COLUMBIA | SIPA
Center for Development Economics and Policy

CDEP-CGEG WORKING PAPER SERIES

CDEP-CGEG WP No. 42

The Selection and Causal Effects of Work Incentives on Labor Productivity: Evidence from a Two-stage Randomized Controlled Trail in Malawi

Hyuncheol Bryant Kim, Seonghoon Kim, and Thomas T. Kim

August 2017

COLUMBIA | SIPA
Center on Global Economic Governance

The Selection and Causal Effects of Work Incentives on Labor Productivity: Evidence from a Two-stage Randomized Controlled Trial in Malawi*

Hyuncheol Bryant Kim†, Seonghoon Kim‡, and Thomas T. Kim§

22 July 2017

Abstract

Different work incentives may affect labor productivity differently. We implement a two-stage field experiment to measure effects of career and wage incentives on labor productivity through self-selection and causal effect channels. First, workers were hired with either career or wage incentives. After employment, a random half of workers with career incentives received wage incentives and a random half of workers with wage incentives received career incentives. We find that career incentives attract higher-performing workers than wage incentives but do not increase productivity conditional on selection. Instead, wage incentives causally increase productivity for existing workers. Observable characteristics are limited in explaining the selection effect.

Keywords: Career Incentive, Wage Incentive, Internship, Self-selection, Labor Productivity
JEL Classification: J30, O15, M52

* We are grateful to the following staff members of Africa Future Foundation for their excellent field assistance: Narshil Choi, Jungeun Kim, Seungchul Lee, Hanyoun So, and Gi Sun Yang. In addition, we thank Jim Berry, Syngjoo Choi, Andrew Foster, Dan Hamermesh, Guojun He, Kohei Kawaguchi, Asim Khwaja, Etienne Lalé, Kevin Lang, Suejin Lee, Pauline Leung, Zhuan Pei, Cristian Pop-Eleches, Victoria Prowse, Imran Rasul, Nick Sanders, Slesh Shrestha, and Armand Sim as well as seminar participants at Cornell University, Hitotsubashi University, Hanyang University, Korea Development Institute, National University of Singapore, Singapore Management University, Seoul National University, NEUDC 2016, SJE International Conference on Human Capital and Economic Development, First IZA Junior/Senior Labor Symposium, IZA/OECD/World Bank/UCW Workshop on Job Quality in Post-transition, Emerging and Developing Countries, and UNU-WIDER Conference on Human Capital and Growth for their valuable comments. This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant. All errors are our own.
† hk788@cornell.edu; Department of Policy Analysis and Management, Cornell University
‡ seonghoonkim@smu.edu.sg; School of Economics, Singapore Management University
§ krutaeks@gmail.com; Department of Economics, Ohio State University

1. Introduction

Work incentives are an essential means of human resource management. For example, career incentives (tenure and promotion) and financial incentives (higher wage, cash bonus, and employee stock option) incentives are common types of incentives to achieve higher productivity. Firms often consider both types of incentives to recruit productive workers and motivate employees to exert more effort. However, it is rare to find a study that evaluates the relative importance or effectiveness of career incentives and financial incentives.

In addition, work incentives can affect labor productivity through *selection* and *causal effect* channels. Lazear (2000) shows that performance pay, as opposed to fixed hourly wages, increases labor productivity not only because it attracts more productive workers but also because it motivates workers to increase their effort levels. A better understanding of these two channels through which incentives affect labor productivity would allow firms to design optimal hiring and compensation strategies to maximize labor productivity and reduce the need for costly screening processes. However, empirical evidence on these channels is limited because it is difficult to separately isolate the productivity-enhancing effect of work incentives (*causal effect*) and the change in productivity through endogenous worker sorting (*selection effect*).[1]

In this study, we provide experimental evidence on how career and wage incentives affect labor productivity through worker selection and causal effect channels. We conduct a two-stage randomized controlled trial to separately isolate the selection and causal effects of these incentives in collaboration with Africa Future Foundation (AFF), an international non-governmental organization (NGO).[2] The career incentives we study are a future job prospect and a recommendation letter, which are typical benefits of an internship position.[3] Wage incentives in our setting are a lump-sum salary and performance-related bonus payment. While a promotion and a pay raise often occur jointly in the workplace, we study career and financial incentives as distinct components of work incentives.

Our experimental design was implemented in the context of a recruitment drive of census enumerators. To hire enumerators for a population census of Chimutu, a rural district in Malawi, AFF approached 440 randomly selected recent high school graduates from the pool of the 2011 secondary school survey participants in the AFF's project areas. They had graduated from secondary school several months before the experiment, and most of them did not have job experience. As shown in Figure 1, in the first stage, study subjects were randomly assigned to one of two groups: (i) those who received a job offer with career incentives of an internship (hereafter the *Internship* group) and (ii) those who received a job offer with wage incentives (hereafter the *Wage* group). Those assigned to the *Internship* group received an internship opportunity that

[1] The *causal* effect refers to the difference in labor productivity when incentives affect performance holding employee composition constant. The *selection* effect refers to the difference in labor productivity driven by workers' self-selection into the job.

[2] Our two-stage experimental design is similar to recent experimental studies in development economics (Karlan and Zinman, 2009; Ashraf et al., 2010; Cohen and Dupas, 2010; Beaman et al., 2015).

[3] An internship is a temporary position that can be paid or unpaid, and is distinguished from a short-term job in that it emphasizes on-the-job training for students or entry-level workers. According to a 2011 survey of the US-based National Association of Colleges and Employers, more than 50% of graduating college students had internship experiences (Nunley et al., 2016). Internship programs are also widely available in Malawi in the public, private, and NGO sectors. For example, about 20% of regular workers in AFF are hired through the internship program.

comes with (a) a potential long-term employment opportunity at AFF as a regular employee and (b) a recommendation letter specifying their relative job performance.[4] A one-time temporary work opportunity with a lump-sum wage and a bonus payment based on job performance was offered to those assigned to the *Wage* group. Individuals who accepted the job opportunity in the first stage proceeded to enumerator training and the second-stage randomization. In the second stage, a randomly selected half of the workers in the *Internship* group additionally received the same wage incentives of the *Wage* group without prior notice after completing the training. In the same manner, a randomly selected half of the workers in the *Wage* group additionally received the same internship incentives as the *Internship* group without prior notice.

As shown in Figure 1, this research design allows us to obtain two sub-groups, *Group 2* (*G2*) and *Group 3* (*G3*), which have identical incentives (both career and wage incentives) during the work, but the channels through which they were attracted to the job are different. As a result, we isolate the selection effect on labor productivity by comparing *G2* and *G3*.[5] In addition, we estimate the causal effects of career incentives on job performance by comparing *Group 1* (*G1*), who only have career incentives, and *G2*. *G1* and *G2* workers became enumerators through career incentives but only *G2* received additional wage incentives. Hence, any difference in performance between *G1* and *G2* can be interpreted as a causal effect of wage incentives estimated among the *Internship* group. Similarly, we estimate the causal effects of wage incentives on job performance by comparing *G3* and *Group 4* (*G4*) who only have wage incentives. *G3* and *G4* workers became enumerators through wage incentives but only *G3* received additional career incentives. Any difference in performance between *G3* and *G4* can be interpreted as a causal effect of career incentives estimated among the *Wage* group.

The nature of an enumerator job is multidimensional in that enumerators are expected to conduct interviews quickly and accurately. In addition, they work for an NGO that serves local communities in rural Malawi, and thus, they are expected to give good impressions to the NGO's beneficiaries. As a result, we measure job performance by the number of surveys conducted per day (survey quantity), survey error rate (survey quality), and subjective performance evaluations (SPEs) by census respondents as well as AFF supervisors.

We use three data sources: AFF's administrative data of candidates' training and enumerators' job performance; Chimutu population census data; and AFF's surveys of the study participants, which collect rich information on observable individual characteristics. Our rich and high-frequency data enable us to estimate selection and causal effects of work incentives precisely and to explain its potential mechanism.

Out of 440 randomly selected recent male high school graduates whom AFF approached for the baseline survey of this study without prior notice of job opportunity, 362 (82.3%)

[4] An entry-level regular position (enumerator or data entry clerk) at AFF has career advancement prospects that lead to more advanced positions, such as head enumerator, junior project assistant, senior project assistant, and project manager. AFF did not explicitly state the actual probability of hiring for the *Internship* group. We acknowledge that changing probabilities of hiring after the internship might affect effort levels, but we do not compare different levels of the same incentive, but rather two different types of incentives.

[5] Due to the nature of our experimental design, the selection effect of either career incentives or wage incentives can be evaluated against the other type of incentives.

participated in the baseline survey.[6] None of them previously worked for AFF and only 8.8% were currently working as they had graduated from their secondary school just 5 months previously. Of 176 participants assigned to the *Wage* group, 74 (42.0%) accepted a job offer by joining the training session. Of 186 participants assigned to the *Internship* group, 74 (39.8%) took up the job offer. The fact that the take-up rates were similar between the two groups suggests that we are comparing two distinct types of incentives whose perceived market values are similar. Out of 148 trainees, 11 dropped out from the training. Thus, 137 enumerators worked in the field for an average 18 days interviewing 21,561 households.[7]

We reach three main conclusions using high frequency and quality data on multidimensional labor productivity. First, we find that career incentives, compared to wage incentives, attract workers with higher labor productivity in terms of survey quality and quantity as well as SPEs by census respondents through the self-selection mechanism. However, we find that the causal effects of career incentives are limited in improving labor productivity, while wage incentives causally increase labor productivity. As a result, job performance measured by survey quality and quantity as well as SPEs by respondents is highest among *G2* enumerators who were hired through the career incentive channel and additionally received wage incentives. Third, we find that observable individual characteristics are limited in explaining the estimated selection effect, suggesting a limitation of screening based on observable characteristics and a need for a self-selection mechanism that can attract productive workers with desirable characteristics.

We contribute to the existing literature by providing real-world evidence on both career and wage incentives for labor productivity. There is a large body of existing experimental and quasi-experimental research on the role of incentives in job performance. However, to the best of our knowledge, our study is unique in that it is the first of its kind to compare career and financial incentives directly in the same setting. Even though a better understanding of these work incentives would allow employers to design an optimal mixture of compensation strategies, there is no research that directly compares career and financial incentives in the same setting. In addition, we carefully separate the selection and causal effect channels of these work incentives through two-stage randomization, which sheds light on the design of a better compensation strategy. In addition, existing research to separately measure the selection and causal effects of work incentive is also rare. Last but not least, this study provides the first empirical evidence that credibly examines the selection and causal effects of an internship on job performance.

Our study is related to several strands of the literature. First, it is closely related to the literature that estimates the selection and causal effects of incentives on job performance. Lazear (2000) separately isolates worker selection and causal effects of a financial incentive using non-experimental panel data on job performance from a large manufacturing factory in the U.S. which changed from a fixed salary to a piece rate. Guiteras and Jack (2015) separately isolate worker selection and causal effects of a financial incentive by experimentally varying the wage level of

[6] There were 536 eligible study subjects who were male and fresh high school graduates in the AFF's project areas. Out of the 536, AFF provided job offers to a randomly selected group of 440. However, all 536 subjects were invited to participate in the same baseline survey. Individual characteristics and the balance between the two groups (440 vs. 96) are shown in Table A.1.

[7] Throughout this paper, *target study participants* refer to the 440 individuals who were invited to participate in the baseline survey, *study participants* refer to the 362 individuals who participated in the baseline survey; *trainees* refer to the 148 individuals who joined the training; and *enumerators* refer to the 137 individuals who worked in the field.

daily workers and identifying their reservation wages in rural Malawi.[8] Gagliarducci and Nannicini (2013) also separately identify the selection and causal effects of a financial incentive on the performance of politicians by exploiting policies that discontinuously change salaries by population size and a term limit on re-election. All of these studies shed light on the relative importance of the selection and causal effects of incentives on labor productivity.

Some studies focus on the effects of incentives at the recruitment stage on worker selection and job performance (i.e., selection effect). Dohmen and Falk (2011) show that sorting of workers largely explains higher labor productivity under a variable-payment scheme compared to a fixed-payment scheme in a laboratory experiment setting with payment based on a simple math test. In a natural field experiment, Dal Bó et al. (2013) show that a higher wage attracts more qualified applicants without the cost of losing workers with strong public service motivation in the context of a recruitment drive of Mexico's public sector workers. They present the difference in observable baseline characteristics but lack data on actual job performance. Ashraf et al. (2016) similarly show that salient career incentives attract more productive workers without discouraging those with pro-social preferences from applying for a job in the context of a recruitment drive of community health workers in Zambia. On the other hand, Deserranno (2016) finds that the expectation of a higher salary for a newly created public health worker position discourages job applications from candidates with pro-social preferences in Uganda.

In addition, another strand of the literature focuses on causal effects of work incentives.[9] The literature on causal effects of work incentives focuses mainly on financial incentives. For example, Gneezy and List (2006) empirically test the gift exchange theory originally developed by Akerlof (1984) and show that workers exert more efforts when they receive a financial incentive ("gift") from their employers. Shearer (2004) presents experimental evidence from Canadian tree planters that piece rates induce more effort than fixed wages. In addition, several studies estimate the causal effects of performance pay on productivity of agents in the public service or NGO (Glewwe et al., 2010; Duflo et al., 2012; Fryer, 2013; Ashraf et al., 2014).

Lastly, our study is related to the literature on internships. Most existing studies on internships are mainly descriptive (Brooks et al., 1995; D'Abate et al., 2009; Liu et al., 2014). A rare exception is Nunley et al. (2016), which sends out fake résumés with randomly changed characteristics of applicants. They find that a résumé with internship experience receives 14% more callbacks from potential employers but a major limitation of the résumé audit study is that it cannot analyze job performance.

The remainder of the paper is structured as follows. Section 2 outlines the research context and design. Section 3 describes the data and reports sample statistics. Section 4 presents the main results on labor productivity and discusses the findings. Section 5 presents the results of additional analysis including the impact of supervisor visits and short-term effects of enumerator job

[8] They use the Becker–DeGroot–Marschak (BDM) mechanism (Becker et al., 1964) to identify the reservation wage. By comparing the work performance of workers with an identical reservation wage but different actual wages for bean-sorting work, they isolate the causal effect of a higher wage on labor productivity. However, Bohm et al. (1997) and Horowitz (2006) discuss that BDM may not be incentive compatible in practice, and thus, could bias measuring reservation prices. Moreover, since revealing a reservation wage is not part of the ordinary employment process, job applicants might not be comfortable revealing their true reservation wage. Our research design separately isolates the selection and causal effects of work incentives without relying on the indirect inference of an unobserved worker characteristic.

[9] Oyer and Schaefer (2011) and Bandiera et al. (2011) provide an excellent survey of the literature.

experience. Section 6 concludes.

2. Research Context and Design

2.1. Research Context

Malawi is one of the least developed countries in the world with GDP per capita in 2015 of US$382 (World Bank, 2016). Among 20–29 years old males, 19.6% completed secondary school education according to the 2010 Malawi Demographic and Health Survey. Employment in the official sector is 11% and the median monthly income is US$28.8 (13,420 MWK) (National Statistical Office of Malawi, 2014).[10]

AFF conducted a district-wide population census of Chimutu, a rural district located outside of the capital city of Malawi, to collect demographic and socio-economic information of households in January 2015. Chimutu district consists of 52 smaller catchment areas and there are about 94,000 people in about 24,000 households. AFF planned to complete a census within a month by hiring more than 130 enumerators.

The enumerator position at AFF could be an attractive starting job for entry-level young workers because it offers a competitive salary and confers career-advancing incentives. For example, AFF's many regular staff members were initially recruited as enumerators. A primary role of the census enumerators was to interview household heads to collect basic demographic, socioeconomic, and health information. During the census period, enumerators stayed at a house in the assigned catchment area rented by AFF. Enumerators were asked to survey at least eight households per day. They worked alone in the field but supervisor teams visited them without prior notice. Since enumerators interviewed many residents in remote villages to collect a variety of personal and complex information, the job required both cognitive and interpersonal skills as well as physical endurance.

