Managing Remote Team Coordination: Experimental Evidence on Constrained and Flex Scheduling*

Early Stage Draft, Please Do Not Cite Without Permission

Tanjim Hossain† Elizabeth Lyons‡

February, 2023

Abstract

A frequently cited benefit workers associate with remote work is flexible work schedules that allow them to balance work and non-work time demands. While flexible schedules may allow employers to attract and retain workers who value them, they may also make coordination among work teams more challenging by introducing scheduling uncertainty and communication lags. We study these trade-offs by comparing the coordination and performance impacts of flex work schedules that do not specify when workers in teams should collaborate on their assignments relative to constrained schedules that do specify time windows for collaboration on team performance among remote workers. Using a field experiment among teams of online language translators, we find that flexible schedules lead to somewhat lower team performance than constrained schedules. We find heterogeneity in this treatment effect that is consistent with flexible schedules being challenging for teams more likely to have other communication barriers. In particular, we find mixed gender teams perform significantly worse under flex schedules, but same gender teams do not. We are currently investigating whether these effects are temporary or persist over time.

JEL Codes: J24, M12, M54

*We thank seminar participants at the University of Soutern California, Harvard, and conference participants at the 2023 ASSA annual meeting for helpful comments. All errors are our own.

†University of Toronto email: Tanjim.Hossain@Rotman.Utoronto.Ca

‡School of Global Policy & Strategy, UC San Diego; email: lizlyons@ucsd.edu.

1 Introduction

Collaborative production is considered essential for the majority of jobs in the US (Freeman et al., 2022; Weidmann and Deming, 2021). Despite it's ubiquitous use, effective team production is challenging. A growing body of literature has highlighted some of the factors that can limit successful collaboration, including barriers to efficient team formation (Hossain et al., 2019), and communication barriers within teams (Hjort, 2014; Lyons, 2017; Weidmann and Deming, 2021). The growth in remote and hybrid work arrangements (Bloom et al., 2022; Dhawan and Chamorro-Premuzic, 2018) has introduced additional potential barriers to successful teamwork by reducing the frequency of face-to-face interactions which may increase the costs of information exchange (e.g. Catalini, 2018).[1]

An argument in favor of expanding the use of remote and hybrid work is that many employees prefer it and, thus, offering it can reduce recruitment and retention costs (Barrero et al., 2022). One of the dimensions of these work arrangements that particularly appeals to workers is the more flexible work schedules typically associated with them. For instance, the Gartner 2021 Digitial Work Survey found that flexible work hours was the most frequently cited reason workers think remote work improves their productivity,[2] and prior evidence has demonstrated a preference for flexible hours independently of the location of work (Felfe, 2012; McNall et al., 2009). However, while remote and hybrid work facilitates flexibility in work planning more easily than on-site work[3], flexible work hours are not a necessary component of flexible work location arrangements. Employers could, for instance, allow for remote or hybrid work to reap benefits from lower office space costs and geographically unconstrained pools of job applicants, but restrict the hours during which employees are required to work.

Given that flexible work hours could exacerbate coordination costs associated with remote work by increasing the uncertainty associated with team communication, understanding whether the potential benefits it may provide individual workers through higher job satifsaction outweighs the barriers to coordination it could generate is critical for understanding how best to manage remote and hybrid workers. Moreover, the impact of response time lags and uncertainty on team performance has implications for our understanding of team work more generally.

In this paper, we aim to isolate the role of flexible hours in remote team work in order to investigate

[1] To our knowledge, there is not evidence that remote and hybrid work arrangements have reduced the prevalence teamwork. For instance, a 2021 Gartner survey found that 80% of surveyed workers in the US, Europe, and Asia-Pacific use software that supports collaboration (see https://www.gartner.com/en/newsroom/press-releases/2021-08-23-gartner-survey-reveals-44-percent-rise-in-workers-use-of-collaboration-tools-since-2019).

[2] https://www.gartner.com/smarterwithgartner/digital-workers-say-flexibility-is-key-to-their-productivity.

[3] Once workers arrive at a work-specific site, the likelihood that they pursue non-work activities while there is lower than if they were to work from wherever they would otherwise pursue non-work activities.

1) whether the extent of coordination differs between teams with more or less work schedule flexibility, 2) whether teams perform differently with more of less work schedule flexibility, and 3) whether the answers to the first two questions differs by the extent of ex-ante communication frictions within a team.

To test our research questions, we ran a field experiment on a large online labor market (Burbano, 2021; Lyons, 2017) that allowed us to assign workers into teams of two and randomly assigned them to complete the task within a specific or constrained time window or allowed them to work on the task at any time within the work period in a flexible way. Workers were hired to translate an academic blog post from English into one of two languages; Swahili and Bengali. Teammates had not previously worked together when we hired them, and were each assigned to translate one half of the blog post. To generate incentives for team collaboration, we asked teammates to ensure they work together so that the final document is cohesive and we offered a 25% bonus for submitting a single cohesive document, rather than two separate halves. Each team was given a link to a private Slack channel to host their communication about the task.[4] Teams had a total of 24 hours from the time they were hired until the job deadline.

The job we assigned workers requires effective coordination for cohesive output. First, the way words and expressions are translated across the two documents halves should be consistent for cohesion and, second, the transition from the first half to the second half of the document should flow seamlessly. The individual halves do not depend on collaboration for successful completion. Collaboration could enhance individual performance if teammates provide each other useful input on their respective halves. It could also harm individual performance if time consuming coordination reduces time spent on translation. Translation involves problem solving and there is likely variation across workers in which problems are relatively easy or hard, and in how long the task takes to complete (Nitzke, 2019). Thus, there is ex-ante uncertainty over the necessary extent and the optimal timing of useful communication for cohesive document (Dessein and Santos, 2006). Our task is, therefore, similar to team projects with moderate interdependence in which some division of work across teammates is desirable and the need to communicate depends on the realization of information that arises during project execution.[5]

To generate flexible and constrained work schedules, one half of teams were not told what time period they should be on Slack to work with their teammates and the other half were told which two-hour window they should join Slack to work together. We refer to these two groups as *Flex* and *Constrained*, respectively.

[4] As we explain in further detail in section 2, the reason we assigned each teammate one part of the task rather than allowing teammates to decide on the appropriate division of work is because we wanted to reduce opportunities for free riding and, instead, focus on mutually beneficial collaboration.

[5] When continuous communication is needed for task completion, uncertainty about scheduling may be resolvable upfront as work cannot begin until teammates are collaborating. Similarly, when the task can be fully divided up front and no collaboration is subsequently needed, uncertainty about scheduling is unlikely to be a barrier to successful task completion.

Before assigning work time windows in the *Constrained* group, we asked workers in the group which time windows they were available to work and chose the first window of overlappig availability between teammates. Moreover, the job posts that advertised the constrained schedule job opportunities specified that workers would be asked to work within a two hour window to ensure this would not be a surprise to them, and to capture the potential for differential selection into *Flex* and *Constrained* jobs. Variation in the translation languages teams were hired for allows us to test whether the impacts of flexible scheduling differs across work settings. Otherwise, the job advertisement, instructions, and tasks teams received were identical across *Flex* and *Constrained* teams.

Our analysis of the results from this experiment demonstrates three main results. First, we find some evidence of negative selection into *Flex* jobs—hired teams made up of workers who applied to the *Flex* jobs have significantly less education, and sizeably but insignificantly less UpWork experience and lower UpWork ratings. Moreover, we find that this differential selection is primarily driven by applicants to the Swahili translation jobs and not the Bengali translation jobs. However, controlling for team characteristics does not change our estimated effects of flex scheduling on team coordination or performance, suggesting that selection into schedule types is not the primary driver of our findings. Second, we find weak evidence that teams in *Flex* are less likely to coordinate with each other, and weak evidence that they perform worse on the task, due to a combination of lower translation quality and lower document cohesion. Third, we find significant heterogeneity in the impacts of flex scheduling on team performance. In particular, we find that Swahili translator teams coordinate significantly less and perform significantly worse in *Flex* than in *Constrained*, but that Bengali translator teams perform equally well in both. Similarly, we find mixed gender teams coordinate significantly less and perform significantly worse in *Flex* than in *Constrained*, but that same gender translator teams perform equally well in both. We also find evidence that flexible schedules lead to worse performance, but no less coordination, among teams made up of two females.

Our findings contribute to our growing understanding of impediments to successful teamwork. Prior studies have demonstrated, for instance, that national diversity (Lazear, 1999; Lyons, 2017), ethnic diversity (Hjort, 2014), gender diversity (Berge et al., 2016; Kelemen et al., 2020), geographic dispersion (Hinds and Mortensen, 2005), and mode of communication (Maznevski and Chudoba, 2000) can be barriers to effective team coordination. Our study builds on these studies to examine how a relatively low-cost managerial decision, whether or not to constrain the timing during which teamwork should occur, contributes to effective coordination and mediates the effects of diversity on coordination. Thus, our study also contributes to literature on the organization and management of teams that goes beyond managing barriers to coordination

and, for instance, considers interventions for reducing moral hazard (Holmstrom, 1982; Rayo, 2007), and improving the allocation of tasks (Burgess et al., 2010; Dessein and Santos, 2006; Van de Ven et al., 1976).

