-weighted to give equal weight for all the workers.

Table 13: Selection on Initial Ability- Controlling for Characteristics

Panel A: Regressions with ATE

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Regression Specification	**Pref_Home**	**(SE)**			
	Baseline	0.116***	(0.033)			

Panel B: Controlling for individual hypothesis

	Hypothesis controlled for	**PCA**	**Pref_Home**	**(SE)**	**Characteristics**	**Pref_Home**	**(SE)**
(1)	Costs	1st principal comp	0.116***	(0.033)	distance	0.116***	(0.033)
(2)	Low Dscpln				last_min_effrt	0.105***	(0.03)
(3)					time_discount	0.110***	(0.03)
(4)		1st principal comp	0.101***	(0.032)	All characteristics	0.102***	(0.032)
(5)	Status				off_jobs_num	0.091***	(0.03)
(6)					fam_inc_scal	0.115***	(0.03)
(7)		1st principal comp	0.109***	(0.033)	All characteristics	0.089***	(0.033)
(8)	Outside Opt				fulltime_pref	0.107***	(0.03)
(9)					commit_prof_asp	0.118***	(0.03)
(10)		1st principal comp	0.116***	(0.032)	All characteristics	0.108***	(0.031)
(11)	Home Press				fam_care	0.110***	(0.03)
(12)					married	0.113***	(0.03)
(13)					child_yes	0.117***	(0.03)
(14)					age_scale	0.103***	(0.03)
(15)		1st principal comp	0.109***	(0.034)	All characteristics	0.084**	(0.034)
(16)	Female Const				female	0.115***	(0.03)
(17)					female_fam_care	0.112***	(0.03)
(18)					female_married	0.120***	(0.03)
(19)					female_child	0.117***	(0.03)
(20)					female_age_scale	0.117***	(0.03)
(21)		1st principal comp	0.118***	(0.033)	All characteristics	0.096***	(0.032)

Panel C: Controlling for all hypothesis

Regression Specification	PCA	**Pref_Home**	**(SE)**	Characteristics	**Pref_Home**	**(SE)**
Controlled for all Hypothesis	All 1st principal components	0.088***	(0.032)	All characteristics	0.051*	(0.031)

Notes: This table contains estimates of effect of workers selecting home work based on initial ability when controlled for various worker characteristics. The regression specification is given by equation 5. Only the relevant coefficient and the corresponding standard error is reported for each regression. The coefficient represents the selection effect based on initial ability and is presented for various regression in column (3) and (6), and the corresponding standard errors are presented in columns (4) and (7), respectively. Panel A presents the baseline effect when no characteristics are controlled for. In panel B, we present the selection effect where characteristics representing single hypothesis are controlled for. Columns (5) lists the characteristic that is controlled for with each section separated by dashed lines representing one hypothesis. Column (6) and (7), represent the corresponding coefficient and standard error of the estimated selection effect. The final line of each section denoted by "All characteristics" in column column (5), represents the selection effect when controlled for all characteristics listed in the particular hypothesis section. Columns (2)-(4) represents selection effect when controlled for the first principal component of set of all characteristics representing a particular hypothesis. Finally, Panel C represents results of selection effect when we control for all hypothesis. columns (2)-(4) uses all the 1st principal components as a control where as columns (5)-7) uses all the characteristics as controls. All regressions control for Speed Test fixed effect, which account for variation that occurs in productivity due to 3 different types of typing speed test performed by workers prior to beginning the work, and wave fixed effect, which account for hiring batch of workers. The row denoted by Baseline represents the selection effect from baseline specification represented earlier in Table 9 Column (3). Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted.