Study participants to whom AFF offered an enumerator job were drawn from the sample who participated in the 2011 secondary school student survey of 7,971 secondary school students in four districts in Malawi, including Chimutu. This 2011 survey was a baseline survey for the AFF's previous research program that randomly provided HIV/AIDS education, male circumcision, and financial support for female education in their catchment areas.[11] Of the 536 males who participated in the 2011 secondary school survey and graduated from secondary schools in July 2014, AFF randomly selected 440 as target study participants. The 440 target study participants participated in the survey (i.e., the baseline survey of this study) without notice of a potential job offer. This sample recruitment approach allowed AFF to hire workers familiar with the census area. The NGO considered only males due to security concerns in the field. In addition,

[10] MWK denotes Malawi Kwacha. As of January 1, 2015, US$1 was equivalent to 466 MWK. Throughout the paper, we use this as the currency exchange rate.

[11] AFF's catchment area includes the four districts are Chimutu, Chitukula, Tsbango, and Kalumb. For details of AFF programs, see Data Appendix A.4.

the NGO required secondary school graduation as proof of minimum cognitive skill requirements.

Outside options for the enumerator job are other formal sector jobs, household farming, and repeating secondary school. For example, at the time of the baseline survey, 4.7% of our study participants were working for pay in formal sectors, 4.3% were working for their family business (mainly farming), and 15.8% were attending vocational schools or colleges.[12] About 60% were actively searching for jobs.

In a study that involves job applications or job offers, there might be concern that individual characteristics of job applicants would be systematically different from those of non-applicants. Applicants could be more likely to possess the necessary skills, have better access to the information (at least for a job vacancy), and/or be less likely to be happy with their existing positions if they are currently working for another employer. Hence, the estimation of selection effects of any work incentives is inherently local to job applicants. Therefore, our sample recruitment strategy has the following advantages. First, we observe the population of a young cohort whose members are potentially interested in a job opportunity in the local labor market, contrary to existing studies that observe only job applicants. This sampling feature allows us to have a better sense of the external validity of our findings. Second, approaching those who just graduated from secondary school is relevant to an internship, which mainly targets young and entry-level workers.

2.2. Experimental Design

As discussed in the introduction, we implement a two-stage randomized trial to separately identify selection and causal effects of career and wage incentives. As shown in Figure 1, in the first stage, study participants were randomly assigned to the *Wage* group or the *Internship* group.[13] In the second stage, of those who accepted AFF's conditional job offer, a randomly selected half of the *Internship* group and the *Wage* group additionally received career and wage incentives, respectively.[14]

This two-stage randomization creates four study groups: *G1* received career incentives only while *G2* received both career and wage incentives. The incentive scheme of *G1* and *G2* mimics unpaid and paid internship arrangements, respectively. Similarly, *G4* received wage incentives only while *G3* received both career and wage incentives. Regardless of the study groups, the enumerators performed the same tasks during the census.

Comparing *G2* and *G3* isolates the selection effect of the career incentives evaluated against the wage incentives because they have identical incentives at work, but the channels

[12] Repeating the last year of the secondary school to improve their secondary school graduation exam scores is not uncommon in Malawi because the graduation exam score plays an important role in signaling the job candidate's ability. Thus, it could be an important alternative to working for the AFF as a short-term enumerator. In the baseline survey, 1.3% were repeating the secondary school, even though they have already graduated from secondary school. This number increases to 27% in the follow-up survey implemented one year after the experiment.

[13] Note that participants were randomly pre-assigned to either one of the two groups before they participated in the baseline survey. At the end of the survey, they were given a job offer according to the pre-assigned group. The job offers were given regardless of participants' employment and schooling status at the time of the survey.

[14] A job offer was valid conditional on the successful completion of the training. For the sake of simplicity, we refer to a conditional job offer simply as a job offer henceforth.

through which they were recruited were different.[15] In addition, comparing *G1* and *G2* isolates the causal effects of the wage incentives among those who took up the job offer with career incentives in the first stage, and comparing *G3* and *G4* isolates the causal effects of the career incentives among those who took up the job offer with wage incentives in the first stage. The differences in productivity between *G1* and *G4* can be interpreted as the combination of the selection and causal effects.

2.2.1. Recruitment and baseline survey

We describe the research stages in chronological order as shown in Table 1. As stated in the introduction and earlier in this section, AFF invited 440 males who met the eligibility criteria for the baseline survey (Row A) and 362 (82.3%) participated in the baseline survey (Row B).[16] In addition, AFF invited study subjects soon after the census was completed between April and June 2015 to measure time and risk preferences and rational decision-making ability.[17] We further discuss the data collected from these surveys in Section 3.

To minimize unexpected peer effects among workers with different incentives, the baseline survey was conducted separately for the *Internship* group and the *Wage* group. The training was also provided separately. Even though we cannot fully exclude possible interaction between study groups, it is unlikely that study participants in different study group actively interact with each other, because only about 14 people per school (440 from 31 schools) were invited for the study.

2.2.2. First-stage randomization

AFF supervisors explained the details of the enumerator job at the end of the baseline survey. As described in Section 1, target study participants received a job offer with different work incentives.

Of 220 target study participants assigned to the *Wage* group, 176 (80.0%) showed up for the baseline survey (Row B) and were given a short-term job offer, each with a fixed salary of 10,000 MWK (US$21.5) for 20 working days and performance pay of 500 MWK (US$1.1) for every extra 8 households after the first 160 households.

Of 220 target study participants assigned to the *Internship* group, 186 (84.5%) showed up for the baseline survey (Row B) and were given a job offer with career incentives which consists of a recommendation letter and the prospect of working at AFF as a regular staff member. We randomized the first-stage incentives in advance and then invited different first-stage groups on different dates. In addition, study participants were not aware of the other type of incentives when they received an offer.

The base wage of 10,000 MWK (US$21.5) was competitive for young workers who had just graduated from secondary schools because the median monthly salary of secondary school graduates in 2013 was 12,000 MWK (US$25.8), according to the Malawi Labor Force Survey (NSO, 2014). In addition, the prospect of a regular entry-level staff position at AFF whose entry-

[15] The comparison of *G2* and *G3* can be interpreted as the selection effect of the wage incentives evaluated against the career incentives, but for the sake of convenience, we mainly focus on the career incentives.

[16] Those who did not participate in the survey were unreachable (45%), or refused to participate (13%), or could not participate in the survey because they were at school (32%) or working (10%).

[17] This survey was conducted to measure time and risk preferences and rational decision-making ability after the census was completed under the assumption that these measures are not affected by our interventions. 334 (76%) out of 440 target study participants participated in the survey.

level monthly salary is 26,000 MWK (US$55.8) could be attractive.[18] AFF notified the *Internship* group that there would be a chance of a long-term contract, without specifying the precise probability, depending on job performance during the contract period and AFF's job vacancies. Working as an intern without knowing the exact probability of hiring is close to the general internship setting. The recommendation letter was signed jointly by the director of AFF and the head of the Chimutu district. Lastly, one-time transportation support, on average about 1,500 MWK (US$3.2), was given to both *Wage* and *Internship* groups depending on the distance from the worker's home and the dispatched village.

2.2.3. Training

Those taking up the job offer were required to participate in the 1-week training program in January 2015. It was designed to equip trainees with the necessary skills and knowledge for the census work. The training outcomes were measured by a quiz score and the proportion of erroneous entries in a practice survey.

Out of the 186 study participants in the *Internship* group, 74 (39.8%) participated in the 1-week training session, as did 74 out of 176 (42%) study participants in the *Wage* group (Row C).[19] The job take-up rates (training participation rates) between the *Internship* group and the *Wage* group were not statistically different, which could imply that the perceived value of career and wage incentives evaluated by the offer take-up rate were similar. However, 11 trainees from the *Internship* group were not hired because of low training performance, while no one failed from the *Wage* group (Row D). As a result, in total, 137 enumerators were finally hired, 63 of which were from the *Internship* group and 74 from the *Wage group* (Row E). Therefore, we do not observe the job performance of 11 trainees from the *Internship* group who failed the training requirement. We discuss this further in Section 4.

2.2.4. Second-stage randomization

Second-stage randomization was conducted during the training and the randomization results were announced after the training completion and before the dispatch to the catchment area. The wage incentives were given to a randomly selected half of the *Internship* group. Similarly, the career incentives were given to a randomly selected half of the *Wage* group. Trainees were not aware of additional incentives during the training. No one refused to accept the additional incentives, which implies that the composition of worker characteristics between *G1* and *G2* and between *G3* and *G4* remain the same.

Right after the second-stage randomization, they signed the employment contract, which describes the incentives and benefits of the position, as shown in Figures A.1, A.2, and A.3. For example, the employment contract of *G1* explicitly states that enumerators will not be given any financial compensation and will be provided with a recommendation letter and a potential job opportunity based on their performance. It may help to prevent enumerators believing that additional changes in the work condition may happen in the future. It is also noteworthy that the

[18] Those promoted to a project manager position at AFF were paid between US$100 and US$160 (46,600 MWK and 74,560 MWK) per month during the study time.

[19] The take-up rate seems low given the low employment rate in our setting. However, the low take-up rate is plausible, because household farming is one of the important outside options.

contract explains main job performance measures: survey quality, survey quantity, and SPEs.

2.2.5 Census and post-enumeration survey

Enumerators were dispatched to 52 catchment areas in January 2015. Enumerators were randomly assigned to catchment areas stratified by population and land size, and they worked independently. Enumerators in the same catchment area have the same incentives to prevent unexpected peer effects. In addition, enumerators were not assigned to areas from which they originally came, as locality could affect their performance.

The specifics of the census questionnaire include a variety of individual- and household-specific characteristics such as demographics, wealth, employment and income, and health. It took about 25 minutes on average to interview a household. In total, enumerators surveyed 21,561 households during the contract period.

AFF supervisor teams visited enumerators to monitor and guide enumeration work on randomly selected dates without prior notice. Supervisors are AFF's regular staff members, each of whom has at least 3 years of experience conducting field surveys. Each supervisor team consists of two supervisors.

AFF randomly assigned five supervisor teams to 52 catchment areas for their visits. Most enumerators met a supervisor team at least once during the census period; 37% of the enumerators met supervisors twice and the remaining 60% met supervisors once.[20] Enumerators were aware of supervisor visits, but did not know the exact date. Supervisors joined each enumerator for interviews of about three households, addressed common errors, and provided overall comments at the end of the visit.

Shortly after the completion of the census (March–May 2015), AFF conducted a post enumeration survey (PES) to correct errors found in the original census interview and to find omitted households and measure SPE by revisiting all households in Chimutu. AFF announced a PES plan to evaluate the performance before the field dispatch to prevent enumerators from outright cheating or fabricating census interview sheets. PES was conducted by some enumerators from *G1*, *G2*, and *G3* (those with career incentives) who were temporarily hired to conduct PES.[21]

As stated in the employment contract, AFF provided recommendation letters to the enumerators with career incentives (*G1*, *G2*, and *G3*) in May 2015. The letter specified the job description of an enumerator and his relative job performance measured by survey quality, survey quantity, and SPEs.[22]

2.2.6. Follow-up surveys

AFF conducted follow-up surveys with the study participants after 1 and 2 years from the

[20] A supervisor team failed to visit one catchment area where four enumerators were assigned because the enumeration work was completed right before the randomly selected first visit took place.

[21] Hiring enumerators as regular staff members required the careful calculation of job performance after the completion of the census, which requires a couple of months. Meanwhile, AFF hired 43 PES enumerators among 98 census enumerators with career incentives (*G1*, *G2*, and *G3*) on a temporary basis (2–3 months) through a simple performance evaluation based on SPE by supervisors and error rates measured from five randomly selected surveys.

[22] For example, if an enumerator has higher job performance than the average, the letter specifies a very strong recommendation. If an enumerator has performance below the average, the letter specifies a somewhat lukewarm recommendation.

experiment (May 2016 and June 2017) over the telephone. The telephone survey was administered to investigate whether the short-term job experience affected participants' future labor market outcomes.

3. Data

We use data from various sources, including baseline and follow-up surveys, AFF's administrative data on training and job performance, and the Chimutu population census.[23]

Data from the surveys include the following. First, we use data from the 2011 secondary school student survey. It contains rich information on a variety of areas covering demographics, socioeconomic status, health, and cognitive ability. Second, we use data from the 2014 baseline survey, which collects information on demographics, education, employment history, cognitive abilities, non-cognitive traits, and HIV/AIDS related outcomes. We measure cognitive ability by a cognitive ability index, defined as the average z-score of the Raven's matrices test score, the math and English scores of the 2014 Malawi School Certificate of Education (MSCE) test, and the verbal and clerical ability test scores of the O*NET test, following the approach of Kling et al. (2007).[24] Non-cognitive traits include self-esteem, intrinsic motivation, extrinsic motivation, and the Big Five personality test (extraversion, openness, conscientiousness, agreeableness, and neuroticism). Third, the additional baseline survey conducted in April–June 2015 collected data on risk and time preferences and rational decision-making ability using the tests recently developed by Choi et al (2014).[25] Finally, we use data from the follow-up surveys to measure labor market outcomes 1 and 2 years after the short-term employment.

Columns (2) and (3) of Table 2 present the baseline characteristics of each study group. The results of first- and second-stage randomization balance are presented in Columns (4), (5), and (6), respectively. Panel A represents individual baseline characteristics of study participants. Study participants are about 20 years old and only 9% work in the official sector reflecting weak labor demand in Malawi.[26] Column (4) compares the *Internship* and *Wage* groups. Columns (5) and (6) compare *G1* and *G2*, and *G3* and *G4*, respectively. Data Appendix A.1 provides the specific definition of the variables presented in Panel A. Panel B represents the catchment area characteristics where enumerators were dispatched. It shows that the number of households per

[23] We calculate the average characteristics of the catchment area based on the census data. These characteristics were used as the control variables in the main regression analysis.

[24] AFF had access to the administrative MSCE score data via the cooperation of the Ministry of Education of the Republic of Malawi. Raven's progressive matrices test is a non-verbal test of thinking and observation skills. The MSCE is a test that all Malawian students must take to graduate from secondary school. The MSCE score we use is a standardized test score of mathematics and English, which are mandatory subjects of the test. The O*NET® test is a tool for career exploration developed through the U.S. Department of Labor. We use verbal and clerical perception ability test scores of O*NET®, which are directly related to enumerator job characteristics. Data Appendix A.1 provides the definitions of cognitive ability measures.

[25] As explained in Subsection 2.2.1, risk and time preferences, and rational decision-making ability were measured after the census was completed. We included these measures in the randomization balance test under the assumption that these traits were not affected by our experiment. Data Appendix A.1 provides the details of how we measure them.

[26] The employment rate of baseline survey non-participants is similar. We reached non-participants via phone calls and 9.7% of them told us that they did not attend because they were working.

catchment area is about 213 for the *Internship* group and 239 for the *Wage* group.[27]

The results confirm that the study groups are well balanced: the proportion of statistically significant mean difference at the 10% significance level is 2 out of 27 (7.4%) in Column (4), 3 out 27 (11.1%) in Column (5), and 4 out of 27 (14.8%) in Column (6).[28]

In addition, we examine whether the baseline survey participants and nonparticipants are systematically different. Table A.2 shows that they are not statistically different from each other in most dimensions except for the household asset score.

Training outcomes are measured by a quiz score and the proportion of erroneous entries in a practice survey.[29] The quiz tested specific knowledge on the census details.[30] Main job performance during the census is measured based on the census survey data in three dimensions: 1) quantity, 2) quality, and 3) SPEs. Quantity is measured by the number of households surveyed by each enumerator per day. Quality is measured by the proportion of systematically inconsistent or incorrect entries in the census questionnaire specific to each household surveyed.[31] SPEs are measured by census respondents. During the PES, a census respondent was asked to evaluate how carefully the enumerator explained the questions.[32] In addition to the three main types of job performance, another SPE was measured by AFF supervisors: After the completion of the census, 12 supervisors jointly evaluated the overall work attitude of each enumerator. Enumerators were aware that job performance evaluations because the employment contract states that job performance is measured by three main types of job performance (survey quantity and quality as well as SPE by the respondents).