We also contribute to the literature on alternative or non-traditional work arrangements by analyzing a frequently provided and commonly stated benefit of remote work arrangements. Prior work has demonstrated that offering occasional flexible work hours can improve worker and team performance (Angelici and Profeta, 2020), however, the studied workers were primarily non-remote. Whether these benefits extend to fully remote teams is unclear given that they have fewer opportunities to overcome delayed or uncertain response times during the course of the work week.[6] Perhaps most related to our study is Yang et al. (2022) who analyze the communication patterns of Microsoft employees following the COVID-19 induced shift to remote work. The study finds that remote work is associated with less synchronous communication, more asynchronous communication, and narrower communication networks. These findings are consistent with remote work increasing the uncertainty of the value of communication efforts. We examine the extent to which constrained schedules may be able to reduce this uncertainty. Moreover, by allowing for differential selection into flexible and constrained jobs, we are able to study the extent to which any increase in coordination barriers due to flexibility are offset by improved worker satisfaction, an often cited benefit of alternative work arrangements (e.g. Bloom et al., 2022). To examine whether the negative effects of flexible schedules are mitigated as team members get to know each other, we are currently running a second round of our RCT in which our previously hired teams are block re-randomized into *Flex* or *Constrained* schedules.[7]

2 Experimental Design

There are several empirical challenges associated with linking flexible schedules to performance outcomes among remote teams of workers using administrative data. For example, more effective managers may be better at matching schedule types to tasks and also be better at hiring and motivating workers. Thus, we may observe that schedule types more likely to be adopted by better managers (either flexible or constrained) are also associated with better team outcomes but for reasons other than the schedules themselves. Similarly, and relatedly, schedule types may systematically differ across task types making it challenging to understand the extent to which a task type is driving performance relationships between scheduling and outcomes. In addition, data that includes both measures of team coordination and performance among teams working on similar tasks but under different scheduling arrangements is challenging to access.

[6] Importantly, many remote workers are onboarded fully remotely and, thus, have minimal face time with their teammates (Groysberg, 2020).

[7] The results from this round of the study will be available by May, 2023.

To address these challenges and collect data on characteristics, coordination, and performance among teams of workers performing identical tasks but under different scheduling arrangements, we ran a field study on a large, global online labor platform.

2.1 Experimental Setting

The platform we are running our study on allows workers to post jobs to attract remote applicants from around the world. In general, job postings include a summary of the task, whether the project will pay hourly wages (input-based pay) or a fixed price for output (output-based pay), the amount the project will pay (hourly wage or fixed price), and the approximate time commitment associated with the task.

To apply for jobs, applicants submit proposed bids (either an hourly wage or fixed price for output) and cover letters explaining why they are suited for the job. They also advertise their capabilities through their profile pages which are observable to potential employers. This profile pages include information on prior work experience on the platform and any related prior employer feedback they have received, educational attainment, a summary of skills, applicant locations, and a photo. In addition to financial incentives on any given job, workers on the platform are incentivized to perform well on a task in order to receive positive employer ratings and feedback for their profile pages. This feedback can significantly increase future work prospects (Pallais, 2014).

The platform supports team work and monitoring through a digital work room that takes screenshots of worker progress and reports mouse clicks and key strokes, and allows teammates to collaborate. To reduce the risks that input-based pay workers are wasting time, hourly wage workers can be required to work in the digital work room to complete their tasks. Fixed price worker pay cannot be conditional on work performed in the digital work room.

The task that we hired workers to complete is a translation task that requires workers to translate a four-page blog post summarizing an academic study from English into either Swahili or Bengali.[8] We have selected a translation task for several reasons. First, there is a clear measure of performance on the task; specifically, how readable the translated text is (Arnold et al., 1995; Weng et al., 2019). Second, though measuring performance is relatively straightforward, it requires creativity and problem solving (e.g. Al-Awawdeh, 2021) and, thus, there is ex-ante uncertainty about how workers should approach the task and when and why they might benefit from advice from teammates. Importantly, for reasons that continue to be relevant today, Google translate and other freely available machine translation software are not yet able

[8]The English text we asked hired workers to translate is provided in Appendix Appendix B.

to effectively translate English into Bengali and Swahili (De Pauw et al., 2011; Hasan et al., 2019).[9]

In addition, focusing our task on specific languages ensures we will hire teammates who speak similar languages and are, thus, able to communicate. Moreover, it reduces the likelihood that applicants from very different time zones apply to this job (and, in practice, time zone variation within the teams we hired is close to zero). While time zone variation is a potentially interesting source of remote team coordination barriers, it is not the source we are interested in studying in this paper. Also of importance, our task is relatively affordable on the virtual labor market we are using.

2.2 Hiring and Work Process

To hire teams of workers for our task, we posted two jobs at a time — one that advertised the Swahili tranaltion job and one that advertised the Bengali translation job. The job posts specified that we would hire applicants to work in teams of two to complete a translation task, that we would pay up to $20 per worker for a fixed price contract, and that the deadline for submission would be 24 hours after hiring. The full job posting information is provided in Appendix Appendix B. We hired all applicants who bid at most $20 and displayed the ability to communicate in english and in the language the document was to be translated in.

Once hired, workers were told who their teammate was, which section of the blog post they were responsible for, and their job deadline. In addition, they were provided with a link to their team Slack channel, and asked to work with their teammate to ensure the translated document was cohesive. To provide additional incentives beyond the potential positive platform feedback for teammates to work together, they were told they would receive a $5 bonus for submitting a single cohesive translated document by the end of the 24 hour work period. The full instructions workers received are eprovided in Appendix Appendix B.

Upon completing the task, workers were asked to complete a survey on their experience working on the task. The list of survey questions is provided in Appendix Appendix B.

2.3 Experimental Variation

We generate variation in scheduling type by randomizing with a coin flip which of the two job postings we advertised at the beginning of our study offered a flexible scheduling arrangement (*Flex*) and which offered a constrained scheduling arrangement (*Constrained*). After hiring thirty teams for each of the job postings,

[9]To further ensure that machine translation would be noticeable and ineffective, we included English words with multiple possible translations into Bengali or Swahili depending on the context, and we ensured our translation evaluators had copies of the machine translated versions of the blog post to compare submissions against.

we switched and assigned the flexible scheduling arrangement to language it was not initially assigned to, and the constrained arrangement to the other language.

The motivation for this approach to varying schedule types across languages is that we were concerned that posting two jobs per language at one time that differed in their scheduling type would confusion among individual potential applicants and alert them to the possibility that we were focused on scheduling variation. Given that Bengali translators are unlikely to be alerted about or look into Swahili translation jobs and vice versa, we were not concerned about confusion over different schedules across these two jobs. An important concern about this approach is that the pool of workers may differ across languages and over time on the platform. To address this, we added a small pool of teams hired for *Flex* Swahili task at the end of our second round of hiring and we do not find differences in the characteristics of teams hired for *Flex* Swahili jobs across hiring rounds. In addition, we include hiring round fixed effects in all our analyses and also confirm that including these fixed effects does not change our findings.

To implement constrained schedules, we asked workers hired for *Constrained* jobs to work with their teammates within a specific two hour window. This two hour window was selected based on pre-determined overlapping availability between teammates. The specific text we used to determine this overlap and to request that teammates work together given the overlapping time window are provided in the job instructions copied in Appendix Appendix B. Applicants to constrained jobs were made aware that they would be asked about their availability and to work within a 2-hour window based on information included in the job posting (see job posting information in in Appendix Appendix B). Workers in *Flex* were asked to work with their teammate, but were not given a specific time window to do so. Otherwise, *Flex* and *Constrained* job postings, tasks, and instructions were identical.

3 Data and Analysis Plan

3.1 Outcome Variables

Our primary outcomes of interest are 1) whether or not teammates communicated or attempted to communicate with each other, and 2) how effectively teams performed.

To measure team communication, we capture whether both workers in a team join their team Slack channel, and whether any communication within the Slack channel occurs. We use the former measure to capture any chatting that occurs through direct messaging, which we are unable to observe directly.[10] The

[10]It is possible for teammates to send direct messages to each other through the hiring platorm. We are unable to observe these instances of communication.

latter is a more restrictive measure of coordination. As Table 1 demonstrates, both workers in slightly more than one third of teams joined Slack, and about 30% were observed communicating. We supplement these objective measures of communication with responses to a post-task survey question that asked team members if they communicated with each other. This measure likely suffers from response bias because workers were aware that we wanted them to work with their teammates. Consistent with this, in 70% of teams for which at least one worker completed the survey, at least one teammate reports that their team communicated. We, nonetheless, verify whether using it as an outcome changes are findings. We are in the process of analyzing the text in Slack channel communication to better understand the content and timing of messages.