Table 14: Heterogeneity in Treatment

(1)	(2)	(3)	(4)
Panel A: Regressions with ATE			
		ATE	
Baseline		-0.18***	
		(0.05)	
Panel B: Regressions with CATE			
Hypothesis	Characteristic	CATE	Δ CATE
Costs	abv_avg_dist	-0.20***	0.05
		(0.07)	(0.10)
Low Discipline	last_min_effrt_yes	-0.18**	-0.01
		(0.07)	(0.10)
	high_discount	-0.12**	-0.11
		(0.06)	(0.10)
Status	low_fam_income	-0.28***	0.18*
		(0.08)	(0.10)
	no_prior_off_job	-0.16***	-0.05
		(0.06)	(0.10)
Outside Option	fulltime_pref	-0.08	-0.12
		(0.13)	(0.14)
	commit_prof_asp	-0.16***	-0.06
		0.06)	(0.11)
Home Pressure	fam_care	-0.19***	0.14
		(0.05)	(0.16)
	married	-0.17***	-0.09
		(0.06)	(0.12)
	child_yes	-0.17***	-0.06
		(0.05)	(0.14)
	abv_avg_age	-0.16***	-0.07
		(0.06)	(0.10)
Female Constraints	female	-0.16**	-0.07
		(0.07)	(0.10)
	female_fam_care	-0.19***	0.31*
		(0.05)	(0.17)
	female_married	-0.18***	-0.02
		(0.05)	(0.14)
	female_child	-0.18***	-0.07
		(0.05)	(0.15)
	female_old	-0.15***	-0.23*
		(0.05)	(0.13)

Notes: This table contains results from various regressions that we run in order to explore the heterogeneity in Treatment effect. Panel A and B represents partial results from one and twenty-two regressions, respectively. In Panel A we represent the average treatment effect of the baseline regression in column (3) which was earlier reported in Table 6 column (1). In Panel B we represent coefficients (and standard errors) of the variables alloc_home (a binary taking value 1 when worker was randomly assigned to work from home) and the variable alloc_home interacted with a dummy variable representing a relevant subgroup in columns (3) and (4), respectively. The regression specification is given by equation 6. Across all regressions the primary measure is the log of net speed. Net speed is defined as number of accurate characters typed per minute. The first two rows of panel B represent heterogeneity in treatment effect explored on dimension of cost of commute to office by dividing the entire sample into two groups, below- and above-average distance to travel to the office. Similarly, subsequent rows show results from regression exploring heterogeneity within subgroup created on basis of proxy characteristics other five hypothesis. Across both the panels, all regressions account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted. Despite observations being at survey tasks or weeks, all regressions are re-weighted to give equally weight for all the workers.

Table 15: Selection on Treatment- Controlling for Characteristics

(1)	(2) Regression Specification	(3) **Pref_Home*Alloc_Home**	(4) **(SE)**	(5) **Obs**	(6) $\mathbf{R^2}$
Baseline		-0.14*	(0.08)	138,646	28.4
Hypothesis controlled for	Costs	-0.14*	(0.08)	138,646	28.4
	Low Discipline	-0.15*	(0.08)	138,646	28.7
	Status	-0.15*	(0.09)	138,646	29.2
	Outside Option	-0.16*	(0.08)	138,646	28.7
	Home Pressures	-0.17*	(0.09)	138,646	28.9
	Female Constraints	-0.16*	(0.08)	138,646	29.6
Controlled for all hypothesis		-0.18**	(0.09)	138,646	31.7

Notes: This table contains estimates of effect of workers selecting home based work environment in anticipation of its effect on their productivity while accounting for various worker characteristics. Each row represents results from an individual regression where the dependent variable is the primary productivity measure of the log of net speed. The regression specification is given by equation 7. The dependent variable Net speed is defined as number of accurate characters typed per minute. The pertinent coefficient and standard error of pref_home*alloc_home alone are represented in columns (3) and (4), respectively. pref_home*alloc_home is a binary variable which takes value one if worker chose to work at home and was assigned to work at home. It takes value of zero otherwise. All regressions account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. The row denoted as Baseline represents the selection effect from baseline specification represented earlier in Table 12 Column (2). Next six rows represent the selection on treatment effect when individually controlled for group of proxy characteristics for each hypothesis considered to explain the negative selection effect. The final row represents the the selection effect when we control for groups of characteristic representing all hypothesis. The same regressions are re-weighted to give equally weight for all the workers. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all other specifications the unit of observations is survey task attempted.