4. Main Results

4.1. Job Offer Take-up

Column (1) of Table 3 confirms that the job offer take-up rates between the *Internship* and the *Wage* groups are not different. This result suggests that the average perceived market values

[27] See the note to Table 2 for the specific definition of the variables used in Panel B.

[28] The number of siblings, the only unbalanced individual variable in Column (4), eligibility for AFF's past interventions and catchment area characteristics controls are included in all specifications of the main analysis.

[29] The purpose of the practice survey was to practice interview skills before enumerators were dispatched to the field. The practice survey performance was evaluated as follows. First, we randomly matched two trainees. Each trainee in a randomly assigned pair received a pre-filled census questionnaire sheet and a blank survey questionnaire sheet. Then, one trainee interviewed the other matched trainee in the same pair and the latter answered based on the assigned survey sheet. There were two different types of pre-filled questionnaire sheets with different hypothetical household information. Thus, trainees in the same pair acted as if they were two different households. Each trainee in every pair conducted this practice survey by changing roles. After conducting practice survey sessions, supervisors collected the survey sheets and calculated the error rate.

[30] The quiz consists of 12 questions, a mixture of open-ended and true/false type questions. The full text of the quiz is presented in Figure A.4.

[31] For example, if a respondent has a child, the information about her child should be filled in. If not, it is counted as an error. The Data Appendix provides the details about how we calculate the survey error rate.

[32] The question asked was "Whenever you were confused or could not understand the meaning of any question, did the enumerator carefully explain the meaning of the questions to you?" Enumerators who conducted the PES sought to interview the original census respondent particularly for this question. However, the original respondent was not always available, and thus we can analyze SPE by census respondents only when the census respondent and the PES respondent are identical. In addition, we conducted the test of equality between households in which the census respondent and the PES respondent are identical and households of which they are different. The probabilities that an original census respondent is a PES respondent are 77%, 77%, 83%, and 82% for *G1*, *G2*, *G3*, and *G4*, respectively. These rates are significantly different. Hence, the interpretation of the SPE analysis by respondents should be taken with caution.

of a job offer with career incentives and a job offer with wage incentives are similar. Even though the take-up rates are similar across the two groups, it is possible that the composition of job takers between the two groups could be different if different incentives attract workers with different observable and unobservable characteristics. We test multidimensional sorting discussed in Dohmen and Falk (2011) by exploring whether career and wage incentives attract those with different observable characteristics. Columns (2) to (18) of Table 3 show the regression results of the following equation:

$$Accept_i = \alpha + \delta \cdot Internship_i + \lambda \cdot Trait_i + \varphi \cdot Internship_i \cdot Trait_i + \epsilon_i \quad (1)$$

$Accept_i$ is a binary indicator that equals 1 if individual *i* accepted a job offer, and 0 otherwise. $Internship_i$ is a binary indicator if individual *i* belongs to the *Internship* group and the omitted category is the *Wage* group. *Trait* is an individual characteristic variable that we evaluate one by one. ϵ_i is an error term. We test whether career incentives attract workers differently over a variety of individual characteristics including demographic and socioeconomic characteristics, cognitive ability index, and non-cognitive traits.

Our coefficient of interest is φ, which captures whether there is differential take-up of a job offer between the *Internship* group and the *Wage* group by individual traits. We find that none of the estimates of φ across individual traits is statistically significant at the 5% level. This finding implies that observable characteristics are not likely to predict self-selection. Table A.3 provides additional evidence on self-selection by comparing the observable characteristics of job offer takers between the *Internship* group and the *Wage* group. The results in Table A.3 confirm the results in Table 3 that the two groups are not systematically different in terms of both statistical and economic significance (the p-value of the joint F-test of the equality between the two groups is 0.53). We find only one statistically significant result (extroversion) out of 21 traits.[33] In addition, only 4 out of 21 variables have mean differences larger than 0.2 standard deviations.[34]

We acknowledge that study participants could have responded to the self-reported non-cognitive tests in a way they believe desirable from the perspective of a potential employer. This possibility might explain the little difference in terms of non-cognitive traits between the two groups. This is consistent with the real world in which job seekers are not able to manipulate test scores (cognitive ability) in a pre-employment test but might try to respond to a personality test in a way in which they have a desirable non-cognitive skill. However, participants did not know about the possibility of receiving a job offer at the time of the survey.

The absence of systematic differences in observable characteristics does not necessarily mean that unobservable characteristics, training outcomes, and job performance would not be different if some of the unobservable characteristics were to affect training outcomes and job performance.

[33] This finding should be interpreted with caution due to the concern about multiple hypothesis testing.

[34] One may concern on statistical power due to relatively small sample size (N=148). However, for most variables (X out of 21), we are able to detect 10 percent differences between two groups. For example, Column 2 of Table 4 shows we are powered to detect age differences between two groups as far as it is bigger than 0.07 (=0.037*1.96) years, which is a 0.36% change (0.07/20.4)*100. Nonetheless, we cannot fully rule out the possibility that we may not detect somewhat large differences between two groups; therefore, the results should be interpreted with this caveat.

4.2. Training Outcomes

Even though we do not find any differences in observable characteristics between job takers of the two groups, we might find a difference in training outcomes if career and wage incentives attract people with different unobservable characteristics. Panel A of Figure 2 displays the kernel density estimates of the training outcomes measured by the quiz score and the practice survey error rate. Table 4 shows the corresponding results from the following specification:

$$Training_i = \alpha + \beta \cdot Internship_i + \omega_i \quad (2)$$

where $Training_i$ is the training outcomes such as practice survey error rate and quiz score for individual i.[35] For the practice survey error rate regression, we control for a practice survey type (and pair–fixed effect) in the regression.[36]

Panel A of Figure 2 shows that the *Wage* group performs better than the *Internship* group in terms of both quiz score and practice survey error rate. Panel A of Table 4 provides corresponding results from the regression. It confirms that quiz score of the *Internship* group trainees are 2.0 points (23.8%) lower than the *Wage* group trainees (Column (1)). Similarly, the survey error rate is 10.4 percentage points (38.2%) higher among the *Internship* group trainees than the *Wage* group trainees (Column (3)).

At the end of the training, AFF disqualified 11 trainees who did not meet the minimum qualification requirement. As the regression results above indicate, the *Internship* group performed worse than the *Wage* group. All dropouts (11 trainees) came from the *Internship* group and all those in the *Wage* group passed training.

Panel B of Table 4 presents the training outcomes of enumerators dispatched to the field by excluding the 11 training failures. The regression results between the two panels are qualitatively similar, but the magnitude of the coefficient estimates is larger in Panel A than in Panel B because those who failed training are all from the *Internship* group.

The specification used in Columns (2) and (4) is to test whether individual observable characteristics can explain the differences in the training outcomes between the two groups. In addition to the number of siblings, a binary indicator for the survey questionnaire type, and past eligibility status of AFF's HIV/AIDS program included in Columns (1) and (3), the specification also includes age, household asset score, cognitive ability index, and non-cognitive traits, such as self-esteem, intrinsic and extrinsic motivation, and Big 5 personality scales. The specification of column (5) further includes the practice survey pair fixed-effect variables. We find similar coefficient estimates between Columns (1) and (2), and between Columns (3) and (4). For example, observable characteristics explain the difference in quiz score and error rate only by 4.0% (=(2.01-1.93)/2.01) and 14.4% (=(.104-.089)/.104), respectively. This finding implies that observable characteristics are somewhat limited in explaining the difference in the training outcomes.

In summary, we find that those attracted by a job offer that comes with wage incentives

[35] We control for number of siblings, which is not balanced in the baseline, and binary indicator variables of the past eligibility status of AFF's programs.

[36] As explained in footnote 28, a practice survey was conducted between trainee pairs.

outperformed those attracted by a job offer that comes with career incentives in the training. This difference could be because workers with different characteristics might be attracted by different work incentives, thereby creating the difference in the training outcomes (selection effect). Furthermore, those in the *Internship* group have an incentive to exert more effort than the *Wage* group due to the future job prospect of the career incentives (causal effect). That is, in the absence of this causal effect of career incentives, the difference could have become larger. In this sense, this finding can be interpreted as a (negative) selection effect of career incentives dominating a (positive) causal effect of career incentives, if one exists. Thus, the observed difference in training performance could be the lower bound of the selection effect (in magnitude). However, we are not able to disentangle the two effects.

4.3. Selection effect of career incentives on labor productivity

In this subsection, we examine the selection effect of career incentives evaluated against wage incentives on job performance. As previously discussed, *G2* and *G3* have the same incentives at work but the channels by which they were recruited are different. Therefore, we argue that differences in performance are driven by the *selection effect*.

An implicit identifying assumption in this analysis is that perceived work incentives of enumerators in *G2* and *G3* are identical even though the sequences of the provision of career and wage incentives are different. The different sequence could form different perceived expectations of the level of compensation package. If this were the case, it could affect the feeling about what actually occurs, leading to different levels of work efforts, and our estimates of the selection effect would be biased as Abeler et al. (2011) discussed. However, we argue that this is unlikely: if there were such a difference in feeling, we expect that differences in job performance would become smaller over time because the difference in the feeling might diminish with time. However, Figure A.5 shows that the difference in job performance is fairly constant over time.

In addition, the difference sequence could still generate bias if it changes how the incentives are perceived by study participants. For example, those recruited with career incentives might consider the addition of wage incentives as reward for good performance during training while those recruited with wage incentives consider the addition of career incentives as a windfall gain, not reward. However, this cannot be the case, because we clearly indicated to trainees that the additional provision of incentives in the second stage would be randomly determined.

Panel B of Figure 2 suggests that *G2* performs better than *G3* in terms of survey quality and quantity, and SPE by census respondents. This finding is surprising because the *Wage* group had better training outcomes than the *Internship* group did. We test this graphical evidence formally by estimating the following equation:

$$Y_{ijklt} = \alpha + \beta \cdot G2_j + \gamma \cdot H_i + \varphi \cdot Z_k + V_{lt} + \sigma_t + \psi_{ijklt} \quad (3)$$

where Y_{ijklt} is job performance measured in the survey collected from household *i* by enumerator *j* whose supervisor is *l*, in catchment area *k*, surveyed on the *t-th* work day. $G2_j$ is 1 if enumerator *j*

belongs to *G2* and 0 if he belongs to *G3*. H_i is a vector of respondents' household characteristics and Z_k is a vector of catchment area characteristics.[37] σ_t is the survey date fixed effect. V_{lt} is the supervisor team-specific post-visit effect.[38] Standard errors are clustered at the enumerator level. For the dependent variable, we use survey quality measured by the survey error rate ($Error_{ijktl}$), survey quantity measured by the number of surveys per day ($Survey_{jktl}$), SPE by census respondents ($SPErespondent_{ijktl}$), and SPE by supervisors ($SPEsupervisor_{jk}$).[39]

Panel A of Table 5 presents the regression results from equation (3). It shows that *G2* outperforms *G3* in three main types of job performance even though *G3* outperforms *G2* during the training.[40] For example, the error rate is 2.0 percentage points (26%) lower in *G2* than *G3*, as shown in Column (1). The survey quantity of *G2* is higher than that of *G3* by 1.48 households per day (13.8%), as shown in Column (4). In addition, *G2* has a 37% higher SPE score by the respondents (Column (7)). However, the SPE score by supervisors is higher in *G3* than in *G2*.

To assess how much observable individual characteristics and training performance can explain the selection effect estimated in Columns (1), (4), (7), and (10), we additionally control for demographic and socioeconomic status, cognitive ability, and non-cognitive traits (Columns (2), (5), (8), and (11)) as well as training performance (Columns (3), (6), (9), and (12)). As shown in Columns (2), (5), (8), and (11), observable individual characteristics explain little in selection effect on survey quality and quantity. For example, the inclusion of observed individual characteristics does not explain the estimated selection effect of an internship on survey quality at all and it explains survey quantity only by 4.7% (=(1.48-1.41)/1.48).[41] Additionally controlling for training performance is still limited in explaining the selection effect. This result is consistent with the facts that observable characteristics of job takers between the *Internship* and the *Wage* group are not different, and that the *Internship* group whose training outcomes were better than the *Wage* group outperformed the Wage group for on-the-job performance.

In Table A.4, we report additional performance outcomes to decompose the objective performance outcomes (survey quality and quantity). In terms of survey quality, we decompose the types of errors into incorrectly entered entries (e.g., filling in 179 for a person's age) and incorrectly missing entries (e.g., a child is present in the household but his/her age is missing). To better understand how survey quantity changes, we conduct regression analysis on three time–use variables such as total work hours per day, average survey time per household, and intermission time between surveys.[42] Work hours per day are the difference between the beginning time of the first survey and the end time of the last survey of the day. Survey time per household is the length of each survey. Intermission time is defined as the difference between the beginning time of a

[37] Respondent's household characteristics include the fixed effect for family size. Catchment area characteristics include the total number of households, asset score, birth rate, malaria incidence, rate of birth with the assistance of a health professional, and death rate.

[38] $V_{lt} = \eta_1 I(t>First) + \eta_2 I(t>Second)$ where First and Second are the dates of supervisor team *l*'s first and second visits, respectively, to enumerator *j*.

[39] We do not control for σ_t and V_{lt} when we analyze *SPE(supervisor)* because it does not vary over time and catchment area.

[40] See Figure A.6 for the training outcomes of each study group including *G2* and *G3*. Note that the second stage randomization groups (*G1–G4*) were announced after the training.

[41] However, the individual characteristics can explain about 13% and 27% of the selection effect on SPEs by census respondents and by supervisors, respectively.

[42] The survey beginning and end times were recorded as a part of the census questionnaire. However, there was a sizable number of missing values in the interview beginning and end times, so we imputed the missing values. See Data Appendix A.3 for the imputation process. However, the results remain similar even if we do not use the observations with imputed time values.

survey and the end time of the previous survey.

Panel A of Table A.4 indicates that the selection effect of career incentives on survey quality in Table 5 is mostly driven by the decrease in errors of incorrectly missing entries (Column (3)). In addition, the selection effect of career incentives on survey quantity mainly comes from a decrease in intermission time between surveys (Column (9)). The average intermission time for *G2* enumerators is shorter by 4.2 minutes (an 18.3% decrease) compared to *G3* enumerators. In other words, those recruited by career incentives increase the survey quantity by exerting more effort to reduce the time taken to move to the next household, but not increasing total work hours, or decreasing survey time. Again, observable enumerator characteristics do not explain differences between *G2* and *G3* much.

In summary, we find that career incentives of an internship significantly improve labor productivity through the self-selection of workers. In addition, the observable characteristics and training outcomes do not play an important role in explaining the selection effect on job performance. This result implies the importance of a self-selection mechanism during the hiring process, which can attract workers with desirable unobservable characteristics. Our findings suggest that screening via the observables by employers could be limited unless they can measure workers' productivity directly. Pre-employment screening for highly productive employees could be particularly difficult when hiring entry-level workers who have no record of accomplishment of past job history or credentials to verify their unobserved productivity.

4.4. Causal effects of work incentives on labor productivity

To measure causal impacts of career incentives on labor productivity, we compare job performance of enumerators who receive both wage and career incentives (*G3*) and that of enumerators with wage incentives only (*G4*). Similarly, we measure causal impacts of wage incentives by comparing job performance between enumerators with only career incentives (*G1*) and enumerators with both career and wage incentives (*G2*).