We capture performance primarily in three ways. First, we measure the cohesion of a translated document using an indicator equal to one if a team submitted a single, fully translated document. In total, 34% of teams submitted a cohesive document.

Second, we measure the quality of the translation output using the average ratings from a panel of four expert translators (Arnold et al., 1995; Weng et al., 2019). The same four experts reviewed all the documents translated into Swahili, and a separate four expert translators reviewed all the documents translated into Bengali. Each document was rated on a scale of 1-9, with 9 representing a perfectly clear and understandable document without any inappropriate grammar or words and 1 representing a completely incomprehensible document. The full rating scale definition is provided in Appendix B. For teams in which workers submitted their individual 2-page translations separately, rather than submitting a single fully translated document, quality is measured as the average quality of the two halves across the four evaluators. We did not ask the translators to consider that the two documents were not cohesive in their evaluations in order to have a measure of individual quality that was separate from our measure of cohesion. This allows us to capture the possibility that team work positively or negatively contributes to individual teammembers' work.[11] Similarly, if only one half of the document was submitted, we measure quality based on the evaluations of that half of the document. We separately capture whether an incomplete submission was made in a variable equal to one if only one half or none of the translated document was submitted, and zero otherwise. On average, translation quality was about 6 out of 9 and 15% of teams had incomplete submissions.

Lastly, we combine our quality measures into an aggregate "high performance" measure that is equal to one if a team submitted an above sample median quality and cohesive translated document. About 24% of team submissions meet this criteria of high performance.

[11] On one hand, communication may allows team members to provide each other with useful advice on their respective document sections. On the other hand, efforts to communicate may take time away or distract from individual translation work.

3.2 Independent Variables

Our primary independent variable of interest is whether or not a team work under a flexible or constrained schedule. As discussed in section 2, we hired a small pool of Swahili translators in our flexible schedule condition at the end of our second hiring round.

Our two other independent variables of interest are whether the team was hired to translate the document into Bengali or into Swahili, and whether or not a team is made up of two members of the same gender. We capture the gender of workers from their names and profile pictures on their hiring platform profile page. There were signicantly more males than females who applied for the jobs, so 62% of teams are made up of two workers from the same gender, and only about 6% of these are made up of two females.

We also capture proxies for team capabilities based on the information provided on workers' hiring platform profile pages. In particular, we combine team member experience on the platform to generate a total measure of platform work experience per team. On average, teams have about 6 previous jobs on the platform. To capture performance on prior tasks, we average the aggregated ratings team members received from prior employers on the platform. There are no ratings for workers with no prior experience and, thus, we only have this quality measure for about 30% of teams. Consistent with low online ratings being uncommon, the average platform rating is 4.44 out of 5 (e.g. Pallais, 2014). We have very little variation in the amounts workers bid on the job, but we are able to observe the hourly wages workers advertise on their profile pages as a signal of what they expect to earn from work. To the extent that this captures workers' perceptions about their relative abilities, it may provide useful information for understanding differences across workers. The mean advertised wages among teams in our sample, calculated as the average wages of teammates within each team, is about 13.50USD.

We supplement these platform-specific proxies for ability with average education within the team. We generate a measure of education for each worker based on the level of education they report on their platform profile pages that is equal to one for workers with at most a bachelor's degree, two for workers with at most a master's degree, and three for workers with at most a doctorate. Average team education is about 0.95, or about a bachelor's degree for both team members. We also generate an indicator variable equal to one if either member of a team lives in a major city and zero otherwise. We consider the major cities in our sample to be Dhaka, Kolkata, Dar es Salaam, and Nairobi. These workers may have more experience working for foreign employers, and may differ in their translation capabilities if they are differentially exposed to English relative to workers in more rural regions. About 70% of teams have at least one team member living in these major cities.

Summary statistics for our independent variables are included under "Work & Team Characteristics" in Table 1.

3.3 Analysis Plan

The objective of our study is to estimate the overall differential performance impact of two different type of work schedules among teams of workers. This overall effect includes any effects due to differential individual motivations, worker-types selecting into the jobs, and team coordination. While the primary contribution of our analyses is to provide causal evidence on this aggregate effect, in an effort to disentangle these underlying mechanisms, we begin by assessing whether team characteristics differ across the two job types and by job type and language. Because we do not post jobs in the same language that offer different schedules at the same time, we cannot conclude that workers are choosing between the two types of schedule and demonstrating their preference through their application. However, if similar worker types are on the platform over time, differences in team characteristics across the job types are suggestive of differential worker-type selection across *Flex* and *Constrained*.

To assess the effects of flex versus constrained schedules on team coordination and performance, we estimate the following baseline regression model:

$$Y_i = \alpha + \beta_1 \text{Flexible Schedule}_i + \beta_2 \text{Bengali}_i + hiringround_i + \varepsilon_i, \tag{1}$$

where Y_i is a measure of team i communication or performance, $FlexibleSchedule$ equals one if team i is under *Flex* and zero if it is under *Constrained*, and *Bengali* equals one if team i was hired to translate the document from English to Bengali and zero if it was hired to translate the document from English to Swahili. We include the translation language indicator in all our regression analyses because this variation is part of our randomization.[12] We control for hiring round fixed effects to address changes in applicant pools or circumstances across our three rounds of hiring.

We also analyze whether the inclusion of team characteristics as control variables to our baseline model changes our estimated treatment effects. If differential worker selection into *Flex* and *Constrained* treatments drives differential performance effects across the two schedules, conditional on our measures of team characteristics capturing some of the relevant variation in worker selection, including these controls should change our treatment effects (e.g. Oster, 2019). Thus, comparing coefficients with and without these controls

[12]Given that our RCT has a factorial design, the interaction between schedule type and language is necessary for assessing the significance of our findings (Muralidharan et al., 2019). As we show in Table 5, this interaction is not zero for some outcomes and, thus, we verify the robustness of our findings to the inclusion of the interaction as a control in our subsequent analyses.

allows us to provide suggestive evidence on the extent to which different scheduling arrangements improve performance through the types of workers they attract.

Ongoing data collection and cleaning will allow us to assess whether communication patterns are suggestive of coordination challenges, whether our sample workers differentially select out of a follow-up job opportunity depending on whether they are offered a flexible or constrained schedule[13], and whether negative effects of flex scheduling on coordination are mitigated as team members have more experience working together. These analyses will be included in the next draft of this paper.

4 Results

4.1 Team Characteristics Across Flexible and Constrained Schedules

Differences in team characteristics across the flex and constrained schedules are presented in Table 2. They demonstrate that gender composition is similar across the two types of schedules, as is whether or not a team member is from a major city, platform experience, and platform ratings. The average advertised wage is about 1.50USD higher among flex teams, but this difference is not significant. The only statistically significant difference among our observables is in average education, with teams of workers who applied to the flex schedule job having significantly less education than teams of workers who applied to the constrained schedule job.

We also examine team characteristic differences across flexible and constrained applicants by translation language (see Appendix Table A1). These differences demonstrate that the sample average differences in education across flex and constrained teams are driven by Swahili translators. They also demonstrate that Bengali female translators are more likely to apply to flex jobs than constrained jobs, but that this difference is, if anything, reversed among Swahili translators.[14]

Combined, these comparisons suggest some differential selection into scheduling types by applicants that differ across the language types. This suggests that preferences for flex schedules may be context specific.

[13] In our second round of hiring, we are contacting our sample workers with a job opportunity in a randomized order and, thus, we are able to observe selection directly.

[14] While teams made up of two females are no more common in flexible jobs than constrained jobs among the Bengali translators, teams of two females are very uncommon to begin with. Where it is apparent that Bengali translation female applicants are less likely to apply to constrained positions is in the increased likelihood of mixed gender teams in flex schedules.

4.2 Differences in Coordination and Performance Across Flexible and Constrained Schedules

Table 3 presents average differences in coordination and performance across teams working under the flexible and constrained schedules. On average, teams in the flexible treatment coordinate less and perform worse across most measures, but these differences are not significant.

We also present differences in survey responses across the two schedule types. Teams assigned flexible work schedules appear to have higher opinions of their teammate's efforts. In particular, flex schedule teammembers are more likely to both report that their teammate did the same amount of work as them. While the likelihood that both teammates report doing more work than the other is significantly higher among flex teams, the averages are very low. Moreover, the likelihood that at least one teammate reports doing more work than the other is significantly lower in flex schedules. These patterns may suggest that flex schedule teammates had better experiences in the job, or that each teammate is less aware of what the other worked on.