Table 16: Heterogeneity in Selection on Treatment

(1)	(2)	(3)	(4)
Panel A: Baseline Regression			
		Pref_home*Alloc_home	
Baseline		-0.14*	
		(0.082)	
Panel B: Regressions with CATE			
Hypothesis	Characteristic	CATE	Δ CATE
Costs	abv_avg_dist	-0.02	-0.21
		(0.11)	(0.17)
Low Discipline	last_min_effrt_yes	-0.17	0.04
		(0.11)	(0.17)
	high_discount	-0.14	0.00
		(0.10)	(0.16)
Status	low_fam_income	0.09	-0.43**
		(0.12)	(0.17)
	no_prior_off_job	-0.18*	0.11
		(0.10)	(0.18)
Outside Option	fulltime_pref	-0.19	0.04
		(0.17)	(0.18)
	commit_prof_asp	-0.17*	0.07
		(0.10)	(0.19)
Home Pressure	fam_care	-0.11	-0.73***
		(0.09)	(0.15)
	married	-0.11	-0.12
		(0.09)	(0.20)
	child_yes	-0.07	-0.48**
		(0.09)	(0.21)
	abv_avg_age	-0.01	-0.37**
		(0.10)	(0.18)
Female Constraints	female	-0.22*	0.11
		(0.13)	(0.17)
	female_fam_care	-0.12	-0.80***
		(0.09)	(0.16)
	female_married	-0.14	0.00
		(0.09)	(0.22)
	female_child	-0.10	-0.31
		(0.09)	(0.22)
	female_old	-0.13	-0.02
		(0.10)	(0.22)

Notes: This table contains results from various regressions that we run in order to explore the heterogeneity in selection on treatment effect. Panel A and B represents partial results from one and twenty regressions, respectively. In Panel A we represent the selection on treatment effect of the baseline regression in column (3) which was earlier reported in Table 12 column (2). In Panel B we represent coefficients (and standard errors) of the variables pref_home*alloc_home (a binary taking value 1 when worker was randomly assigned to work from home and chose to work from home) and pref_home*alloc_home interacted with relevant subgroup in columns (3) and (4), respectively. The regression specification is given by equation 8. Across all regressions the primary measure is the log of net speed. Net speed is defined as number of accurate characters typed per minute. The first two rows of panel B represent the heterogeneity in selection on treatment effect explored on dimension of cost of commute to office by dividing the entire sample into two groups, below- and above-average distance to travel to the office. Similarly, subsequent rows show results from regressions exploring heterogeneity within subgroup created on basis of proxy characteristics other five hypothesis and the first principal component of proxy group. Across both the panels, all regressions account for variation arising from type of survey section being attempted, duration of employment and hiring batch of workers using section, week and wave fixed effects respectively. Standard errors (in parentheses) are clustered at individual level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all specifications the unit of observations is survey task attempted. Despite observations being at survey tasks or weeks, all regressions are re-weighted to give equally weight for all the workers.

A Appendix

A.1 Pictures of the Office and a few of Home work settings

[Figure A.1 about here.]

A.2 User interface of data entry tasks

[Figure A.2 about here.]

A.3 Ad types

[Figure A.3 about here.]

[Table A.1 about here.]

A.4 Waves

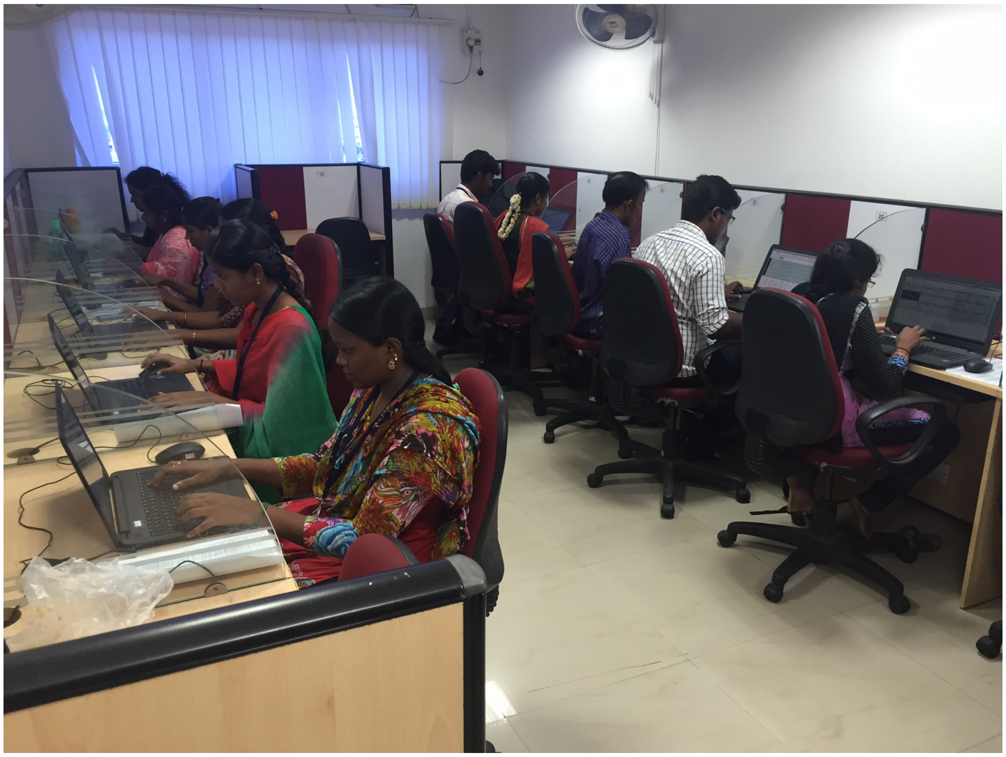
[Table A.2 about here.]