Panels B and C of Table 5 report the causal effects of career and wage incentives on job performance, respectively. Panels C and D in Figure 2 present the corresponding graphical evidence. In Panel B of Table 5, we find no evidence that career incentives improve job performance (survey quality, survey quantity, and SPE by respondents) although career incentives could motivate enumerators to exert more effort. Contrary to our expectation, we find a decrease of survey quantity (a decrease of 0.594 households per day, or 5.2%). This is mainly driven by an increase of intermission time (by 6.57 minutes), as shown in Panel B of Table A.4. This leads to a total increase of intermission time of 69.0 (=6.57*(11.5−1)) minutes per day, which dominates a 46.9-minute increase of total work hours. However, we find that SPE measured by supervisors significantly increases by 52.3%. In summary, career incentives given to existing workers hired through the wage incentive channel do not improve labor productivity, but they induce enumerators to have better evaluation from supervisors.

A possible explanation is that it is difficult (costly) for those hired through wage incentives, and whose work performance is lower compared to those hired through career incentives, to

improve work performance at least in the short run. They rather exerted efforts on the relationship with supervisors. Another possibility explanation is that career incentives might not be very appealing to enumerators recruited through wage incentives conditional on the self-selection. For example, they might have not have needed a job for a longer period because they might have been looking for only a short-term position due to another long-term commitment. Another explanation is that marginal effects of incentives in the second stage could be smaller, since they are a measure for people who had already received the incentive from the first stage and thus, were already motivated enough.

There may be concern that despite high frequency data, the relatively small number of enumerators allows for the detection of only relatively large effects and makes it difficult to interpret null results. Indeed, we are slightly under-powered in the regression analysis of Panel B in the sense that the size of the standard errors is not small enough to capture the small effect (if any) of the work incentives. For example, we are able to capture the causal impacts of career incentive on survey quality and quantity only if the change is greater than 26.3% (=0.011×1.96/0.082) and 10.1% (=0.594×1.96/11.5), respectively.[43] However, there is no reason to believe that additional career incentives *decrease* labor productivity.

Panel C of Table 5 shows that wage incentives improve job performance. We find that survey errors decrease by 2.8 percentage points (a 37.3% decrease) and survey quantity increases by 1.2 households per day (a 12.1% increase) without statistically significant changes in SPEs. Panel C of Table A.4 shows that the decrease in the survey error rate is explained by both a decrease in incorrect entries (Column (1)) and illogical missing entries (Column (3)). In addition, we find that an increase in survey quantity is driven by a decrease in actual survey time (i.e., each survey was completed more quickly) as shown in Column (7), Panel C of Table A.4.

This finding is consistent with the gift exchange model of the efficiency wage theory formulated by Akerlof (1984), by which a worker exerts more efforts upon receiving a gift from an employer that exceeds the minimum level of compensation for the minimum level of effort.

4.5 Discussion of Selection and Causal Effects

As stated in the introduction, one of the major contributions of our study is to separately identify the worker's self-selection (*G2* vs. *G3*) and causal effects of work incentives (*G1* vs. *G2* and *G3* vs. *G4*). Panel D of Table 5, which compares *G1* versus *G4*, resembles the combined effects of selection and causal effects on productivity in that participants were attracted to accept a job offer via different incentives and the incentives at work also remained different. We find no significant difference in the combined effects between *G1* and *G4* implying the importance of separating selection and causal effects. It is noteworthy that the combined effects of career incentives (Panel D) are not necessarily a simple sum of the selection effect (Panel A) and causal effect (Panel B). In addition, the study sample used in Panel D is different from that in Panels A and B. In Panel A, we use *G2* and *G3* to estimate the selection effect of career incentives. In Panel

[43] Other than causal effect of career incentives (Panel B), all other survey quantity and quality outcomes (Panel A and C) are statistically significant.

B, we use *G3* and *G4* to estimate the causal effect of career incentives. Thus, the sum of the coefficient estimates in Panels A and B are not the same as the coefficient estimates in Panel D.

Another point to discuss is unbalanced dropouts of the trainees. All 11 trainees who were dropped were from the *Internship* group. One may argue that if the labor productivity of the dropouts were lower than that of the hired enumerators, the performance-improving selection effects of an internship would be overestimated. However, we do not consider that any particular adjustment is necessary in the main analysis because screening out trainees who did not meet the minimum requirement is a regular business practice. Nevertheless, we re-estimate equation (3) after dropping 11 trainees with the lowest training scores from the *Wage* group.[44] Panel A of Table A.5 shows that the results on selection effect remain mostly robust; the size of the coefficients for the selection effect on survey quality becomes smaller while that on survey quantity becomes larger. A possible explanation is that training performance is not a precise predictor of job performance, consistent with the finding in Table 5 that the additional control of training outcomes does not change the main results. We find similar results on causal effects (Panels B and C) and combined effects (Panel D).

The last point to discuss is why *G2* outperforms *G3* for actual job performance while the *Wage* group outperforms the *Internship* group during the training. A potential explanation is that different skill sets required in each setting. The test taken during the training was in a classroom setting while job performance resulted from actual interactions with respondents in the field. Thus, it is plausible that those selected through the internship could have comparative advantages in on-the-job performance but not in tests in a classroom setting. A critical characteristic of an enumerator is the skill to ask strangers sensitive questions about their households. This kind of skill might not be captured easily in a test taken in a laboratory setting. In addition, it is possible that the *Internship* group initially had lower performance in the training but caught up with the *Wage* group later in the field due to a steeper learning curve. However, this is less likely, as we find no evidence for the catch-up. Job performance between the *Internship* and the *Wage* groups over the survey period remained constant over time (see Figure A.5. for the performance trend). The last explanation is that screening out 11 trainees in the *Internship* group could serve as a reminder or a credible threat to those with career incentives that only some of them would be hired as regular workers in AFF. However, this is also unlikely, since we find career incentives are limited at improving the job performance of existing workers in Panel B of Table 5.

[44] We drop six from *G3* and five from *G4*.

5. Additional Analysis

In this section, we present empirical evidence on the effect of supervisor visits on job performance and the 1-year effect of short-term job experience on employment.

5.1. Impacts of Supervisor Visits on Job Performance

As stated in Subsection 2.2.5, five supervisor teams visited enumerators on randomly selected dates to monitor and advise enumerators during the survey. Figure 3 shows how a supervisor visit affected job performance of enumerators before and after the visit. Panels A and B illustrate changes in job performance over time around the first and second supervisor visits, respectively. A vertical line at day zero represents the day of a supervisor visit. The figure shows that the survey error rate decreases over time, especially right after the supervisor visit, but survey quantity and SPE remain similar. However, with regard to the 95% confidence intervals in the figure, the changes are not statistically significant due to large standard errors. To quantify the changes in job performance after supervisor visits, we estimate the following equation:

$$Y_{ijklt} = \alpha + \eta_1 \cdot First_{lt} + \eta_2 \cdot Second_{lt} + \gamma \cdot H_i + \varphi \cdot Z_k + \delta \cdot X_j + \psi \cdot t + \mu_{ijktl} \quad (4)$$

where Y_{ijklt} is a job performance outcome for the survey collected from household *i* by enumerator *j* whose supervisor is *l*, in catchment area *k*, observed on the *t-th* work day. *First* and *Second* are binary indicators of surveys collected on the dates after the supervisor's first and second visits, respectively. X_j is a vector of individual characteristics including the standard set of control variables used above, and study group dummies.

Table 6 reports the regression results of equation (4). Columns (2), (4), and (6) include survey date-fixed effects instead of a linear time trend. Column (1) shows that the error rate decreases by 0.8 percentage points (a 10.8% improvement) after the first supervisor team visit although the increase is insignificant. In column (3), survey quantity increases by 1.06 households per day (a 9.7% increase) after the first visit.[45] This finding is not surprising because the supervisor visit could help enumerators accumulate job-specific human capital.

Lastly, the second visit does not lead to any statistically significant change in productivity. A possible explanation is that the additional visit does not lead to much skill acquisition. Another possible explanation is that only 37% of enumerators had a second visit, so the lack of statistical power could make it difficult to estimate the impact of the second supervisor visit precisely.

[45] However, the effects on both survey quality and quantity become smaller and insignificant when the survey date fixed effect is included. In addition, we find that SPE measured by respondents is negatively affected by the supervisor visit and the effect becomes statistically significant and large when we control for the survey date fixed effect.

5.2. Short-run Impacts on Employment

We conducted telephone-based follow-up surveys after 1 and 2 years from the completion of the original contract to examine the effects of short-term work experience on employment. Panels A to D of Table 7 present the 1- and 2-year effects on labor market outcomes by comparing the *Internship* group and the *Wage* group regardless of their offer take-up status.

AFF rehired 43 enumerators on a temporary basis among the 98 individuals with career incentives. The extended contract offered a daily wage of 500 MWK to conduct the PES or other surveys implemented by AFF. After the completion of the PES, some enumerators who remained for other surveys went on a labor strike asking for a steep wage increase, while AFF was still digitizing job performance. AFF decided not to rehire any enumerator from the study sample as a regular worker. As a result, at the time of the 1-year follow-up phone survey, none of the original enumerators with career incentives was working for AFF.

We find evidence that the *Internship* group is more likely to work for pay by 5.4 percentage points than the *Wage* group after 1 year from the short-term employment. The result remains similar after controlling for a variety of individual characteristics. If this is the case, we expect that an increase in employment is driven by the individuals who actually worked as enumerators. However, we acknowledge that this analysis cannot reveal a causal relationship, as the offer take-up decision was endogenously made by each participant. In Panel B, we find that study participants who accepted an internship offer are almost 10 percentage points more likely to work for pay compared to the *Wage* group participants. In addition, we find that the coefficient estimate of participants who declined the internship offer on employment is statistically non-significant and much smaller. However, as shown in Panels C and D, the impacts were attenuated over time in the sense that both economic magnitude and statistical significance after two years were substantially smaller.

A possible explanation for the 1-year impacts is that a recommendation letter provided to the *Internship* group enumerators could have helped them to find a job.[46] Another possible explanation is that the *Internship* group could have accumulated more skills useful in the labor market during the census because the nature of the career incentives made them work harder. On the other hand, a lack of 2-year impact reported in Panels C and D suggests that the internship experience for entry-level young workers in developing countries can be an effective way to facilitate school-to-work transition at least for a short period, but it does not improve longer-term labor market outcomes. In addition, it is possible that short-term job experience of less than 2 months does not allow workers to accumulate enough human capital to reap the labor market return over a longer period.[47]

[46] Abel et al. (2017) reported experimental evidence that is consistent with our finding that a recommendation letter increases employer callback rates by 60% in South Africa.

[47] It is noteworthy that the response rate of the 1-year follow-up survey was 98% but the 2-year follow-up survey response rate was 77%. Hence, the 2-year analysis is relatively under-powered compared to the 1-year analysis. In

6. Conclusion

There are two major channels through which an organization can increase the job performance of employees. The first is to offer attractive incentives to hire productive workers (selection effect) and the second is to motivate existing employees to exercise more efforts via providing more incentives at work (causal effect). This study analyzes how career and wage incentives affect labor productivity. Specifically, we separately estimate the selection and causal effects of work incentives on job performance through a two-stage randomized controlled trial in the context of a recruitment drive of census enumerators in Malawi.

We find that the *Internship* group (those attracted by career incentives) outperformed the *Wage* group (those attracted by wage incentives) at work. Observable individual characteristics are limited in explaining the difference in labor productivity due to self-selection. The fact that neither observable characteristics nor training outcomes predict actual job performance implies that screening via observable characteristics is imperfect, particularly when hiring entry-level workers who have no track record of past job history or credentials to verify their unobserved productivity. Our findings imply that it is difficult for an employer to foresee desirable attributes of an effective enumerator. Furthermore, these findings suggest the importance of a recruitment strategy to attract workers with strong unobservable skills via self-selection (e.g., an internship).

Regarding the causal effect of career incentives, we find no positive evidence for the causal effects of career incentives on job performance conditional on selection except for the subjective evaluation of work attitude by supervisors. Our findings suggest that career incentives are effective in improving labor productivity mainly through the selection effect channel. Lastly, we find that financial incentives in the form of a cash bonus can be an effective means to improve the job performance of existing workers as the gift exchange model predicts.

There are limitations to our study. First, we acknowledge that the approach by which we estimate the causal effects of incentives might not perfectly characterize the real world. In the real world, workers usually do not receive additional incentives without prior notice. Even if there were performance pay, workers are usually aware of it when they join the workplace. Second, the length of job we study is a relatively short-term position (less than 1 month). As such, we cannot study whether the estimated selection and causal effects of career and wage incentives remain constant over longer periods. The short-term nature of our study also limits the analysis of the effects of work incentives on retention. Third, we do not directly observe the individual's perception of the value of work incentives. In addition, we do not measure how career and wage incentives change workers' belief about the probability of retention by AFF. Hence, we do not know whether the selection effect of career incentives operates through the expectation of a job prospect at AFF or a potentially favorable recommendation letter. Fourth, non-cognitive traits used in this study are self-reported psychometric scales measured based on a paper test. It would be interesting to know whether such paper-based and self-reported non-cognitive traits are highly correlated with non-cognitive traits measured in other settings. Fifth, the relatively small number of enumerators may

both follow-up surveys, the response rates are well balanced across treatment arms.

prevent us from interpreting relatively small and insignificant effects, especially in estimating the causal effects of career incentives. However, most major outcomes (i.e., selection effects and causal effects of wage incentives) are large enough to detect their effects.

A better understanding of selection and causal effects of work incentives would allow employers to design optimal employment strategies. One of the major challenges of economic development is low labor productivity. In addition to the problem of absenteeism, difficulty of effective job applicant screening and lack of motivation among existing workers are key drivers of low labor productivity in developing countries. Based on our findings, we argue that active adoption of career incentives in the workplace as a hiring strategy could be an effective means to increase labor productivity of an organization. In addition, an unexpected cash bonus could be an effective means to motivate existing employees further.

The findings of this study specifically reflect the selection and causal effects of an enumerator job in Malawi. Hence, our findings might not be directly generalizable in developed countries and to labor markets in which worker quality is easily observable (e.g., manual labor). However, our analysis has several implications for jobs in different settings in which employers have difficulties screening out productive workers with little employment history and motivating existing workers. Testing our research questions in other settings, such as other countries and different types of jobs, would help us establish the external validity of our findings.

References

Abel, Martin, Burger, Rulof, and Piraino, Patrizio, 2017. The Value of Reference Letters—Experimental Evidence from South Africa. Working Paper.

Abeler, J., Falk, A., Goette, L. and Huffman, D., 2011. Reference points and effort provision. The American Economic Review, 101(2), pp. 470-492.

Akerlof, George A., 1984. Gift Exchange and Efficiency–Wage Theory: Four Views. *American Economic Review: Papers and Proceedings*, 74(2): 79–83.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2016. Do-gooders and Go-getters: Career Incentives, Selection, and Performance in Public Service Delivery. Harvard Business School Working Paper.

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. 2014. No Margin, No Mission? A Field Experiment on Incentives for Public Services Delivery. *Journal of Public Economics*, 120 (Dec 2014): 1-17.

Ashraf, Nava, James Berry, and Jesse M. Shapiro. 2010. Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100(5): 2383-2413.

Bandiera, Oriana, Iwan Barankay and Imran Rasul. 2011. Field Experiments with Firms. *Journal of Economic Perspectives*, 25(3): 63-82.

Beaman, Lori, Dean Karlan, Bram Thuysbaert, and Christopher Udry. 2015. Self-Selection into Credit Markets: Evidence from Agriculture in Mali. NBER Working Paper No. 20387. National Bureau of Economic Research.

Becker, Gary, Morris H. DeGroot, and Jacob Marschak. 1964. Measuring Utility by a Single-response Sequential Method. *Behavioral Science*, 9: 226–236.

Brooks, L., Cornelius, A., Greenfield, E. and Joseph, R. 1995. The relation of career-related work or internship experiences to the career development of college seniors. *Journal of Vocational Behavior*, 46(3), pp. 332-349.