Table A2, in the appendix, presents average differences in coordination and performance across teams working under *Flex* and *Constrained* schedules by Bengali and Swahili translators. These differences demonstrate that Bengali translation teams coordinate and perform very similarly across the two schedules. In contrast, Swahili translation teams under *Flex* coordinate less on average, though not significantly so. Their translation quality is significantly worse under *Flex* than *Constrained*. Moreover, while flexible schedule Bengali translators are more likely to complete the survey than those under constrained schedules, the reverse is true for Swahili translators. To the extent that completing the survey captures engagement with the job, this differential completion suggests that flex schedules are associated with higher engagement among Bengali workers, and less engagement among Swahili workers.

We analyze coordination and performance differences more systematically by estimating equation 1 and present these findings in Table 4. Panel A of Table 4 presents the estimates from regressions with no controls for team characteristics, and Panel B presents the estimates from regressions with controls for team characteristics.[15] All regressions include hiring round fixed effects. The estimated coefficients on the flexible schedule coefficients are very similar in regressions with and without controls or team characteristics suggesting that differential selection into the schedule types is not the primary driver of average team coordination and performance.

[15] We do not control for team average UpWork ratings because we have observations of this characteristics for less than half our sample. However, our estimates are unchanged if we use a measure of UpWork experience that combines experience and average ratings and assigns a value of '0' to teams with no ratings as a control.

These estimates demonstrate that flexible schedules reduce the likelihood that teams are observed communicating over Slack, and significantly so when team characteristics are controlled for. In addition, they demonstrate that high performance is less likely among teams working in flexible schedules though, again, the statistical significance of this coefficients changes when team characteristics are controlled for. The direction of coefficients on the flexible schedule estimates are all consistent with worse team coordination and performance under *Flex* than *Constrained*. However, coefficient standard errors demonstrate that these estimates are generally noisy. Combined, our estimates suggest a weakly negative effect of flexible schedules on remote team coordination and performance.

One possible reason that the relationship between scheduling and outcomes is not stronger than what we find is that coordinating schedules in our context is not important for team performance. For example, if team coordination does not help improve the translation output, ease of scheduling team interactions should not impact performance. To examine whether coordination in our context is positively or negatively related to performance, we analyze the relationship between whether or not both team members in a team join Slack or not and our measures of team performance. This analysis demonstrates a consistently positive relationship between joint attempts to coordinate and team performance (see Appendix Table A3). This supports our assumption that the translation task we assign benefits from team coordination.

4.3 Heterogeneous Treatment Effects

An alternative potential explanation for the relatively weak relationship between flexible schedules and team outcomes that is consistent with average performance differences across translation languages and schedules is that flexible schedules are more or less of a barrier to coordination in different types of teams. In particular, among teams that have communication barriers unrelated to scheduling, the uncertainty flexible schedules adds to communication attempts may be challenging to overcome. In contrast, among teams with relatively few communication barriers, this uncertainty may be balanced out by flexible schedule benefits like increased motivation.

4.3.1 Heterogeneous Treatment Effects by Translation Language

To test for potential heterogenous effects of flexible scheduling on team work, we begin by analyzing any differential effects by Swahili and Bengali translation teams. These results are presented in Table 5. We again separately present the estimates from regressions with (Panel B) and without (panel A) controls for team characteristics to assess whether differential selection impacts treatment effects. As in our average

treatment estimates, our coeffients estimates by language are not sensitive to the inclusion of these controls.

Consistent with mean differences across flexible and constrained schedules across the two languages, our coefficient estimates demonstrate that flexible schedules led to less coordination and worse team performance among Swahili teams but not among Bengali teams. In particular, teams of Swahili translators are significantly less likely to communicate and produce signicantly worse quality translations when working under flexible schedules relative to constrained ones. In contrast, Bengali translation teams coordinate and perform no differently across the two types of schedules. One potential reason that flexible schedules have negative implications for Swahili translation but not Bengali translation teams is because the Swahili teams are more likely to be ethnically diverse (Awaworyi Churchill, 2017; Miguel, 2006). We will test for this directly by analyzing whether ethnic diversity based on teammate last names drives the differential impacts of *Flex* over *Constrained* by language.

4.3.2 Heterogeneous Treatment Effects by Gender Diversity

We also analyze differential effects of flexible scheduling by gender homogeneous and gender diverse teams. These estimates are presented in Table 6 with Panel A and Panel B presenting coefficients without and with team characteristic controls, respectively. These results demonstrate that flexible scheduling has significantly negative coordination and performance effects on mixed gender teams, but no significant impacts on same-gender teams. In particular, both teammates on mixed gender teams are less likely to join Slack, they are less likely to communicate with each other, less likely to produce a cohesive translated document, and less likely to produce high overall quality output. We verify that these results are robust to including an interaction between the flexible schedule indicator and translation language (see Appendix Table A5) (Muralidharan et al., 2019).

Given the small sample of teams made up of two females in the non-gender diverse group of teams, our estimates of the negative effects of flex scheduling may be driven by females performing less well under flexible than constrained schedules rather than by gender diversity. We explore this possibility by restricing the sample to same gender teams and interacting an indicator for teams made up of two males with the indicator for flexible schedule. Results of this analysis are presented in Appendix Table A4, and demonstrate that flex schedules do not decrease the likelihood of coordination or of submitting a cohesive document among male or female teams. A possible reason for these differences is that flexible schedules require more time to overcome coordination challenges and females, on average, have less slack time (e.g. **?**Goldin, 2020) making work completion in flexible schedules harder. However, we also find that female teams are less likely to

complete their translation task and more likely to perform worse overall when in flexible schedules relative to constrained ones. An important caveat is that there are only 13 two-female teams each in our flexible and constrained groups.

Another possibility for why mixed gender teams are less likely to communicate is because of gender discrimination by either male or female translators. In particular, upon learning the gender of their teammate (as revealed by names), teammates may opt not to try to coordinate if they believe a given gender is less capable. We examine whether the lack of attempts to coordinate in mixed gender teams are disproportionately driven by either males or females by examing whether one or the other is less likely to join the team Slack room when in a mixed gender team. We do not find evidence that the lack of coordination in these teams is predominantly driven by one or the other gender. Moreover, we do not find any evidence from survey responses that mixed or same gender teams had different perspectives on their teammates' contributions across flexible and constrained schedules (see Appendix Table A6).[16]

5 Conclusion

In this paper, we present the results of a field experiment designed to test the implications of flexible scheduling on the performance of remote teams. While flexible scheduling is frequently cited by workers as an important benefit of remote work, it is not a necessary condition of working remotely and it is theoretically unclear whether the coordination challenges it introduces for teams our outweighed by potential benefits like enhanced worker motivation.

We run our experiment among teams of Swahili and Bengali translators on an online labor market. We allows for differential selection into schedule types by advertising that teams will be expected to work within a given time window among constrained job postings, and not among flexible job postings. This allows us to estimate the combined effect of flexible schedules on performance that could include effects due to differential selection, motivation, and coordination.

Our analysis of the results from this experiment demonstrates three main results. First, we find some evidence of negative selection into flex jobs—hired teams made up of workers who applied to the flex jobs have significantly less education, and sizeably but insignificantly less UpWork experience and lower UpWork ratings. Moreover, we find that this differential selection is primarily driven by applicants to the Swahili translation jobs and not the Bengali translation jobs. However, controlling for team characteristics does

[16]This analysis is restricted to teams in which both workers filled in the survey and, thus, may be biased towards not finding negative sentiments if workers who complete the survey are generally more satisfied with the job than those who do not. However, even among this sample, we find significant negative effects of flexible scheduling among gender diverse teams.

not change our estimated effects of flex scheduling on team coordination or performance, suggesting that selection into schedule types is not the primary driver of our findings. Second, we find weak evidence that teams in the flexible schedule condition are less likely to coordinate with each other, and weak evidence that they perform worse on the task, due to a combination of lower translation quality and lower document cohesion. Third, we find significant heterogeneity in the impacts of flex scheduling on team performance. In particular, we find that Swahili translator teams coordinate significantly less and perform significantly worse in flex than in constrained schedules, but that Bengali translator teams perform equally well in both. Similarly, we find mixed gender teams coordinate significantly and less perform significantly worse in flex than in constrained schedules, but that same gender translator teams perform equally well in both. We also find evidence that flexible schedules lead to worse performance, but no less coordination, among teams made up of two females.

Possible explanations for our heterogeneous treatment effect findings are that team diversity introduces coordination barriers that flex scheduling exacerbates. Alternatively, differential treatment effects by language could be driven by work culture differences. Differential performance by gender diverse and homogenous teams may be partly due to females having less time to overcome scheduling uncertainty than males.

We are currently running a second round of hiring with the same work teams in our current data to examine whether the negative effects of flexible scheduling are reduced as teammates have more experience working together, and to better understand differential selection across jobs. We are also categorizing worker last names into ethnicities to analyze whether ethnic diversity is driving our findings on the differential effects of flexible scheduling across work contexts, and analyzing the timing and content of Slack communication to understand how coordination was undertaken.