[Table A.3 about here.]

A.5 Examples of tasks by difficulty

[Figure A.4 about here.]

Figure A.1: Pictures of the Office and a few of Home work settings



(a) The Office

(b) Home work setups

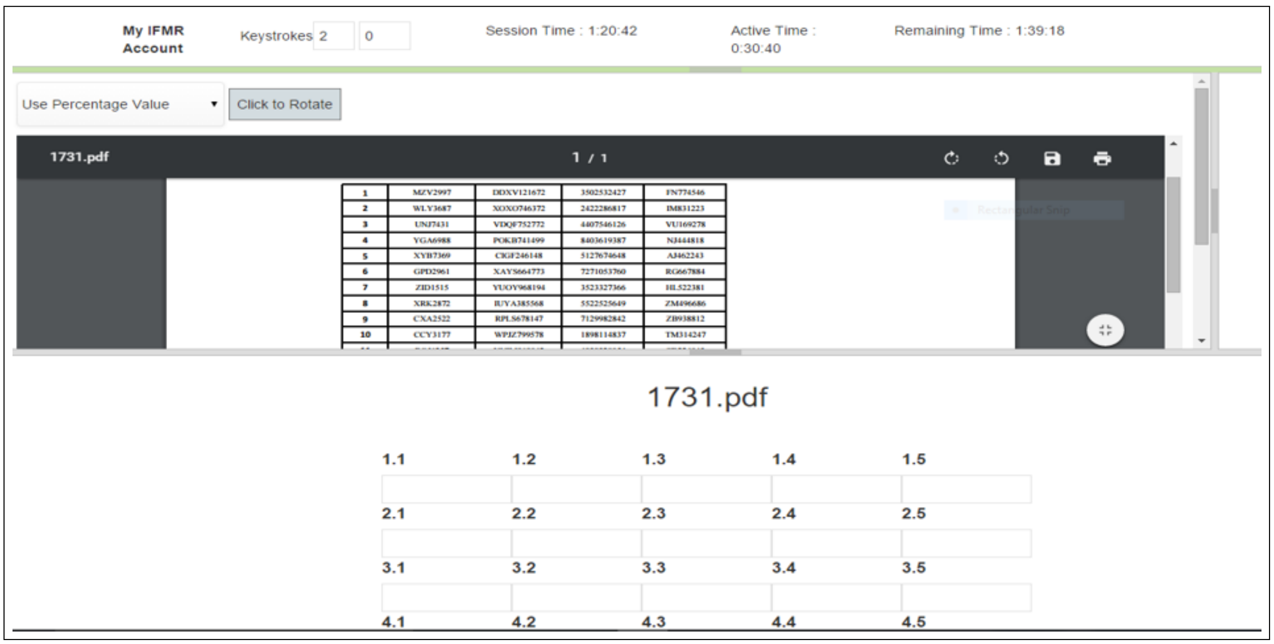


Figure A.2: User interface of a sample data entry tasks in the proprietary software