Bohm, P., J. Lindén, and J. Sonnegård. 1997. Eliciting Reservation Prices: Becker–Degroot–Marschak Mechanisms vs. Markets. *The Economic Journal*, 107(443): 1079–1089.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. Who is (More) Rational? *American Economic Review* 104, No. 6: 1518-1550.

Cohen, J. and Dupas, P. 2010. Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, pp. 1-45.

D'Abate, C.P., Youndt, M.A. and Wenzel, K.E., 2009. Making the most of an internship: An

empirical study of internship satisfaction. *Academy of Management Learning & Education*, 8(4), pp. 527-539.

Dal Bó, Ernesto, Frederico Finan, and Martin A. Rossi, 2013. Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *Quarterly Journal of Economics*, 128(3): 1169–1218.

Deserranno, Erika. 2016. Financial Incentives as Signals: Experimental Evidence from the Recruitment of Health Workers. Working Paper.

Dohmen, T. and Falk, A. 2011. Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101(2), pp. 556-590.

Duflo, E., Hanna, R. and Ryan, S.P. 2012. Incentives work: Getting teachers to come to school. The *American Economic Review*, 102(4), pp. 1241-1278.

Fryer Roland. 2013. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*. 31(2): 373-427.

Gagliarducci, Stefano, and Tommaso Nannicini. 2013. Do Better Paid Politicians Perform Better? Disentangling Incentives From Selection. *Journal of the European Economic Association*.

Glewwe, P., Ilias, N., and Kremer, M., 2010. Teacher incentives. *American Economic Journal: Applied Economics*, 2(3), pp. 205-227.

Gneezy, Uri, and John A. List. 2006. Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments, *Econometrica*, 74(5): 1365-1384.

Guiteras, Raymond P., and B. Kelsey Jack. 2015. Incentives, selection and productivity in labor markets: Evidence from rural Malawi. NBER Working Paper No. 19825. National Bureau of Economic Research.

Horowitz, J.K., 2006. The Becker–DeGroot–Marschak Mechanism is Not Necessarily Incentive Compatible, Even for Non-random Goods. *Economics Letters*, 93(1): 6–11.

Karlan, D. and Zinman, J., 2009. Expanding credit access: Using randomized supply decisions to estimate the impacts. *Review of Financial studies*, 23 (1): 433-464.

Kling, Jeffrey, Jeffrey Liebman and Lawrence Katz. 2007. Experimental Analysis of Neighborhood Effects. *Econometrica*, 75(1): 83–119.

Lazear, Edward P. 2000. Performance Pay and Productivity. *American Economic Review*, 90(5): 1346–1361.

Liu, Y., Ferris, G.R., Xu, J., Weitz, B.A. and Perrewé, P.L., 2014. When ingratiation backfires: The role of political skill in the ingratiation–internship performance relationship. *Academy of Management Learning & Education*, 13(4), pp. 569-586.

National Statistical Office (NSO). 2014. Malawi Labour Force Survey 2013. Zomba, Malawi.

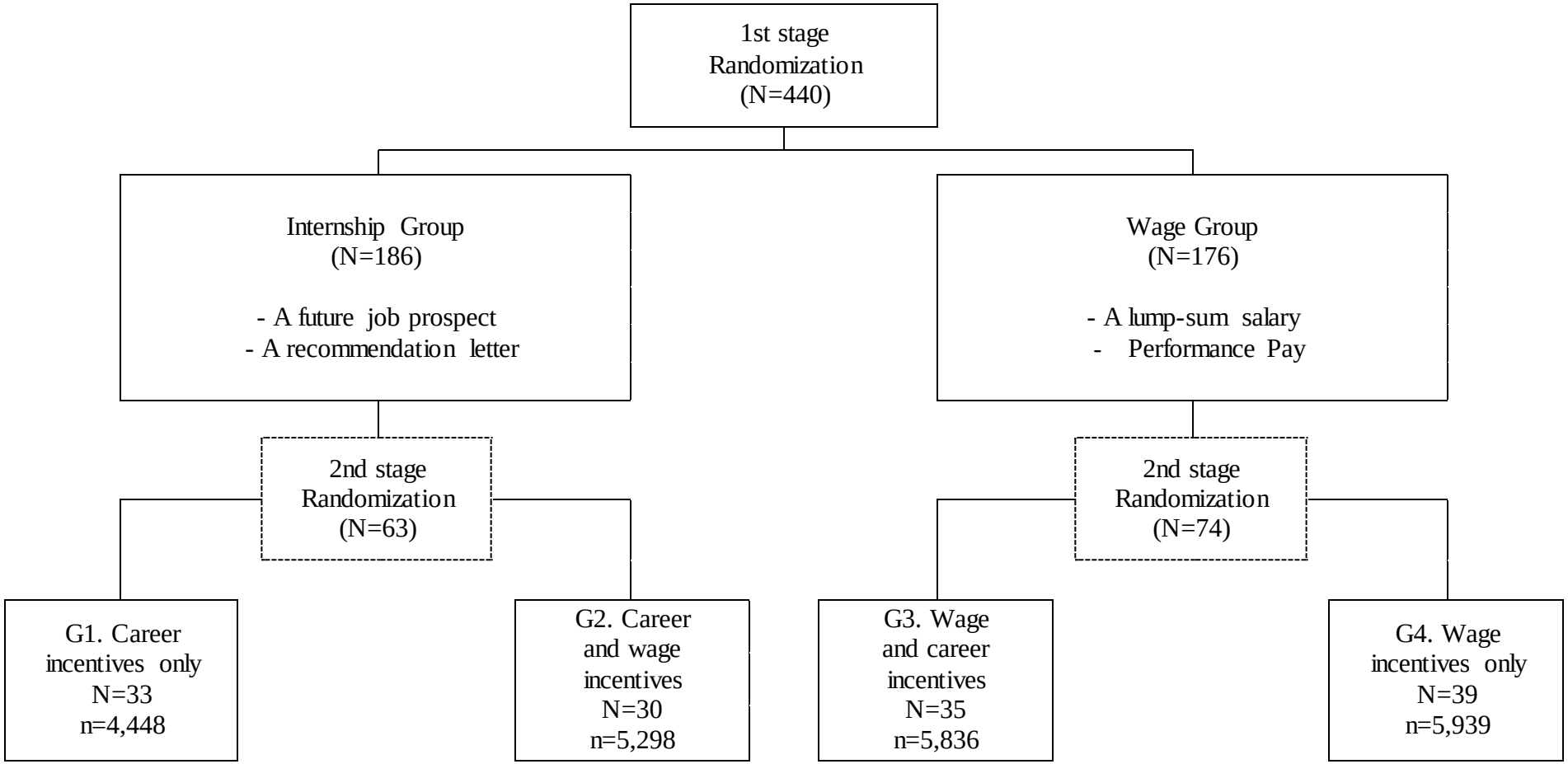
Nunley, J.M., Pugh, A., Romero, N. and Seals, R.A. 2016. College major, internship experience, and employment opportunities: Estimates from a résumé audit. *Labour Economics*, 38, pp. 37-46.

Oyer, Paul and Schaefer, S. 2011. Personnel Economics: Hiring and Incentives, In: Orley Ashenfelter and David Card, editors: *Handbook of Labor Economics*, Vol 4b, Great Britain, North Holland, pp. 1769-1823.

Shearer, B. 2004. Piece rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *Review of Economic Studies*, 71(2), pp. 513-534.

World Bank. 2016. World Development Indicators. http://data.worldbank.org/country/malawi

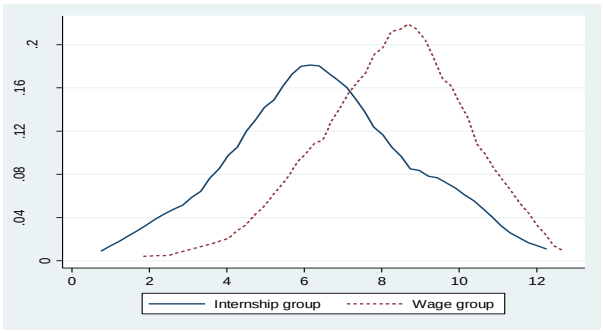
Figure 1: Experimental Design



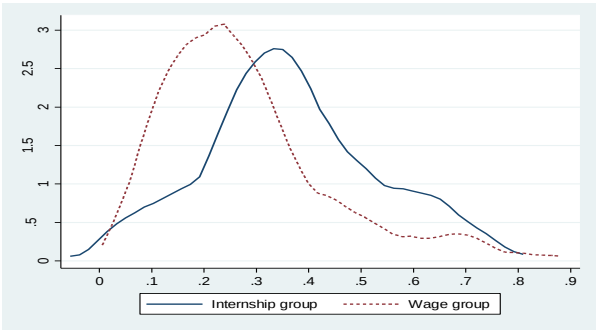
Notes: Upper case N indicates the number of participants in each stage. Lower case n indicates the number of surveys conducted by census enumerators.

Figure 2: Training Performance, Selection, Causal, and Combined Effect

Panel A: Training performance (*Internship* group vs. *Wage* group)

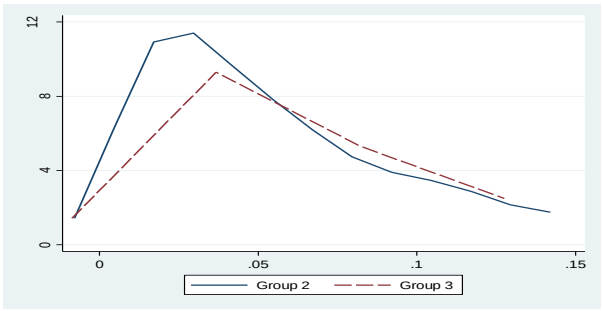


Quiz score

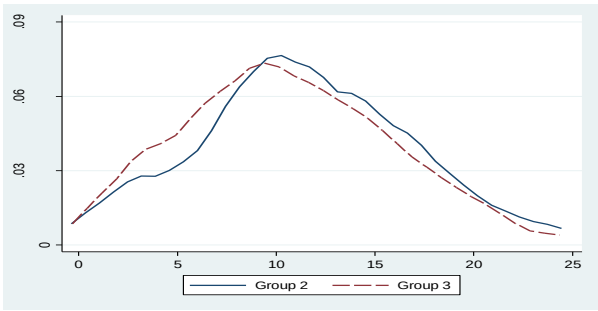


Practice survey error rate

Panel B: Selection Effect (G2 vs. G3)

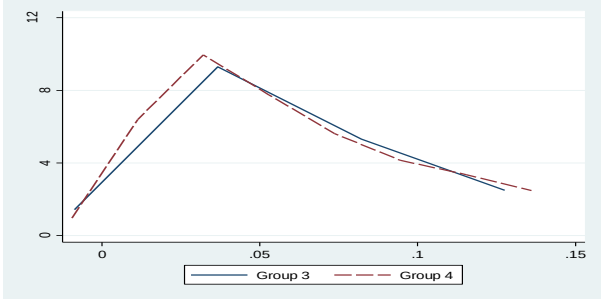


Survey quality (error rate)

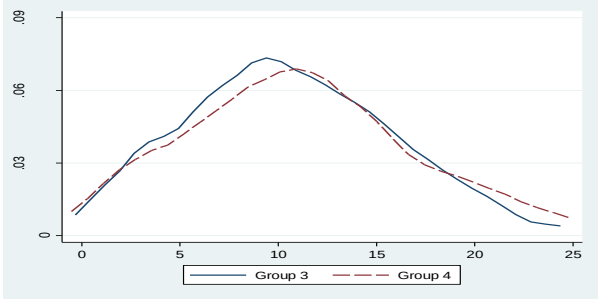


Survey quantity (number of surveys per day)

Panel C: Causal Effect of Career Incentives (G3 vs. G4)

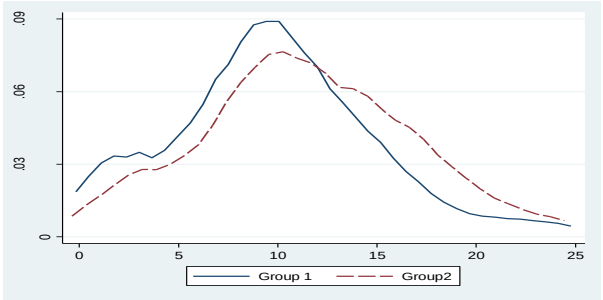


Survey quality (error rate)

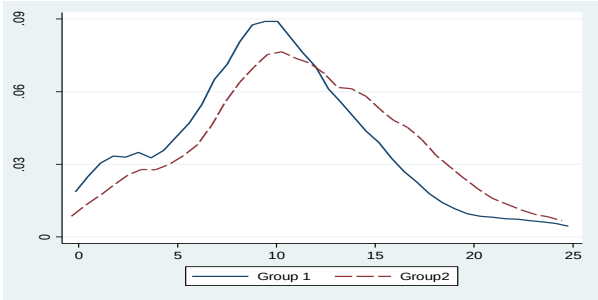


Survey quantity (number of surveys per day)

Panel D: Causal Effect of Wage Incentives (G1 vs. G2)



Survey quality (error rate)

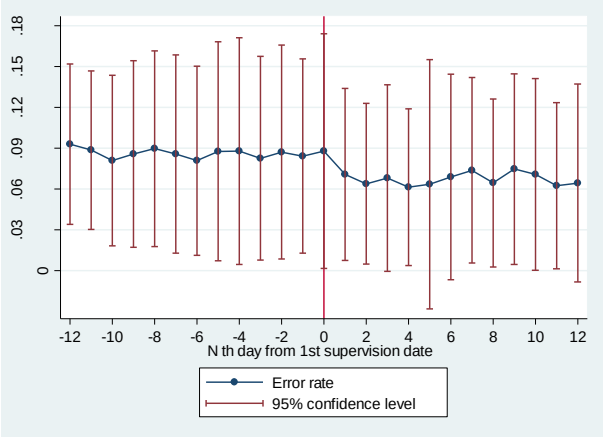


Survey quantity (number of surveys per day)

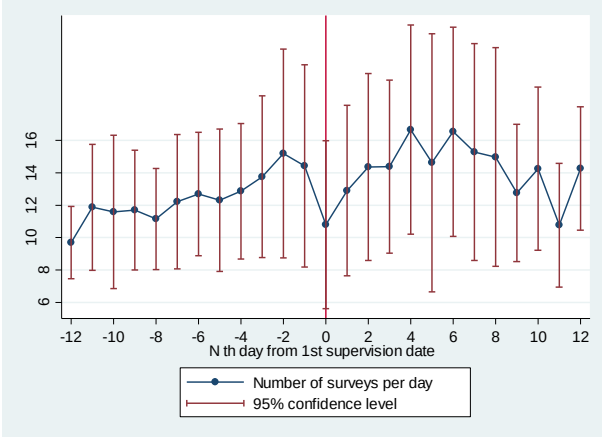
Notes: Panel A presents kernel density estimates of quiz score and practice survey error rate during the training. The *Internship* group received an unpaid job offer with career incentives in the first stage, while the *Wage* group received a non-renewable paid job offer in the first stage. Panels B, C, and D present kernel density estimates of survey quality and survey quantity. Groups 1 and 2 received career incentives in the first stage, but only Group 2 received additional wage incentives in the second stage. Groups 3 and 4 received wage incentives in the first stage, but only Group 3 received additional career incentives in the second stage.