6 Tables

Table 1: Sample Summary Statistics

	Mean	Std. Dev.	N
Outcomes			
Teammates Joined Slack	0.342	0.475	260
Any Slack Channel Communication	0.309	0.463	259
Cohesive Output	0.340	0.475	259
Incomplete Output	0.154	0.361	260
Translation Quality	6.058	1.212	230
High Performance (0/1)	0.235	0.425	260
Work & Team Characteristics			
Flexible Schedule	0.538	0.499	260
Bengali Translation	0.465	0.500	260
No Gender Diversity	0.619	0.487	260
Both Female	0.100	0.301	260
Both Male	0.519	0.501	260
At Least 1 from Big City	0.712	0.454	260
Average Team Education	0.952	0.457	260
Total Platform Experience	6.256	15.266	250
Average UpWork Job Rating	4.443	1.328	74
Average Advertised Wage	13.584	8.138	256
Survey Responses			
Both teammates report doing most work	0.021	0.144	142
Both teammates report doing equal work	0.641	0.481	142
At least 1 teammate reports doing most work	0.359	0.481	142
At least 1 teammate reports communication	0.709	0.455	227
Both teammates complete survey	0.592	0.492	260

Note: This table presents our sample summary statistics.

Table 2: Work Characteristics by Treatment

	Constrained	**Flexible**	p-val of Diff
Bengali Translation	0.508	0.429	0.200
No Gender Diversity	0.642	0.600	0.492
Both Female	0.108	0.093	0.679
Both Male	0.533	0.507	0.675
At Least 1 from Big City	0.692	0.728	0.515
Average Team Education	1.008	0.904	0.065*
Total Platform Experience	7.195	5.482	0.378
Average Upwork Rating	4.620	4.257	0.243
Average Advertised Wage	12.769	14.271	0.142

Note: This table presents our average team characteristics by flexible and constrained treatments to examine whether different workers select into different job types. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Outcomes by Treatment

	Constrained	**Flexible**	p-val of Diff
Teammates Joined Slack	0.350	0.336	0.810
Any Slack Channel Communication	0.336	0.286	0.383
Cohesive Output	0.361	0.321	0.501
Incomplete Output	0.15	0.157	0.874
Translation Quality	6.092	6.022	0.658
High Performance	0.258	0.214	0.405
Both teammates report doing most	0.000	0.043	0.077*
Both teammates report doing equal work	0.570	0.714	0.073*
At least 1 teammate reports doing most work	0.431	0.286	0.073*
At least 1 teammate reports communication	0.693	0.726	0.590
Both teammates complete survey	0.693	0.726	0.590

Note: This table presents our average outcomes by flexible and constrained treatments. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Average Effects of Flex Scheduling on Team Coordination & Performance

	(1) Joined Slack	(2) Slack Communication	(3) Cohesive	(4) Incomplete	(5) Quality	(6) High Peformance
Panel A: No Controls for Team Characteristics						
Flexible	-0.064	-0.093	-0.073	0.030	-0.131	-0.093*
	(0.060)	(0.058)	(0.061)	(0.046)	(0.157)	(0.054)
Bengali	-0.011	0.034	-0.140**	0.027	-0.573***	-0.106*
	(0.060)	(0.058)	(0.061)	(0.046)	(0.158)	(0.054)
Observations	246	246	246	246	217	246
R^2	0.066	0.059	0.032	0.010	0.075	0.050
Panel B: With Controls for Team Characteristics						
Flexible	-0.065	-0.103*	-0.074	0.013	-0.152	-0.090
	(0.061)	(0.059)	(0.062)	(0.046)	(0.158)	(0.055)
Bengali	-0.039	0.007	-0.136**	0.033	-0.490***	-0.102*
	(0.064)	(0.062)	(0.065)	(0.049)	(0.169)	(0.058)
No Gender	0.017	0.012	0.016	-0.065	-0.118	0.005
Diversity	(0.062)	(0.060)	(0.063)	(0.047)	(0.162)	(0.056)
At Least 1	-0.114*	-0.091	0.021	-0.002	0.425**	0.007
Big City	(0.068)	(0.066)	(0.069)	(0.052)	(0.180)	(0.062)
Average	-0.077	-0.149**	-0.020	-0.071	0.068	0.025
Education	(0.068)	(0.066)	(0.069)	(0.051)	(0.184)	(0.062)
Platform	0.001	0.001	0.000	0.000	0.004	0.001
Experience	(0.002)	(0.002)	(0.002)	(0.002)	(0.006)	(0.002)
Average	-0.000	-0.001	-0.000	0.004	0.003	0.001
Wage	(0.004)	(0.004)	(0.004)	(0.003)	(0.012)	(0.003)
Observations	246	246	246	246	217	246
R^2	0.084	0.088	0.033	0.032	0.104	0.053

Note: Standard errors in parentheses. Regressions include hiring round fixed effects. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Average Effects of Flex Scheduling on Team Coordination & Performance by Translation Language

	(1) Joined Slack	(2) Slack Communication	(3) Cohesive	(4) Incomplete	(5) Quality	(6) High Peformance
Panel A: Without Controls for Team Characteristics						
Flexible Schedule	-0.077	-0.145*	-0.043	0.055	-0.368*	-0.103
	(0.080)	(0.078)	(0.081)	(0.061)	(0.213)	(0.072)
Bengali Translation	-0.028	-0.031	-0.103	0.059	-0.831***	-0.118
	(0.089)	(0.087)	(0.091)	(0.068)	(0.223)	(0.081)
FlexibleXBengali	0.031	0.119	-0.067	-0.058	0.514	0.022
	(0.121)	(0.118)	(0.122)	(0.092)	(0.315)	(0.109)
Flexible+Flex XBengali=0	0.608	0.767	0.229	0.970	0.529	0.322
Observations	246	246	246	246	217	246
R^2	0.066	0.063	0.034	0.011	0.087	0.050
Panel B: With Controls for Team Characteristics						
Flexible	-0.078	-0.159**	-0.048	0.045	-0.412*	-0.103
	(0.082)	(0.079)	(0.084)	(0.062)	(0.216)	(0.075)
Bengali	-0.054	-0.060	-0.105	0.071	-0.759***	-0.116
	(0.091)	(0.088)	(0.093)	(0.070)	(0.228)	(0.083)
FlexibleXBengali	0.029	0.127	-0.059	-0.073	0.565*	0.027
	(0.123)	(0.120)	(0.126)	(0.094)	(0.322)	(0.112)
Flexible+Flex XBengali=0	0.705	0.820	0.259	0.715	0.661	0.810
Observations	246	246	246	246	217	246
R^2	0.084	0.092	0.034	0.034	0.117	0.053

Note: Standard errors in parentheses. Regressions include hiring round fixed effects, and Panel B includes the controls included in the regressions presented in Panel B of Table 4. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Average Effects of Flex Scheduling on Team Coordination & Performance by Gender Diversity

	(1) Joined Slack	(2) Slack Communication	(3) Cohesive	(4) Incomplete	(5) Quality	(6) High Peformance
Panel A: Without Controls for Team Characteristics						
Flexible Schedule	-0.193**	-0.210**	-0.200**	0.016	-0.343	-0.197**
	(0.097)	(0.094)	(0.098)	(0.074)	(0.256)	(0.087)
No Gender Diversity	-0.100	-0.096	-0.100	-0.072	-0.282	-0.086
	(0.092)	(0.090)	(0.094)	(0.071)	(0.232)	(0.084)
FlexibleXGender Diversity	0.213*	0.193	0.209*	0.019	0.340	0.171
	(0.124)	(0.121)	(0.126)	(0.095)	(0.330)	(0.112)
Flex+FlexXNo Diversity=0	0.793	0.823	0.904	0.561	0.988	0.706
Observations	246	246	246	246	217	246
R^2	0.078	0.069	0.044	0.017	0.082	0.059
Panel B: With Controls for Team Characteristics						
Flexible	-0.181*	-0.202**	-0.207**	0.007	-0.455*	-0.203**
	(0.098)	(0.095)	(0.100)	(0.075)	(0.257)	(0.089)
No Gender Diversity	-0.089	-0.079	-0.106	-0.070	-0.367	-0.099
	(0.094)	(0.091)	(0.095)	(0.072)	(0.233)	(0.085)
FlexibleXNo Gender Diversity	0.189	0.164	0.218*	0.010	0.495	0.185
	(0.126)	(0.122)	(0.128)	(0.096)	(0.334)	(0.114)
Flex+FlexXNo Diversity=0	0.787	0.712	0.900	0.790	0.969	0.799
Observations	246	246	246	246	217	246
R^2	0.093	0.095	0.045	0.032	0.113	0.063

Note: Standard errors in parentheses. Regressions include hiring round fixed effects and controls for translation language, and Panel B includes the controls included in the regressions presented in Panel B of Table 4. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Al-Awawdeh, Nabil, "Translation Between Creativity and Reproducing An Equivalent Original Text," *Psychology and Education Journal*, 2021, *58* (1), 2559–2564.