Figure A.3: Newspaper Ads sample

(a) Office-based work ad

(b) Home-based work ad

Figure A.4: Examples of data entry tasks by difficulty

(a) strings of random alpha-numeric characters vs alpha-numeric and special characters

Easy task:

S.no வ. எண்	Information 1 தகவல் 1	Information 2 தகவல் 2	Information 3 தகவல் 3	Information 4 தகவல் 4	Information 5 தகவல் 5
1	EQY6267	UGKI733669	8981753224	WM578562	OZ441532
2	DHQ1499	TSUA974617	4773422856	QD647325	NV663391
3	YAN8395	YJVW199368	6553632731	CW344523	UC189451
4	SQN6386	ZNCQ587129	3070840773	KW478175	XG635848
5	HQT4833	LYHS997811	2157713174	CN687268	LY694874

Hard task:

<table>
<tr><th>S.no வ. எண்</th><th>Information 1 தகவல் 1</th><th>Information 2 தகவல் 2</th><th>Information 3 தகவல் 3</th><th>Information 4 தகவல் 4</th><th>Information 4 தகவல் 4</th></tr>
<tr><td>1</td><td>?z j?G~I_</td><td>oaFeDc-,lg</td><td>4:bcw AsBoe</td><td>-r*/o~n8</td><td>)ekO</td></tr>
<tr><td>2</td><td>X tw +DfoB</td><td>b!`x #*Rk,s</td><td>VLW eNoCArM</td><td>|s.,J@=u</td><td>8X Z -o</td></tr>
<tr><td>3</td><td>5,~;honQ</td><td>k4TF|"?y#[</td><td>%~4w q?5@:t</td><td>1.$}d"^M</td><td>x 2tY9</td></tr>
<tr><td>4</td><td>]nX ^x %1m</td><td>\#*vz Z .#no</td><td>YbBm+44P35</td><td>il{")=cY</td><td>4G*l</td></tr>
<tr><td>5</td><td>tNP$k#|C</td><td>2-/)6aCn)o</td><td>mw 9gOe[lC`</td><td>h$K~DUQr</td><td>ps;M:</td></tr>
</table>

(b) Type-set vs Handwritten text

Easy task:

their mother tongue and unsure in the official language. To remedy the situation we need a radically new approach to the teaching of languages. It is essential that children are taught only in their mother tongue and simultaneously learn Hindi up to grade six. This will give them the necessary grounding in their own milieu, their own folklore, mythology and literature, and help them develop a love and respect

Hard task:

mashindano. ambayo wewe kushinda au kupoteza. Kama unafikiri maishani mchezo, kwa hiyo ni pia ni muhimu kuuliza ni aina gani ya mchezo. Baadhi ya michezo ni alicheza kwa ajili ya kujifurahisha peke yake. Baadhi ya michezo ni distinctively kijamii (daraja): baadhi ni makusudi

Table A.1: Main result separated by home- and office-based work ads

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample		All			Home Ads			Office Ads	
Effect	TE	SAB	SOT	TE	SAB	SOT	TE	SAB	SOT
Dependent Variable		Log(Net Speed)			Log(Net Speed)			Log(Net Speed)	
pref_home		0.12***	0.067		0.16***	0.088		0.089**	0.069
		(0.033)	(0.058)		(0.059)	(0.078)		(0.038)	(0.078)
alloc_home	-0.18***		-0.13**	-0.19**		-0.17**	-0.17**		-0.11
	(0.050)		(0.055)	(0.072)		(0.078)	(0.065)		(0.071)
c.pref_home#c.alloc_home			-0.14*			-0.17			-0.099
			(0.082)			(0.11)			(0.11)
Constant	3.67***	3.22***	0.96*	3.55***	3.12***	0.85**	3.78***	3.28***	1.11
	(0.058)	(0.032)	(0.49)	(0.088)	(0.051)	(0.33)	(0.074)	(0.041)	(0.83)
Observations	138,646	704	138,646	47,253	269	47,253	91,393	435	91,393
R-squared	0.194	0.148	0.268	0.204	0.165	0.290	0.195	0.163	0.260
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the main results of the paper for subsamples split by ad types. Columns (1)-(3) present the main results for the workers where as columns (4)-(6) and (7)-(9) present the same results for home based and office based work ads, respectively. Columns (1), (4) and (7) present the treatment effect regressions for the three samples. Columns (2), (5) and (8) present the regression estimating the selection at baseline effect. Columns (3), (6) and (9) present the selection on treatment effect regressions. All regression are based on 8 weeks work data except the ones in columns (2), (5) and (8), which are based on the 3 speed test conducted for each worker. Variable pref_home is a dummy variable taking value equal to one when the worker requested to work form home and is zero otherwise. Variable alloc_home takes value equal to one when the work is randomly assigned to work from home and is equal to zero is the worker is randomly assigned to work from office. Standard errors (in parentheses) are clustered at wave level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all specifications the unit of observations is survey task attempted. All the regressions are re-weighted to give equal weight for all the workers.