Figure 3: Impact of Supervisor Visits on Job Performance

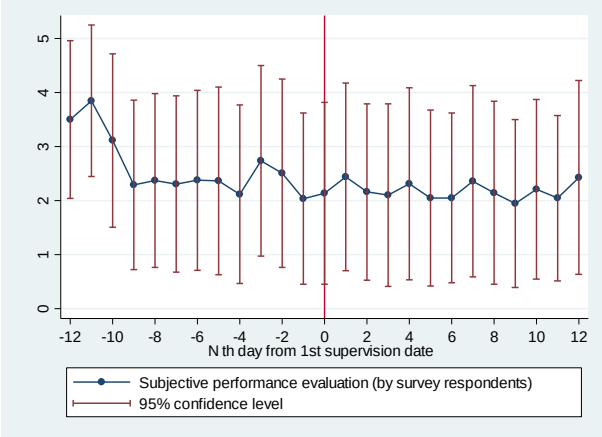
Panel A: Impacts of the first supervisor visit



Survey quality (error rate)

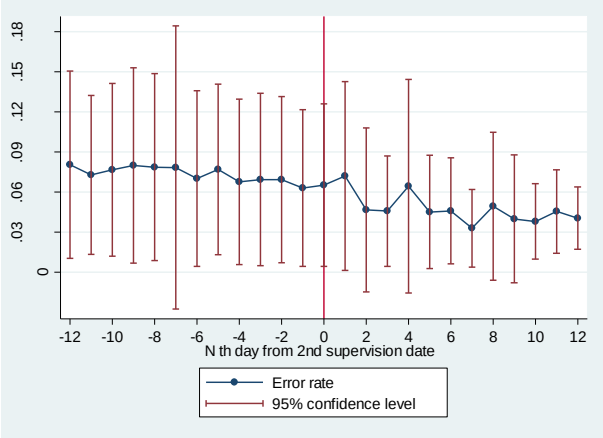


Survey quantity (Number of surveys per day)

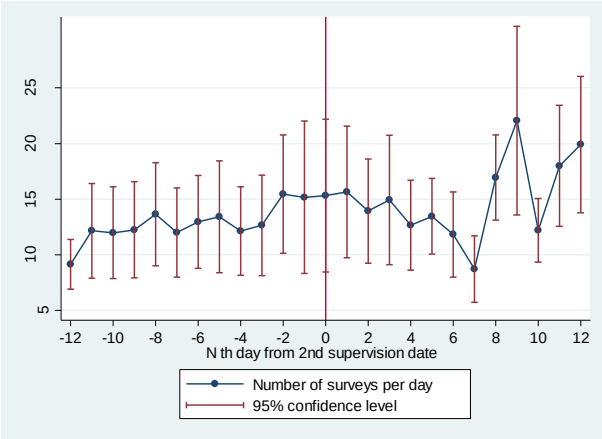


SPE by respondents

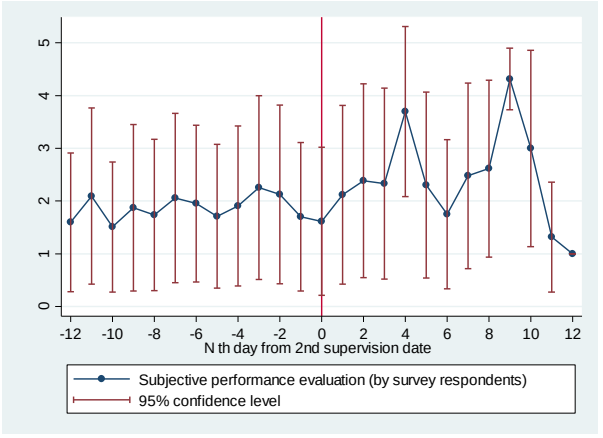
Panel B: Impacts of the second supervisor visit



Survey quality (error rate)



Survey quantity (Number of surveys per day)



SPE by respondents

Notes: The blue horizontal lines in each panel indicate the survey date-specific average of job performances before and after the supervisor visit at day 0. The red vertical lines with caps indicate 95% confidence intervals.

Table 1 Experiment Stages

Stage of experiment			Number of individuals					
			G1 (career incentives only)	G2 (career incentives and additional wage incentives)	G3 (wage incentives and additional career incentives)	G4 (wage incentives only)	p-value	Total
A	Target study subjects	2011 Dec	220		220		-	440
B	Study participants (baseline survey participants)	2014 Dec	186 (84.1%)		176 (80.0%)		.265	362
C	Trainees	2015 Jan	74 (39.8%)		74 (42.0%)		.663	148
D	Trainees who failed training		11		0		-	11
E	Enumerators	2015 Jan-Feb	63 (33.9%)		74 (42.0%)		-	137
			33	30	35	39		
F	Number of surveys		4,448	5,298	5,836	5,939	-	21,521

Note: The proportions of individuals remaining over experiment stages are in parentheses. The number of participants in the stage B is divided by the number of participants in the stage A, and the number of participants in the stages C and E are divided by the number of participants in the stage B.

Table 2 Randomization Balance Check

Variable	Number of observations	Internship group	Wage group	Mean difference (p-value) Internship vs Wage	Mean difference (p-value) G2 vs G1	Mean difference (p-value) G3 vs G4
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 2014 baseline survey						
Age	362	20.5	20.4	.065	-.200	-.207
		(.120)	(.126)	(.707)	(.629)	(.520)
Number of siblings	362	4.60	4.17	.432**	5.00	-.158
		(.132)	(.134)	(.022)	(.315)	(.650)
Asset score	362	1.09	1.19	-.102	.133	.048
		(.066)	(.067)	(.282)	(.489)	(.799)
Currently working	362	.097	.074	.023	.036	-.006
		(.022)	(.020)	(.455)	(.514)	(.913)
Self-esteem	362	19.4	19.3	-.158	.441	-.768
		(3.86)	(3.51)	(.684)	(.662)	(.341)
Intrinsic motivation	362	3.10	3.09	.010	.033	-.075
		(.330)	(.351)	(.644)	(.642)	(.372)
Extrinsic motivation	361	2.84	2.84	.000	.031	.004
		(.281)	(.285)	(.896)	(.646)	(.956)
Extroversion	358	3.61	3.47	.140	.055	-.246
		(1.12)	(1.20)	(.237)	(.851)	(.393)
Agreeableness	362	5.10	5.10	.008	.035	-.268
		(.106)	(.103)	(.955)	(.927)	(.408)
Conscientiousness	361	5.69	5.68	.010	.094	-.054
		(1.34)	(1.36)	(.908)	(.778)	(.850)
Emotional stability	360	5.08	5.06	.020	.064	-.190
		(1.49)	(1.42)	(.905)	(.866)	(.591)
Openness to experiences	362	5.14	5.10	.043	-.094	-.268
		(.114)	(.103)	(.778)	(.779)	(.408)
Time preference	334	.394	.398	-.004	.072*	.013
		(.011)	(.011)	(.783)	(.050)	(.697)
Risk preference	335	.629	.642	-.012	.008	-.033*
		(.007)	(.006)	(.181)	(.714)	(.077)
Rational decision-making ability	334	.817	.836	-.019	.037	-.007
		(.012)	(.011)	(.234)	(.353)	(.820)
Cognitive ability index	362	-.019	.049	-.068	.092	.001
		(.047)	(.049)	(.314)	(.556)	(.995)

Table 2 Randomization Balance Check (continued)

Variable	Number of observation	Internship group	Wage group	Mean difference (p-value) Internship vs Wage	Mean difference (p-value) G2 vs G1	Mean difference (p-value) G3 vs G4
	(1)	(2)	(3)	(4)	(5)	(6)
Male circumcision treatment	362	.425	.460	-.035	-.006	-.226*
		(.036)	(.038)	(.498)	(.962)	(.042)
HIV/AIDS education treatment	362	.511	.443	.068	-.009	.030
		(.037)	(.038)	(.199)	(.943)	(.800)
Scholarship treatment	362	.414	.500	-.086	.021	-.024
		(.036)	(.038)	(.101)	(.868)	(.838)
Transportation reimburse	362	1525	1547.7	-22.7	-103.9	-57.2
		(43.8)	(41.8)	(.708)	(.516)	(.707)
Panel B: Characteristics of dispatched catchment areas						
Number of households per enumerator	137	155.3	159.1	-3.79	40.6***	14.5
		(5.09)	(7.48)	(.676)	(.000)	(.335)
Family size	137	3.94	3.79	.148	.017	.114
		(.068)	(.081)	(.165)	(.170)	(.486)
Household asset score	137	.241	.253	-.012	-.017	.028*
		(.006)	(.007)	(.201)	(.170)	(.058)
Birth rate	137	.071	.065	.006**	.005	.010**
		(.002)	(.002)	(.019)	(.119)	(.026)
Death rate	137	.006	.006	.000	.001	-.001
		(.001)	(.001)	(.981)	(.590)	(.717)
Malaria incidence (under age 3)	137	.525	.513	.012	-.063**	-.018
		(.014)	(.019)	(.615)	(.025)	(.651)
Catchment area size	137	3.11	3.45	-.335	-.361	.238
		(.133)	(.255)	(.248)	(.178)	(.657)
Number of Observations		186	176		63	74

Notes: ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. Asset score is the number of items owned by a household out of the following: an improved toilet, a refrigerator, and a bicycle. See Data Appendix A.1 for detailed definitions of cognitive and non-cognitive trait variables. Male circumcision treatment, HIV/AIDS education treatment, and scholarship treatment are binary indicators for the treatment status of AFF's previous projects. Number of households is the average number of households that each enumerator was supposed to survey. Family size is the average number of family members per household. Household asset score is the number of items owned out of the following: improved toilet, bicycle, lamp, radio, cell phone, bed, and table and chair. Birth rate is the average number of births in the last 3 years per household. Death rate is the number of deaths in the last 12 months per household. Catchment area size is the land size subjectively reported by local health workers and AFF supervisors on a scale from 1 to 10.

Table 3 Job Offer Acceptance by Individual Trait

Dependent Variable (Job offer acceptance)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Age	BMI	Number of siblings	Asset score	Currently working	Self-esteem	Intrinsic motivation	Extrinsic motivation
Trait		.042	-.028	.038*	-.068*	-.107	-.024**	-.012	-.019
		(.030)	(.018)	(.019)	(.040)	(.136)	(.010)	(.108)	(.136)
Internship group	-.024	-.323	-.901*	-.029	-.023	-.025	-.321	.521	.733
	(.052)	(.747)	(.489)	(.131)	(.085)	(.055)	(.278)	(.491)	(.520)
Trait * Internship group		.015	.044*	-.002	-.009	.028	.015	-.176	-.266
		(.037)	(.025)	(.028)	(.054)	(.180)	(.014)	(.157)	(.182)
Constant	.481***	-.372	1.03***	.326***	.558***	.491***	.931***	.517	.537
	(.055)	(.613)	(.357)	(.094)	(.073)	(.057)	(.205)	(.336)	(.387)
Observations	362	362	360	362	362	362	362	362	361
R-squared	.018	.046	.028	.036	.036	.021	.034	.027	.031
Mean (SD)		20.4(1.65)	19.8(2.13)	4.39(1.80)	1.14(.896)	.086(.280)	19.3(3.69)	3.09(.340)	2.84(.282)

Dependent Variable (Job offer acceptance)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Extroversion	Agreeableness	Conscientiousness	Emotional stability	Openness to experiences	Time preference	Risk preference	Rational decision-making ability	Cognitive ability index
Trait	-.058*	-.001	.046*	.011	-.001	.196	.288	.019	-.126**
	(.032)	(.027)	(.026)	(.027)	(.027)	(.284)	(.498)	(.274)	(.053)
Internship group	-.297*	.025	.251	.145	.041	-.096	.388	-.228	-.034
	(.173)	(.196)	(.216)	(.195)	(.187)	(.158)	(.413)	(.305)	(.052)
Trait * Internship group	.077*	-.010	-.049	-.033	-.013	.199	-.644	.257	-.057
	(.046)	(.037)	(.037)	(.037)	(.035)	(.384)	(.640)	0.363)	(.073)
Constant	.683***	.486***	.223	.426***	.485***	.407***	.299	.502**	.490***
	(.126)	(.148)	(.152)	(.148)	(.148)	(.130)	(.324)	(.234)	(.054)
Observations	358	362	361	360	362	334	335	334	362
R-squared	.027	.019	.026	.020	0.019	.024	.019	.019	0.060
Mean (SD)	3.54(1.16)	5.11(1.39)	5.68(1.35)	5.07(1.45)	5.36(1.35)	.396(.144)	.635(.083)	.826(.149)	.348(.477)

Notes: Robust standard errors are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. Asset score is the sum of items owned out of improved toilet, refrigerator, and bicycle. See Data Appendix A.1 for the definitions of cognitive and non-cognitive trait variables.

Table 4: Training Performance

Dependent variable	Quiz score		Practice survey error rate		
	(1)	(2)	(3)	(4)	(5)
Panel A: 148 Trainee Sample					
Internship group	-2.01***	-1.93***	.104***	.089***	.234
	(.344)	(.308)	(.026)	(.029)	(.187)
Observations	148	148	148	148	148
R-squared	.228	.520	.114	.239	.800
Wage Group Mean (SD)	8.43 (1.82)		.272 (.142)		
Panel B: 137 Enumerator Sample					
Internship group	-1.44***	-1.45***	.094***	.080***	.229
	(.329)	(.294)	(.028)	(.030)	(.189)
Observations	137	137	137	137	137
R-squared	.163	.490	.099	.243	.856
Wage Group Mean (SD)	8.43 (1.82)		.272 (.142)		
Individual characteristics	No	YES	No	No	YES
Practice survey pair FE	No	No	No	YES	YES

Notes: Robust standard errors are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings and binary indicators for previous AFF programs. The practice survey error rate regression includes a binary indicator for the survey questionnaire type. A practice survey pair is a trainee pair who conducted the practice survey with each other. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items).

Table 5 Selection and Causal Effects of Work Incentives on Job Performance

VARIABLES	Survey quality (error rate)			Survey quantity (number of surveys per day)			Subjective performance evaluation (by survey respondents)			Subjective evaluation of work attitude (by supervisors)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Selection effect (G2 vs G3)												
G2	-.020*	-.021**	-.021**	1.48***	1.41***	1.31**	.783**	.691*	.682*	-.174*	-.137	-.135
	(.011)	(.008)	(.008)	(.516)	(.486)	(.546)	(.387)	(.364)	(.382)	(.100)	(.108)	(.115)
Observations	11,130	11,130	11,130	1,003	1,003	1,003	6,473	6,473	6,473	65	65	65
R-squared	.156	.302	.302	.144	.166	.173	.443	.592	.594	.401	.606	.634
Mean (SD) of G3	.077 (.078)			10.7 (5.45)			2.09 (1.65)			.850 (.163)		
Panel B: Causal effect of career incentives (G3 vs. G4)												
G3	.011	.006	.007	-.594	-.867	-.894	.095	.391	.327	.305***	.277***	.289***
	(.011)	(.010)	(.009)	(.602)	(.623)	(.612)	(.368)	(.351)	(.346)	(.038)	(.048)	(.050)
Observations	11,775	11,775	11,775	1,063	1,063	1,063	7,233	7,233	7,233	74	74	74
R-squared	.181	.265	.273	.149	.185	.189	.379	.492	.499	.619	.681	.693
Mean (SD) of G4	.082 (074)			11.5 (6.36)			2.08 (1.59)			.583 (.119)		
Panel C: Causal effect of wage (G1 vs. G2)												
G2	-.028*	-.019*	-.017	1.19*	1.18	1.18*	.276	.237	.021	-.134	-.151	-.238
	(.017)	(.011)	(.011)	(.619)	(.735)	(.679)	(.546)	(.608)	(.609)	(.155)	(.233)	(.224)
Observations	9,779	9,779	9,779	914	914	914	4,516	4,516	4,516	63	63	63
R-squared	.167	.354	.357	.203	.229	.238	.389	.607	.656	.366	.502	.561
Mean (SD) of G1	.075 (.068)			9.84 (5.19)			2.67 (1.66)			.803 (.162)		
Panel D: Combined effect (G1 vs. G4)												
G1	-.002	-.003	-.004	-1.45	-1.35	-.876	-.269	-.042	-.076	.191***	.202**	.191*
	(.013)	(.013)	(.014)	(.984)	(1.05)	(1.05)	(.474)	(.472)	(.552)	(.067)	(.092)	(.096)
Observations	10,424	10,424	10,424	974	974	974	5,276	5,276	5,276	72	72	72
R-squared	.194	.276	.277	.157	.221	.225	.517	.623	.628	.569	.627	.636
Mean (SD) of G4	.082 (074)			11.5 (6.36)			2.08 (1.59)			.583 (.119)		
Individual characteristics	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings, catchment area characteristics, supervisor team-specific post visit variables, survey date-fixed effect, and binary indicator variables for previous AFF programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Training performances include the quiz score and practice survey error rate. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

Table 6: Impacts of Supervisor Visits

Variable	Survey quality (error rate)		Survey quantity (number of surveys per day)		Subjective performance evaluation (by survey respondents)	
	(1)	(2)	(3)	(4)	(5)	(6)
First visit	-.008	-.005	1.06*	.839	-.349	-.536**
	(.006)	(.006)	(.557)	(.606)	(.244)	(.251)
Second visit	.008	.005	-1.78	-.954	-.118	.149
	(.007)	(.008)	(1.36)	(1.44)	(.275)	(.310)
Observations	20,381	20,381	1,841	1,841	11,099	11,099
R-squared	.221	.228	.086	.125	.273	.296
Linear time trend	YES	No	YES	No	YES	No
Survey date fixed effect	No	YES	No	YES	No	YES
Mean (SD) of the dependent variable	.074 (.071)		10.9 (5.70)		2.29 (1.69)	

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote significance at 1%, 5%, and 10%, respectively. All specifications include catchment area characteristics, study group fixed effect (G1–G4), and binary indicator variables for previous AFF programs. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, malaria incidence among children under 3, and deaths in the last 12 months. Individual characteristics include age, number of siblings, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items).