Angelici, Marta and Paola Profeta, "Smart-working: Work flexibility without constraints," 2020.

Arnold, D, L Balkan, R Lee Humphreys, S Meijer, and L Sadler, "Machine translation: An introductory guide," 1995.

Barrero, Jose Maria, Nicholas Bloom, Steven J Davis, Brent H Meyer, and Emil Mihaylov, "The Shift to Remote Work Lessens Wage-Growth Pressures," Technical Report, National Bureau of Economic Research 2022.

Berge, Lars Ivar Oppedal, Kartika Sari Juniwaty, and Linda Helgesson Sekei, "Gender composition and group dynamics: Evidence from a laboratory experiment with microfinance clients," *Journal of Economic Behavior & Organization*, 2016, *131*, 1–20.

Bloom, Nicholas, Ruobing Han, and James Liang, "How hybrid working from home works out," Technical Report, National Bureau of Economic Research 2022.

Burbano, Vanessa C, "The demotivating effects of communicating a social-political stance: Field experimental evidence from an online labor market platform," *Management Science*, 2021, *67* (2), 1004–1025.

Burgess, Simon, Carol Propper, Marisa Ratto, Stephanie von Hinke Kessler Scholder, and Emma Tominey, "Smarter task assignment or greater effort: the impact of incentives on team performance," *The Economic Journal*, 2010, *120* (547), 968–989.

Catalini, Christian, "Microgeography and the direction of inventive activity," *Management Science*, 2018, *64* (9), 4348–4364.

Churchill, Sefa Awaworyi, "Microfinance and ethnic diversity," *Economic Record*, 2017, *93* (300), 112–141.

de Ven, Andrew H Van, Andre L Delbecq, and Richard Koenig Jr, "Determinants of coordination modes within organizations," *American sociological review*, 1976, pp. 322–338.

Dessein, Wouter and Tano Santos, "Adaptive organizations," *Journal of political Economy*, 2006, *114* (5), 956–995.

Dhawan, Erica and Tomas Chamorro-Premuzic, "How to collaborate effectively if your team is remote," *Harvard Business Review*, 2018.

Felfe, Christina, "The motherhood wage gap: What about job amenities?," *Labour Economics*, 2012, *19* (1), 59–67.

Freeman, Richard B, Xiaofei Pan, Xiaolan Yang, and Maoliang Ye, "Team Incentives and Lower Ability Workers: An Experimental Study on Real-Effort Tasks," Technical Report, National Bureau of Economic Research 2022.

Goldin, Claudia Dale, "Journey across a Century of Women," *NBER Reporter*, 2020, *3*, 1–7.

Groysberg, B, "How remote work changes what we think about onboarding," *Cambridge, MA: Harvard Business School, April*, 2020, *27*.

Hasan, Md Arid, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan, "Neural machine translation for the Bangla-English language pair," in "2019 22nd International Conference on Computer and Information Technology (ICCIT)" IEEE 2019, pp. 1–6.

Hinds, Pamela J and Mark Mortensen, "Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication," *Organization science*, 2005, *16* (3), 290–307.

Hjort, Jonas, "Ethnic divisions and production in firms," *The Quarterly Journal of Economics*, 2014, *129* (4), 1899–1946.

Holmstrom, Bengt, "Moral hazard in teams," *The Bell journal of economics*, 1982, pp. 324–340.

Hossain, Tanjim, Elizabeth Lyons, and Aloysius Siow, "Fairness considerations in joint venture formation," *Experimental Economics*, 2019, pp. 1–36.

Kelemen, Thomas K, Samuel H Matthews, Xin an Zhang, Bret H Bradley, and Huihua Liu, "When does gender diversity enhance team performance? The dual need for visionary leadership and team tenure," *Journal of Applied Social Psychology*, 2020, *50* (9), 501–511.

Lazear, Edward P, "Globalisation and the market for team-mates," *The Economic Journal*, 1999, *109* (454), 15–40.

Lyons, Elizabeth, "Team production in international labor markets: Experimental evidence from the field," *American Economic Journal: Applied Economics*, 2017, *9* (3), 70–104.

Maznevski, Martha L and Katherine M Chudoba, "Bridging space over time: Global virtual team dynamics and effectiveness," *Organization science*, 2000, *11* (5), 473–492.

McNall, Laurel A, Aline D Masuda, and Jessica M Nicklin, "Flexible work arrangements, job satisfaction, and turnover intentions: The mediating role of work-to-family enrichment," *The Journal of psychology*, 2009, *144* (1), 61–81.

Miguel, Edward, "Ethnic diversity and poverty reduction," *Understanding poverty*, 2006, pp. 169–184.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich, "Factorial designs, model selection, and (incorrect) inference in randomized experiments," Technical Report, National Bureau of Economic Research 2019.

Nitzke, Jean, *Problem solving activities in post-editing and translation from scratch: A multi-method study*, Language Science Press, 2019.

Oster, Emily, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

Pallais, Amanda, "Inefficient hiring in entry-level labor markets," *American Economic Review*, 2014, *104* (11), 3565–3599.

Pauw, Guy De, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver, "Towards english-swahili machine translation," in "Research Workshop of the Israel Science Foundation" 2011.

Rayo, Luis, "Relational incentives and moral hazard in teams," *The Review of Economic Studies*, 2007, *74* (3), 937–963.

Weidmann, Ben and David J Deming, "Team players: How social skills improve team performance," *Econometrica*, 2021, *89* (6), 2637–2657.

Weng, Wei-Hung, Yu-An Chung, and Peter Szolovits, "Unsupervised clinical language translation," in "Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining" 2019, pp. 3121–3131.

Yang, Longqi, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht et al., "The effects of remote work on collaboration among information workers," *Nature human behaviour*, 2022, *6* (1), 43–54.

For Online Publication

Appendix A Additional Tables

Table A1: Work Characteristics by Treatment

	Constrained	**Flexible**	p-val of Diff
Bengali Translation Teams			
No Gender Diversity	0.688	0.458	0.010**
Both Female	0.098	0.068	0.549
Both Male	0.590	0.389	0.028**
At Least 1 from Big City	0.639	0.559	0.375
Average Team Education	0.984	0.924	0.501
Total Platform Experience	9.037	6.579	0.438
Average Upwork Rating	4.377	3.676	0.197
Average Advertised Wage	11.603	11.591	0.993
Swahili Translation Teams			
No Gender Diversity	0.593	0.704	0.176
Both Female	0.119	0.111	0.891
Both Male	0.475	0.592	0.169
At Least 1 from Big City	0.746	0.852	0.118
Average Team Education	1.034	0.889	0.050*
Total Platform Experience	5.508	4.700	0.739
Average Upwork Rating	4.953	4.838	0.359
Average Advertised Wage	13.915	16.248	0.123

Note: This table presents our average team characteristics by flexible and constrained treatments and by translation language to examine whether different workers select into different job types. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2: Work Characteristics by Treatment

	Constrained	**Flexible**	p-val of Diff
Bengali Translation Teams			
Teammates Joined Slack	0.311	0.339	0.750
Any Slack Channel Communication	0.300	0.339	0.652
Cohesive Output	0.300	0.237	0.445
Incomplete Output	0.197	0.152	0.528
Translation Quality	5.669	5.884	0.461
High Performance	0.180	0.169	0.877
Both teammates report doing most work	0.000	0.032	0.338
Both teammates report doing equal work	0.414	0.613	0.127
At least 1 teammate reports doing most work	0.586	0.387	0.127
At least 1 teammate reports communication	0.737	0.750	0.874
Both teammates complete survey	0.492	0.644	0.094*
Swahili Translation Teams			
Teammates Joined Slack	0.390	0.333	0.494
Any Slack Channel Communication	0.373	0.247	0.110
Cohesive Output	0.424	0.383	0.627
Incomplete Output	0.102	0.160	0.319
Translation Quality	6.517	6.146	0.004***
High Performance	0.340	0.247	0.237
Both teammates report doing most	0.000	0.051	0.136
Both teammates report doing equal work	0.674	0.795	0.224
At least 1 teammate reports doing most work	0.326	0.205	0.224
At least 1 teammate reports communication	0.649	0.702	553
Both teammates complete survey	0.746	0.519	0.006**

Note: This table presents our average team characteristics by flexible and constrained treatments and by translation language to examine whether different workers select into different job types. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Relationship between Observed Coordination and Team Success

	(1) Cohesive	(2) Incomplete	(3) Quality	(4) High Performance
Teammates Joined Slack	0.472***	-0.126*	0.413*	0.316***
	(0.089)	(0.072)	(0.241)	(0.082)
Flexible Schedule	-0.021	0.005	0.047	-0.069
	(0.071)	(0.057)	(0.194)	(0.065)
Teammates Joined Slack X Flexible Schedule	-0.070	-0.001	-0.577*	-0.004
	(0.118)	(0.096)	(0.347)	(0.110)
Observations	246	246	217	246
R^2	0.209	0.059	0.118	0.166