Table A.2: Changes made across the waves

Detail	Wave 1	Wave 2	Wave 3 pre	Wave 3 post	Wave 4
Work Duration	3 months		2 months		
Salary- Fixed component	None	• Month 1 (M1): INR 8500 ($ 128.8) • M2: INR 4400 ($ 66.7) • M3: None	• M1 : INR 8500 • M2 : INR 8500		
Variable component	Paid 6 paisa per 4 correct characters ($1 per 4000 correct characters)	INR 65 ($ 1)/ DE task completed over given takert of surveys in each week			
Completion bonus	None			INR 2000 on completion of 1st week of work	• High Incentive: INR 2000 on completion of 1st week of work • Low Incentive: INR 2000 on completion of 8th (last) week of work
Targets to retain the job	• Time: 40 hours			• Time: 35 hours (weeks 1-4) • Tasks: 26 surveys (weeks 5-8)	
Selection Criterion	• Age: 18 to 40 years • Education: 9th grade to graduates • Speed: 10-30 words per minute • Residence: Within Chennai • DE work ex: 0-6 months	• Age: 18 to 40 years • Willingness to work in both locations (surveyor assessment) • Time commitment: Full time		• Age: 18 to 40 years • Willingness to work in both locations (Manager assessment)	

Cells left blank imply that no changes to previous setting were made. We use the average exchange rate between Indian Rupee and United States Dollar during the period of experiment which is INR 66 $ 1.

Table A.3: Treatment and Worker Sorting Effects for All Waves

	(1)	(2)	(3)	(4)	(5)	(6)
Wave		3.5 & 4			All Waves	
Effect	TE	SAB	SOT	TE	SAB	SOT
Dependent Variable		Log(Net Speed)			Log(Net Speed)	
pref_home		0.12***	0.067		0.038	0.055
		(0.033)	(0.058)		(0.034)	(0.050)
alloc_home	-0.18***		-0.13**	-0.18***		-0.12***
	(0.050)		(0.055)	(0.043)		(0.045)
c.pref_home#c.alloc_home			-0.14*			-0.11
			(0.082)			(0.072)
Constant	3.67***	3.22***	0.96*	3.65***	3.34***	0.58
	(0.058)	(0.032)	(0.49)	(0.083)	(0.064)	(0.40)
Observations	138,646	704	138,646	212,823	986	212,823
R-squared	0.194	0.148	0.268	0.190	0.225	0.305
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes
Section & Week FE	Yes		Yes	Yes		Yes

Notes: This table presents the main results of the paper replicated for the All waves. Columns (1)-(3) present the main results for the post-retention bonus sample where as columns (4)-(6) present the same results for all the waves. Columns (1) and (4) present the treatment effect regressions for the two samples. Columns (2) and (5) present the regression estimating selection at baseline effect. Columns (3) and (6) present the selection on treatment effect regressions. All regression are based on 8 weeks work data except ones in columns (2) and (5), which are based on the 3 speed test conducted for each worker. Variable pref_home is a dummy variable taking value equal to one when the worker requested to work form home and is zero otherwise. Variable alloc_home takes value equal to one when the work is randomly assigned to work from home and is equal to zero is the worker is randomly assigned to work from office. Standard errors (in parentheses) are clustered at wave level. *, **, and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. For all specifications the unit of observations is survey task attempted. All the regressions are re-weighted to give equal weight for all the workers.