Table 7: Short-term Impacts of Job Experience on Employment

VARIABLES	Currently working for paid job	
	(1)	(2)
Panel A.1: 1-year effect of career incentives (Internship group vs. Wage group)		
Received an internship offer	.054**	.048*
	(.027)	(.027)
Observations	355	349
R-squared	.029	.080
Wage Group Mean (SD)	.041 (.198)	
Panel B: 1-year effect of those who accepted and rejected an internship offer vs Wage group		
Accepted an internship offer	.099**	.091**
	(.045)	(.045)
Declined an internship offer	.025	.018
	(.029)	(.029)
Observations	355	349
R-squared	.038	.090
Omitted group Mean (Standard Deviation)	.041 (.198)	
Panel C: 2-year effect of career incentives (Internship group vs. Wage group)		
Received an internship offer	.015	.018
	(.037)	(.037)
Observations	277	273
R-squared	.008	.100
Wage Group Mean (SD)	.102 (.304)	
Panel D: 2-year effect of those who accepted and rejected an internship offer vs Wage group		
Accepted an internship offer	.032	.027
	(.052)	(.055)
Declined an internship offer	.002	.011
	(.042)	(.041)
Observations	277	273
R-squared	.009	.100
Omitted group Mean (Standard Deviation)	.102 (.304)	
Individual characteristics	NO	YES

Notes: ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include binary indicator variables for previous AFF programs. Individual characteristics include age, number of siblings, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items).

Online Appendix (not for publication)

Appendix Tables

Table A.1: Randomization balance check between treatment and non-selected groups

Variable	Number of observation	Internship + Wage group	Non-selected group	Mean difference (p-value) Internship + Wage vs Non-selected
	(1)	(2)	(3)	(4)
Panel A: 2011 secondary school census survey				
Height	534	164.5	164.5	.047
		(.367)	(.743)	(.955)
Weight	535	53.5	53.9	-.430
		(.342)	(.984)	(.680)
Age in 2011	536	16.1	16.0	.065
		(.070)	(.197)	(.758)
Living with a father	536	.639	.645	-.006
		(.023)	(.050)	(.908)
Living with a mother	536	.747	.667	.081
		(.021)	(.049)	(.134)
Asset score in 2011	530	1.17	1.41	-.240**
		(.042)	(.106)	(.037)
Subjective health is good or very good	536	.433	.538	.104*
		(.024)	(.052)	(.070)
Raven matrix test score	452	20.0	18.7	1.32
		(.244)	(.696)	(.077)
Number of observations	536	440	96	
Panel B: 2014 baseline survey				
Age in 2014	443	20.4	20.0	.395**
		(.087)	(.159)	(.031)
Number of siblings	443	4.39	4.49	.071
		(.094)	(.243)	(.771)
Asset score in 2014	443	1.14	1.22	-.084
		(.047)	(.102)	(.457)
Currently working	442	.086	.099	-.014
		(.015)	(.033)	(.697)
Self-esteem	443	19.3	20.1	-.706
		(.194)	(.421)	(.134)
Intrinsic motivation	443	3.09	3.10	-.005
		(.018)	(.038)	(.912)
Extrinsic motivation	442	2.84	2.81	-.026
		(.015)	(.031)	(.480)

Table A.1: Randomization balance check between treatment and non-selected groups (continued)

Variable	Number of observation	Internship + Wage group	Non-selected group	Mean difference (p-value) Internship + Wage vs Non-selected
	(1)	(2)	(3)	(4)
Extroversion	433	3.54	3.44	.103
		(.061)	(.136)	(.523)
Agreeableness	443	5.10	5.46	-.356**
		(.074)	(.149)	(.034)
Conscientiousness	442	5.69	6.17	-.487***
		(.071)	(.147)	(.002)
Emotional stability	439	5.07	5.31	-.237
		(.076)	(.164)	(.207)
Openness to experiences	443	5.12	5.76	-.332
		(.077)	(.150)	(.115)
Time preference	402	.396	.366	.030
		(.008)	(.016)	(.101)
Risk preference	403	.635	.656	-.020*
		(.005)	(.011)	(.089)
Rational decision-making ability	402	.826	.786	.040*
		(.008)	(.020)	(.068)
Cognitive ability index	443	.014	.049	.084
		(.034)	(.049)	(.297)
Male circumcision treatment	443	.442	.506	-.064
		(.026)	(.056)	(.300)
HIV/AIDS education treatment	443	.478	.506	-.028
		(.026)	(.056)	(.648)
Scholarship treatment	443	.456	.469	-.013
		(.026)	(.056)	(.829)
Transportation reimburse	443	1536	1511.1	24.9
		(30.3)	(69.2)	(.742)
Number of observations	443	362	81	

Notes: ***, **, and * denote the significance level at 1%, 5%, and 10% respectively.

Table A.2: Individual characteristics between baseline survey participants and non-participants

Variable	Participants	Non-participants	Mean difference between participants and non-participants (p-value)
	(1)	(2)	(3)
Height	164.6	164.5	.071
	(.420)	(.818)	(.939)
Weight	53.6	54.1	-.486
	(.377)	(1.09)	(.674)
Age	16.1	16.0	.134
	(.078)	(.222)	(.571)
Living with a father	.667	.622	-.045
	(.054)	(.026)	(.450)
Living with a mother	.740	.679	.061
	(.023)	(.053)	(.296)
Asset score	2.46	3.12	-.656***
	(.086)	(.197)	(.003)
Subjective health (good or very good)	.428	.551	.123*
	(.026)	(.057)	(.051)
Raven's matrices test score	19.9	18.8	1.04
	(.274)	(.785)	(.216)
Number of observations	362	78	

Notes: ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. The statistics are calculated based on data from the 2011 secondary school survey. Columns (1) and (2) show group-specific means and standard deviations. 440 male secondary school graduates were randomly selected to receive a job offer without prior notice, but only 362 showed up on the survey date.

Table A.3: Individual characteristics after job offer acceptance

Variable	Number of observations	Internship offer takers	Wage offer takers	Mean Difference	Standard Deviation
	(1)	(2)	(3)	(4)=(2)-(3)	(5)
Age	148	20.8	20.7	.162	1.46
BMI	148	19.9	19.5	.413	2.08
Number of siblings	148	4.86	4.46	.405	1.70
Asset score	148	.932	1.05	-.122	.804
Currently working	148	.081	.054	.027	.252
Self-esteem	148	19.1	18.6	.521	3.71
Intrinsic motivation	148	3.05	3.08	-.029	.326
Extrinsic motivation	148	2.78	2.83	-.046	.274
Extroversion	148	3.67	3.27	.405**	1.19
Agreeableness	148	5.03	5.10	-.074	1.44
Conscientiousness	148	5.67	5.87	-.196	1.26
Emotional stability	148	4.94	5.12	-.182	1.50
Openness to experiences	148	5.03	5.10	-.074	1.44
Time preference	137	.414	.411	.003	.136
Risk preference	137	.621	.645	-.024*	.079
Rational decision-making ability	137	.831	.834	-.004	.139
Cognitive Ability Index	148	-.199	-.077	-.119	.591
Male circumcision treatment	148	.392	.338	.054	.483
HIV/AIDS education treatment	148	.473	.473	.000	.501
Scholarship treatment	148	.459	.473	-.013	.501
Transportation reimburse	148	1602.7	1652.7	-50.0	628.2
F-statistics (p-value)				.950 (.532)	
Number of Individuals		74	74	148	148

Notes: ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. Asset score is the sum of items owned out of improved toilet, refrigerator, and bicycle. See Data Appendix A.1 for the definitions of cognitive and non-cognitive trait variables. Male circumcision, HIV/AIDS education treatment, and scholarship are binary indicator variables of the past eligibility status of AFF's previous programs.

Table A.4: Selection and causal effects of work incentives on job performance: additional outcomes

VARIABLES	Survey quality				Survey quantity					
	Proportion of entries incorrectly entered		Proportion of entries incorrectly blank		Work hours (in mins)		Survey time per household (in mins)		Intermission time between surveys (in mins)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Selection effect (G2 vs G3)										
G2	.001	-.001	-.021**	-.020***	-1.24	-4.62	-1.51	-1.09	-4.22	-3.90*
	(.003)	(.002)	(.009)	(.007)	(18.3)	(17.2)	(1.14)	(.978)	(2.54)	(2.31)
Observations	11,130	11,130	11,130	11,130	988	988	11,130	11,130	8,224	8,223
R-squared	.107	.242	.148	.264	.146	.178	.282	.324	.019	.029
Mean (SD) of G3	.016 (.018)		.062 (.070)		422.7 (198.7)		25.2 (10.9)		23.1 (50.2)	
Panel B: Causal effect of career incentives (G3 vs. G4)										
G3	.002	.000	.010	.007	46.9***	37.7**	1.09	1.40	6.57***	5.82***
	(.003)	(.003)	(.009)	(.008)	(17.2)	(18.4)	(1.25)	(1.15)	(1.94)	(2.00)
Observations	11,775	11,775	11,775	11,775	1,054	1,053	11,775	11,775	9,040	9,040
R-squared	.161	.298	.161	.222	.146	.168	.250	.268	.019	.026
Mean (SD) of G4	.019 (.021)		.063 (.066)		387.0 (194.8)		23.9 (11.2)		17.3 (44.0)	
Panel C: Causal effect of wage (G1 vs. G2)										
G2	-.006	-.007**	-.022	-.010	21.9	25.2	-2.97*	-1.88	1.01	.339
	(.004)	(.003)	(.014)	(.009)	(24.0)	(23.2)	(1.52)	(1.60)	(3.15)	(3.51)
Observations	9,779	9,779	9,779	9,779	889	888	9,780	9,780	7,203	7,202
R-squared	.102	.235	.148	.299	.190	.223	.305	.341	.021	.032
Mean (SD) of G1	.019 (.019)		.056 (.061)		382.0 (188.2)		27.4 (12.1)		19.3 (41.8)	
Panel D: Combined effect (G1 vs. G4)										
G1	.007*	.012**	-.009	-.016	-17.9	-30.2	2.21	.157	2.39	1.43
	(.004)	(.005)	(.011)	(.011)	(25.7)	(28.8)	(1.43)	(1.50)	(2.32)	(2.18)
Observations	10,424	10,424	10,424	10,424	955	953	10,425	10,425	8,019	8,019
R-squared	.158	.262	.167	.239	.157	.187	.282	.332	.014	.023
Mean (SD) of G4	.019 (.021)		.063 (.066)		387.0 (194.8)		23.9 (11.2)		17.3 (44.0)	
Individual characteristics	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Training performance	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES

Notes: Standard errors clustered at enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10% respectively. All specifications include number of siblings, survey-date fixed effect, catchment area control variables, supervisor team-specific post visit variables, and binary indicators of the past eligibility status of AFF's previous programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Catchment area control variables include the total number of households, the number of family members, asset score (whether to own improved toilet, bicycle, lamp, radio, cell phone, bed, and table and chair), the number of births per household in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

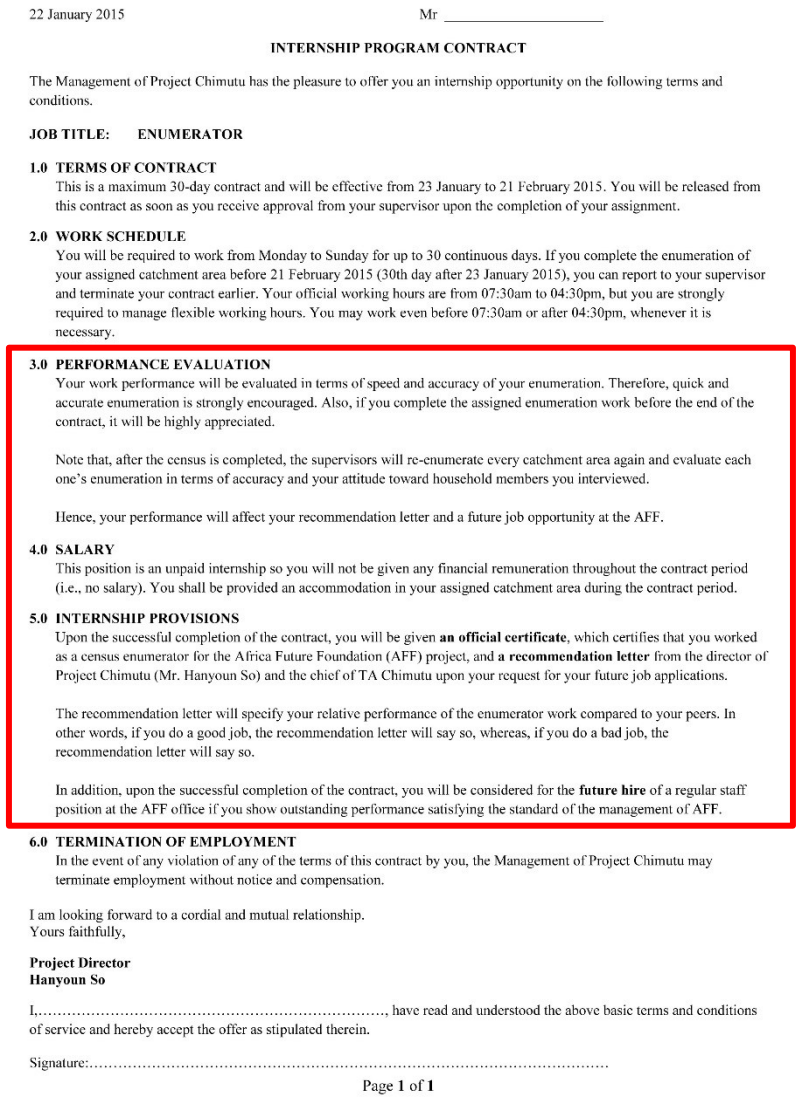
Table A.5: Selection and causal effects of work incentives on job performance after excluding 11 enumerators from the *Wage* group