Note: Standard errors in parentheses. Regressions include hiring round fixed effects and controls for translation language, and Panel B includes the controls included in the regressions presented in Panel B of Table 4. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Relationship between Gender of Same Gender Teams and Team Coordination and Performance

	(1) Joined Slack	(2) Slack Communication	(3) Cohesive	(4) Incomplete	(5) Quality	(6) High Peformance
Flexible Schedule	0.017	-0.090	-0.243	0.271**	-0.435	-0.471***
	(0.193)	(0.184)	(0.200)	(0.134)	(0.481)	(0.174)
Both Male	0.144	0.211	-0.246	0.163	-0.498	-0.326**
	(0.149)	(0.143)	(0.155)	(0.104)	(0.353)	(0.135)
FlexibleXBoth Males	-0.013	0.035	0.304	-0.286*	0.537	0.539***
	(0.211)	(0.201)	(0.218)	(0.146)	(0.523)	(0.190)
Observations	150	150	150	150	132	150
R^2	0.110	0.139	0.047	0.071	0.147	0.109

Note: Standard errors in parentheses. Regressions include hiring round fixed effects and controls for translation language and the controls included in the regressions presented in Panel B of Table 4. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. The sample is restricted to teams with two males or teams with two females. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Average Effects of Flex Scheduling on Team Coordination & Performance by Gender Diversity Including Flexible Schedule and Language Interaction

	(1) Joined Slack	(2) Slack Communication	(3) Cohesive	(4) Incomplete	(5) Quality	(6) High Peformance
Panel A: Without Controls for Team Characteristics						
Flexible Schedule	-0.219*	-0.278**	-0.177	0.054	-0.638**	-0.215**
	(0.113)	(0.110)	(0.115)	(0.087)	(0.309)	(0.103)
No Gender Diversity	-0.097	-0.089	-0.103	-0.076	-0.261	-0.084
	(0.093)	(0.090)	(0.094)	(0.071)	(0.231)	(0.084)
FlexibleXNo Gender Diversity	0.217*	0.203*	0.206	0.013	0.416	0.173
	(0.125)	(0.122)	(0.126)	(0.095)	(0.332)	(0.113)
Flex+FlexXNo Diversity=0	0.983	0.405	0.753	0.339	0.358	0.615
Observations	246	246	246	246	217	246
R^2	0.078	0.074	0.044	0.020	0.094	0.059
Panel B: With Controls for Team Characteristics						
Flexible Schedule	-0.203*	-0.272**	-0.186	0.043	-0.816***	-0.224**
	(0.115)	(0.112)	(0.118)	(0.088)	(0.312)	(0.105)
No Gender Diversity	-0.087	-0.073	-0.108	-0.073	-0.347	-0.097
	(0.094)	(0.091)	(0.096)	(0.072)	(0.231)	(0.085)
FlexibleXNo Gender Diversity	0.193	0.175	0.215*	0.004	0.596*	0.188
	(0.126)	(0.122)	(0.129)	(0.097)	(0.335)	(0.115)
Flex+FlexXNo Diversity=0	0.919	0.286	0.765	0.513	0.362	0.677
Observations	246	246	246	246	217	246
R^2	0.093	0.100	0.046	0.034	0.130	0.064

Note: Standard errors in parentheses. Regressions include hiring round fixed effects and controls for translation language and an interaction between translation language and the flexible schedule indicator, and Panel B includes the controls included in the regressions presented in Panel B of Table 4. High performance is equal to one for teams that submitted a single finished document and received above median quality ratings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Flex Scheduling, Language, and Gender Diversity and Reported Differential Contributions among Teammates

	(1)	(2)
	Both Report Same Workload	
Flexible Schedule	0.136	0.149
	(0.086)	(0.139)
Bengali Translation	-0.210**	-0.213**
	(0.088)	(0.091)
No Gender Diversity	-0.014	-0.005
	(0.084)	(0.121)
FlexibleXNo Gender Diversity		-0.020
		(0.175)
Observations	136	136
R^2	0.095	0.095

Note: Standard errors in parentheses. Regressions include hiring round fixed effects and the controls included in the regressions presented in Panel B of Table 4. The dependent variable is equal to one if both teammates report they did the same amount of work as the other. The sample is restricted to teams in which both workers completed the survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix B Data Appendix

Appendix B.1 Blog Post for Translation in First Hiring Round

Study Finds that Anyone Can Be an Innovator

Professor John Smith holds dual faculty positions at the Policy and Economics departments. His multidisciplinary research interests and include environmental, health, development, and innovation economics, among other areas. Professor Smith's favorite color is Cyan. He plays with his crickets and watches cricket with his daughter in his leisure time.

Professor Jane Doe's research focuses on the intersection between technology and innovation strategy, international management and organizational economics. Her current projects include using data to analyze firm hiring and organization in international labor markets, studying the effects of entrepreneurship training on career decisions, and more. In her spare time, Professor Doe enjoys running with her silver-haired Schnauzer, and working on her root vegetable garden.

******** Students given incentives to innovate are just as skilled as the self-motivated, research finds

What are the traits of an innovator? Is it an inherent or learned quality? Existing theories and empirical research on how innovation occurs largely assume that it is an ingrained quality of the individual and that only people with this innate ability seek and attain jobs that require it; however new research shows this isn't the whole story.

A study undertaken by Professor of Economics John Smith and Professor of Management Jane Doe, tested these previously held notions by creating a contest for the University's engineering and computer science students. The competition was designed to answer the question of: Are persuaded innovators less able to innovate than those who naturally gravitate to innovative activities? As the study text highlights, this question is important because; Innovation has long been viewed as important for productivity and income growth. Organizational and government policy generally aims to foster innovation by targeting the quantity and quality of individuals employed in innovative careers. While innovative output is not concentrated within any single industry or field, one potentially important target for these policies is undergraduate science, technology, engineering, and math (STEM) students who acquire a disproportionate share of the hard skills required to introduce high impact innovations into the economy... despite the skills they accumulate during their Bachelor's degrees, many STEM students do not end up in careers that require them to apply these skills to innovation. For instance, in 2014 the US Census Bureau reported that 74% of people with STEM undergraduate degrees were not employed in STEM jobs despite a robust labor market in this area – nearly 2.4 million STEM job postings in the US are projected to go unfilled in 2018 alone. Part of this under-supply of labor for STEM jobs is likely because STEM degrees allow graduates to pursue a range of lucrative careers that do not require them to innovate, thus making the link between STEM education and innovative output weaker than may be expected.

Whether efforts to encourage more of these graduates to enter innovative careers will increase the innovativeness of the economy depends fundamentally on why they are not entering such careers in the first place. Are they self-selecting out because they know they are not well-suited to the task at hand, or are other factors shaping their career choices?

The mobile application contest the authors ran to test their research question was advertised on various channels and attracted around 100 students. In order to differentiate between self-selected innovators and "induced" innovators, a random subset of eligible students who did not sign up by the contest deadline were offered a monetary incentive of $100 to participate. In total, 190 students signed up.

Submissions between the two groups were evaluated anonymously by technology industry participants who acted as judges for the contest. They evaluated each proposed application across four categories; functionality, user-friendliness, novelty, and potential commercial value.

Though induced participants were less likely to be drawn from majors that provide the most relevant skills for the competition, in particular, electrical engineering and computer science, and had lower cumulative GPAs (specifically, cumulative GPAs of 2.5 and under), the success of these participants appeared statistically indistinguishable from those that were innately drawn to the competition.

"Finding the induced participants may have been technically less well equipped to compete, and equally

as talented, allows us to determine whether individuals who do not choose to innovate in the absence of an intervention are being held back by accurate beliefs about their ability to perform," said Professor John Smith. "This shows that psychological barriers that, if overcome, could meaningfully contribute to the innovation process."

Whether innovators can be created, and how they fare relative to those who self-select into innovative activities also has important implications for public and private policy. In addition, insight into the conditions under which productivity enhancing innovation occurs is critical for understanding economic development and can provide novel insights into the arise of new inventions, according to the authors.

The engineering contest entries were given a score of 1 to 5 on each category for a total score maximum of 20. The developers of the top 3 applications were awarded prize money.

"We selected students at UC San Diego's Jacobs School of Engineering since these students have technical capabilities to produce impactful inventions," Professor Doe said. "In addition, engineers are frequently the targets of interventions to increase innovative activity."

The published study describes the results as;

Despite the mean [characteristic] comparisons suggesting that the induced participants may be less well equipped to compete, the total output of these participants appears statistically indistinguishable from the output of those that self-selected into the competition. In particular, mean submissions across induced and self-selected innovators ... demonstrate that induced participants have slightly lower mean submission rates than self-selected ones but that this difference is far from statistically significant. With our sample size, we would have needed a difference in submission rate of 6 percentage points to detect a significant difference between the induced and self-selected sample which is several times larger than the 1.5 percentage points we observe. Importantly, while we cannot conclude that induced and self-selected participants have different submission rates, we can conclude that inducement increases total innovative output. Induced innovators' likelihood of submitting a project to the contest is significantly larger than zero.

The difference in average innovative output quality conditional on submitting between induced and self-selected participants is reasonably large, with self-selected participant output scoring more than 22% higher on average than induced participant output. Although this difference is not statistically significant (with our sample size, a difference in scores of 1.172 would be statistically significant), it provides suggestive evidence that induced innovators may have lower average quality of innovative output relative to self-selected innovators.

While average outcomes are important for understanding aggregate changes in innovative output caused by increasing the pool of innovators through inducement, whether the distribution of submission quality differs for the self-selected and induced participants is also important for understanding whether inducement can lead to increases in very high impact innovation. This is particularly important for innovation management and policy because the majority of returns from innovation are generated by a small minority of innovations. To investigate distributional differences in the quality of submissions, we plot the distribution of average rankings conditional on submissions for the self-selected and inducement samples. Consistent with the suggestive evidence that the quality of self-selected output is higher than that of the induced sample, it does appear that the induced sample has a higher frequency of low performance relative to the self-selected sample. At the same time, we do not find evidence that the likelihood of very high performance differs meaningfully for the two samples.

The impacts of encouragement on competition outcomes were a bit more surprising, Professors Smith and Doe noted.

The researchers randomly offered encouragement to subsets of both the induced and self-selected contest participants in order to examine the importance of confidence-building interventions on each sample. While encouragement had no impact on performance on average, students with above median GPAs performed significantly worse when they received additional encouragement whereas students with below median GPAs performed significantly better when they received additional encouragement. The study suggests one explanation for this:

Evidence from our post-contest survey responses suggests one reason our encouragement intervention harms the performance of high GPA students – it appears to increase the salience of the time commitment

required for the contest. In particular, we find that the encouragement treatment is associated with an approximately thirty percentage point increase in the likelihood that participants report not submitting a project for consideration by the judges due to time constraints. By contrast we find no relationship between being induced or encouraged and participants reporting that they did not submit due to the difficulty of the contest problem. S12 table reports these estimates. Whether the negative effect of encouragement on high GPA students also relates to the crowding out of intrinsic motivation remains an open question.

The researchers concluded, innovators can be created through inducement subsidies, but that targeting inducement is likely to be more cost effective. In particular, higher GPA students benefitted more from the inducement subsidies than lower GPA students did. The combined findings are summarized in the paper as follows;

Our study provides novel evidence that some STEM students may be selecting out of innovative activities based on their expected performance, while others select out based on the costs of participating. The results demonstrate that innovators can be created by subsidizing their initial entry into innovative tasks, but that targeting inducement towards those who select out due to their expected cost of participation rather than their expected performance is a more effective strategy to promote innovation. That such targeting can be based on relatively easy information to obtain, like training and GPA, suggests that such a strategy may be both practical and cost effective. In addition, we demonstrate that encouragement may also need to be targeted to improve the performance of those students who stand to benefit the most from it, and importantly, to avoid harming those who may respond negatively.

To better understand how innovation contests can be designed to increase participation in innovative activities, Professors Smith and Doe are now studying whether the type of awards offered in innovation contests change who participates in them, and how well they innovate. They are comparing a contest with a single large prize for the top ranked innovator, with a contest that has multiple prizes such that all of the top ranked innovators are awarded some money but the top ranked innovation is awarded less than in the winner-takes-all contest. In this contest, participants are working on cloud and medical technologies.

The winner-takes-all prize structure is riskier in that the likelihood a participant wins anything is lower and is, thus, the researchers expect it will reduce the likelihood that potential innovators invest any effort in the contest, but increase the performance of those who do vie for the prize. In contrast, the contest with multiple prizes provides some insurance that even solutions that are not best-in-class will be awarded some money. Thus, innovators who are less confident in their abilities may be more likely to invest effort in this contest but they may strive less intensively to be the very best.

Appendix B.2 Job Posting Content

Title: English to Bengali/Swahili Translation// Job Descriptors:

General translation services
$20 fixed price
Entry level
Translation deliverables: Localized content
Translation types: Translation
Other skills/expertise: Bengali/Swahili, English, English to Bengali/Swahili Translation// Description: We have four pages of English text that we would like translated to Bengali/Swahili. The text is a blog post written to communicate research findings to the general public. Machine translation is unable to accurately perform the translation.
We are looking to hire two people to complete this together, **so we will be asking which two-hour windows you are available during a 24-hour work window from the start of the contract**. Each person hired will be assigned to translate 2 of the 4 pages of text. *We would like this completed within 24 hours of hiring.**
Please apply if interested! *Bolded text is only included in the constrained schedule job posts. Italicized text is only included in the flexible schedule job posts.

Appendix B.3 Job Instructions

Thank you for applying to this job. Could you please send me 3-7 windows of time during which you're available for at least 2-hours of time to work on this job between 8:00 pm (today's date) and 8:00 pm (tomorrow's date) (Bangladesh/East Africa time)?*

Thank you for applying to and accepting this job! We are trying to better understand how translation performance can be improved and are grateful to have you working on this with us. Please read these instructions carefully.

We already have the machine-translated version of our document and are now only looking for human translations. We are hoping that your translation will be of higher quality than machine-translated versions. The document is a blog post written to communicate research findings to the general public.

We have hired a team of two to complete this translation because we think this type of job benefits from two sets of eyes working through it, and because of the limited window of time we have for the job to be completed. Your teammate on this task is TEAMMATE NAME. Could you please work on translating the first/last two pages (pages 1 and 2/pages 3 and 4) of the document? Your teammate will work on the last/first two pages.

Please coordinate with your teammate to make sure that you are approaching the translation similarly, so that the final translated document is cohesive. You and your teammate can also help each other if you run into any words or phrases that you are unsure how to best translate.

You and your teammate can each earn up to a $5 bonus on top of your contract amount if the final translation reads as a single, cohesive document. Your platform rating on this job will take into account the quality of translation of your translated pages, and how cohesive the final document is.

To make coordination with your teammate easier, we have set up a Slack channel for the two of you. You can communicate with your teammate using this channel: URL. Please join the Slack room, and I will add you to your Slack room. If you haven't use Slack before, you can access the channel through your web browser. Let me know if you have any trouble using it.

The two-hour time window during which both you and your teammate are both available to work on this task is X:00 to Y:00 AM/PM Bangladesh/East Africa time. Please be sure to be in the Slack channel at that time to begin working with your teammate.*

The deadline for the task to be completed is 24 hours from now. Please ensure that either you or your teammate submit the final fully translated text to us by this deadline. So that we can provide each team member with accurate ratings and feedback on the task, please let us know how well you and your teammate worked together after the final translated document has been submitted.

*Bolded text is only included for workers hired under the constrained schedule.

Appendix B.4 Post-Task Survey Questions

Q1 Please enter the name you use on your platform profile here

Q2 Did you communicate with your teammate at all while working on the translation task?

1. Yes, a little bit
2. Yes, a lot

3. No

Display This Question: If Did you communicate with your teammate at all while working on the translation task? != No
Q3 What did you and your teammate talk about?

Display This Question: If Did you communicate with your teammate at all while working on the translation task? != No
Q4 Did you find your communication with your teammate helpful for completing your work successfully?

1. Yes, very helpful
2. Yes, somewhat helpful
3. No, not at all helpful

Display This Question: If Did you communicate with your teammate at all while working on the translation task? = No
Q5 Did you make any attempt to communicate with your teammate?

1. Yes
2. No

Display This Question: If Did you communicate with your teammate at all while working on the translation task? = No
Q6 What was the main barrier to communicating with your teammate?

Q7 What part of the document did you translate?

1. Only the part of the document that was assigned to me
2. Only the part of the document that was assigned to my teammate
3. All of the document
4. None of the document

Q8 Do you think you did more work than your teammate did on the translation task?

1. Yes, I did more work than my teammate did on the task
2. No, my teammate did more work than I did on the task
3. No, we did the same amount of work on the task

Appendix B.5 Translation Evaluation Scale

The translated documents were evaluated along the following scale.

1. None of the document is understandable
2. Almost none of the document is understandable
3. Document is hard enough to understand that overall idea and takeaways are unclear.
4. Overall idea of the document is only vaguely clear after full reading of the document. Majority of the document is difficult to understand.
5. Overall idea of the document is clear, but not immediately so. Poor style, poor word choice, alternative expressions, untranslated words, or incorrect grammar are present throughout the document.
6. Overall idea of the document is immediately clear, but some of the text is hard to understand because of poor style, poor word choice, alternative expressions, untranslated words, or incorrect grammar.
7. Mostly clear and understandable, but with more serious grammatical or word use errors than in category 8
8. Almost or perfectly clear and understandable with minor grammatical or word use errors
9. Perfectly clear and understandable with no inappropriate grammar or word usage