VARIABLES	Survey quality (error rate)			Survey quantity (number of surveys)			Subjective performance evaluation (by survey respondents)			Subjective evaluation of work attitude (by supervisors)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Selection effect (G2 vs G3)												
G2	-.005	-.014*	-.012	1.82***	1.70***	1.60**	.843**	.814**	.700	-.186*	-.156	-.143
	(.010)	(.007)	(.008)	(.540)	(.530)	(.627)	(.399)	(.378)	(.447)	(.104)	(.150)	(.149)
Observations	10,150	10,150	10,150	917	917	917	5,906	5,906	5,906	59	59	59
R-squared	.165	.293	.294	.152	.172	.177	.446	.584	.587	.394	.617	.657
Mean (SD) of G3	.067 (.064)			10.6 (5.60)			2.11 (1.66)			.845 (.169)		
Panel B: Causal effect of career incentives (G3 vs. G4)												
G3	.011	.013	.012	-1.30**	-1.82**	-1.97**	.342	.594	.515	.325***	.308***	.324***
	(.009)	(.010)	(.011)	(.624)	(.764)	(.745)	(.410)	(.371)	(.363)	(.052)	(.076)	(.083)
Observations	9,666	9,666	9,666	876	876	876	5,983	5,983	5,983	63	63	63
R-squared	.197	.258	.260	.178	.207	.215	.348	.518	.526	.610	.692	.713
Mean (SD) of G4	.085 (.076)			11.5 (6.47)			1.94 (1.52)			.596 (.123)		
Panel C: Causal effect of wage (G1 vs. G2)												
G2	-.028*	-.019*	-.017	1.19*	1.18	1.18*	.276	.237	.021	-.129	-.151	-.238
	(.017)	(.011)	(.011)	(.619)	(.735)	(.679)	(.546)	(.608)	(.609)	(.130)	(.233)	(.224)
Observations	9,779	9,779	9,779	914	914	914	4,516	4,516	4,516	63	63	63
R-squared	.167	.354	.357	.203	.229	.238	.389	.607	.656	.344	.502	.561
Mean (SD) of G1	.075 (.068)			9.84 (5.19)			2.67 (1.66)			.803 (.162)		
Panel D: Combined effect (G1 vs. G4)												
G1	.000	.002	.008	-1.32	-1.27	-.387	.013	.666	.767	.154**	.095	.071
	(.013)	(.013)	(.014)	(1.02)	(1.23)	(1.25)	(.439)	(.456)	(.550)	(.063)	(.099)	(.105)
Observations	9,295	9,295	9,295	873	873	873	4,593	4,593	4,593	67	67	67
R-squared	.196	.282	.290	.177	.232	.239	.574	.710	.718	.587	.723	.742
Mean (SD) of G4	.085 (.076)			11.5 (6.47)			1.94 (1.52)			.596 (.123)		
Individual characteristics	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES

Notes: 11 enumerators in the *Wage* group whose training performance is the lowest are excluded. Standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include binary indicators of the past eligibility status of AFF's previous programs, number of siblings, supervisor-specific post-visit fixed effect, survey date-fixed effect and catchment area characteristics which include the total number of households, the number of family members, asset score, number of births in the last 3 years, number of incidences of malaria among children under 3, and number of deaths in the last 12 months. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items).

Appendix Figures

Figure A.1: Contract letter for Group 1 (G1)



22 January 2015 Mr _____

INTERNSHIP PROGRAM CONTRACT

The Management of Project Chimutu has the pleasure to offer you an internship opportunity on the following terms and conditions.

JOB TITLE: ENUMERATOR

1.0 TERMS OF CONTRACT
This is a maximum 30-day contract and will be effective from 23 January to 21 February 2015. You will be released from this contract as soon as you receive approval from your supervisor upon the completion of your assignment.

2.0 WORK SCHEDULE
You will be required to work from Monday to Sunday for up to 30 continuous days. If you complete the enumeration of your assigned catchment area before 21 February 2015 (30th day after 23 January 2015), you can report to your supervisor and terminate your contract earlier. Your official working hours are from 07:30am to 04:30pm, but you are strongly required to manage flexible working hours. You may work even before 07:30am or after 04:30pm, whenever it is necessary.

3.0 PERFORMANCE EVALUATION
Your work performance will be evaluated in terms of speed and accuracy of your enumeration. Therefore, quick and accurate enumeration is strongly encouraged. Also, if you complete the assigned enumeration work before the end of the contract, it will be highly appreciated.

Note that, after the census is completed, the supervisors will re-enumerate every catchment area again and evaluate each one's enumeration in terms of accuracy and your attitude toward household members you interviewed.

Hence, your performance will affect your recommendation letter and a future job opportunity at the AFF.

4.0 SALARY
This position is an unpaid internship so you will not be given any financial remuneration throughout the contract period (i.e., no salary). You shall be provided an accommodation in your assigned catchment area during the contract period.

5.0 INTERNSHIP PROVISIONS
Upon the successful completion of the contract, you will be given **an official certificate**, which certifies that you worked as a census enumerator for the Africa Future Foundation (AFF) project, and **a recommendation letter** from the director of Project Chimutu (Mr. Hanyoun So) and the chief of TA Chimutu upon your request for your future job applications.

The recommendation letter will specify your relative performance of the enumerator work compared to your peers. In other words, if you do a good job, the recommendation letter will say so, whereas, if you do a bad job, the recommendation letter will say so.

In addition, upon the successful completion of the contract, you will be considered for the **future hire** of a regular staff position at the AFF office if you show outstanding performance satisfying the standard of the management of AFF.

6.0 TERMINATION OF EMPLOYMENT
In the event of any violation of any of the terms of this contract by you, the Management of Project Chimutu may terminate employment without notice and compensation.

I am looking forward to a cordial and mutual relationship.
Yours faithfully,

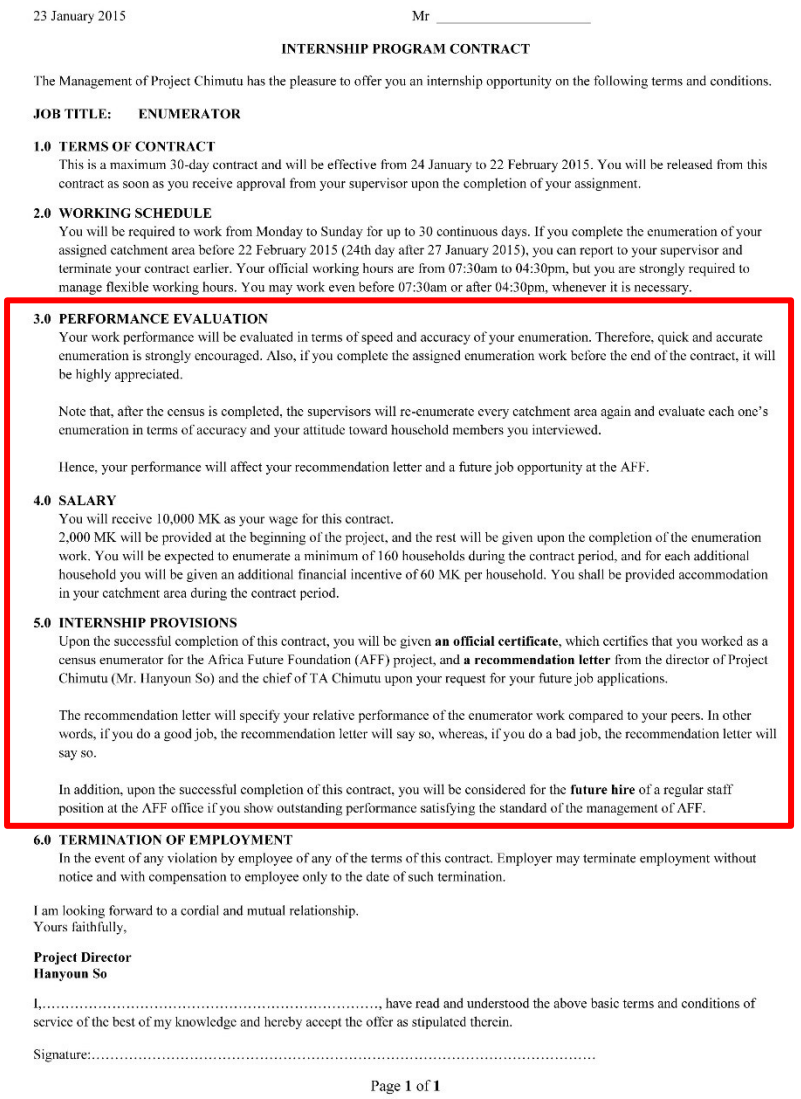
Project Director
Hanyoun So

I,……, have read and understood the above basic terms and conditions of service and hereby accept the offer as stipulated therein.

Signature:……

Page **1** of **1**

Figure A.2: Contract letter for Group 2 (G2) and Group (G3) (the same contract letter for both groups)



23 January 2015 Mr _____

INTERNSHIP PROGRAM CONTRACT

The Management of Project Chimutu has the pleasure to offer you an internship opportunity on the following terms and conditions.

JOB TITLE: ENUMERATOR

1.0 TERMS OF CONTRACT
This is a maximum 30-day contract and will be effective from 24 January to 22 February 2015. You will be released from this contract as soon as you receive approval from your supervisor upon the completion of your assignment.

2.0 WORKING SCHEDULE
You will be required to work from Monday to Sunday for up to 30 continuous days. If you complete the enumeration of your assigned catchment area before 22 February 2015 (24th day after 27 January 2015), you can report to your supervisor and terminate your contract earlier. Your official working hours are from 07:30am to 04:30pm, but you are strongly required to manage flexible working hours. You may work even before 07:30am or after 04:30pm, whenever it is necessary.

3.0 PERFORMANCE EVALUATION
Your work performance will be evaluated in terms of speed and accuracy of your enumeration. Therefore, quick and accurate enumeration is strongly encouraged. Also, if you complete the assigned enumeration work before the end of the contract, it will be highly appreciated.

Note that, after the census is completed, the supervisors will re-enumerate every catchment area again and evaluate each one's enumeration in terms of accuracy and your attitude toward household members you interviewed.

Hence, your performance will affect your recommendation letter and a future job opportunity at the AFF.

4.0 SALARY
You will receive 10,000 MK as your wage for this contract.
2,000 MK will be provided at the beginning of the project, and the rest will be given upon the completion of the enumeration work. You will be expected to enumerate a minimum of 160 households during the contract period, and for each additional household you will be given an additional financial incentive of 60 MK per household. You shall be provided accommodation in your catchment area during the contract period.

5.0 INTERNSHIP PROVISIONS
Upon the successful completion of this contract, you will be given **an official certificate**, which certifies that you worked as a census enumerator for the Africa Future Foundation (AFF) project, and **a recommendation letter** from the director of Project Chimutu (Mr. Hanyoun So) and the chief of TA Chimutu upon your request for your future job applications.

The recommendation letter will specify your relative performance of the enumerator work compared to your peers. In other words, if you do a good job, the recommendation letter will say so, whereas, if you do a bad job, the recommendation letter will say so.

In addition, upon the successful completion of this contract, you will be considered for the **future hire** of a regular staff position at the AFF office if you show outstanding performance satisfying the standard of the management of AFF.

6.0 TERMINATION OF EMPLOYMENT
In the event of any violation by employee of any of the terms of this contract. Employer may terminate employment without notice and with compensation to employee only to the date of such termination.

I am looking forward to a cordial and mutual relationship.
Yours faithfully,

Project Director
Hanyoun So

I,……, have read and understood the above basic terms and conditions of service of the best of my knowledge and hereby accept the offer as stipulated therein.

Signature:……

Page **1** of **1**

Figure A.3: Contract letter for Group 4 (G4)

AFRICA FUTURE FOUNDATION

26 January 2015 Mr _____

TEMPORARY EMPLOYMENT CONTRACT

The Management of Project Chimutu has the pleasure to offer you a temporary employment opportunity on the following terms and conditions.

JOB TITLE: ENUMERATOR

1.0 TERMS OF CONTRACT

This is a maximum 30-day contract and will be effective from 27 January to 25 February 2015. You will be released from this contract as soon as you receive approval from your supervisor upon the completion of your assignment.

2.0 WORK SCHEDULE

You will be required to work from Monday to Sunday for up to 30 continuous days. If you complete the enumeration of your assigned catchment area before 25 February 2015 (30th day after 27 January 2015), you can report to your supervisor and terminate your contract earlier. Your official working hours are from 07:30am to 04:30pm, but you are strongly required to manage flexible working hours. You may work even before 07:30am or after 04:30pm, whenever it is necessary.

3.0 PERFORMANCE EVALUATION

Your work performance will be evaluated in terms of speed and accuracy of your enumeration. Therefore, quick and accurate enumeration is strongly encouraged. Also, if you complete the assigned enumeration work before the end of the contract, it will be highly appreciated.

Note that, after the census is completed, the supervisors will re-enumerate every catchment area again and evaluate each one's enumeration in terms of accuracy and your attitude toward household members you interviewed.

4.0 SALARY

You will receive 10,000 MK as your wage for this contract.
2,000 MK will be provided at the beginning of the project, and the rest will be given upon the completion of the enumeration work. You will be expected to enumerate a minimum of 160 households during the contract period, averaging 8 households per day. When you enumerate more than 160 households, you will be given an additional financial incentive of 500 MK per 8 households. You shall be provided accommodation in your catchment area during the contract period.

5.0 TERMINATION OF EMPLOYMENT

In the event of any violation of any of the terms of this contract by you, the Management of Project Chimutu may terminate employment without notice and compensation.

I am looking forward to a cordial and mutual relationship.
Yours faithfully,

Project Director
Hanyoun So

I,……, have read and understood the above basic terms and conditions of service and hereby accept the offer as stipulated therein.

Signature:……

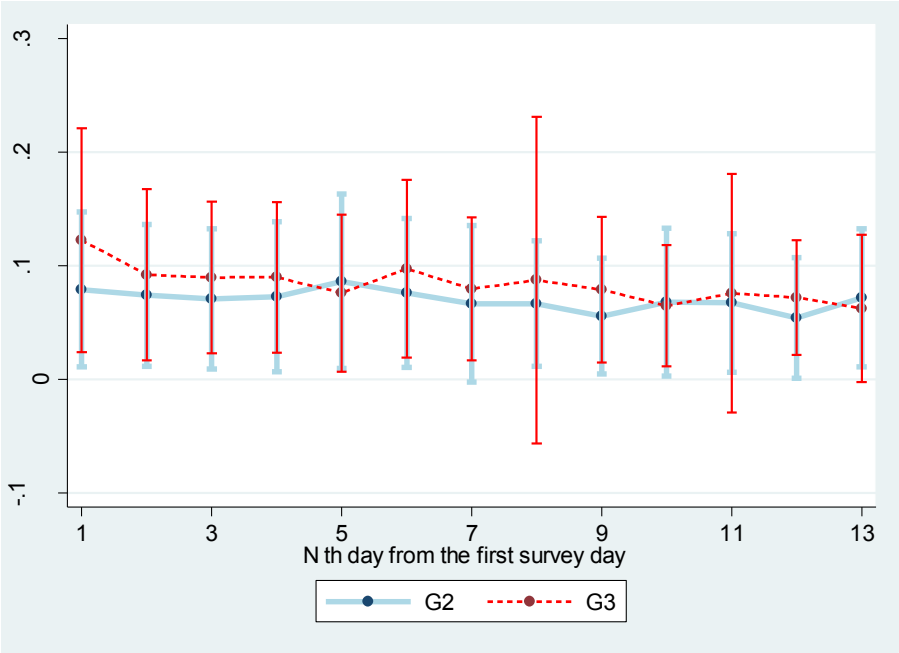
Figure A.4: Training quiz questionnaire

No.	Question	Answer (Point)
1	An important reason for conducting the census is to achieve an improvement of overall quality of health in TA Chimutu. Describe the other two reasons why we conduct the census.	a. To make it possible to reach out to every pregnant woman who wanted to participate in the AFF MCH program. (0.5) b. To enrich the stock of socio-demographic data in T/A Chimutu that is necessary for elaboration of the AFF MCH program. (0.5)
2	Regarding the roles of the enumerator, there are two functions you should NOT perform. Please fill them in the blank spaces below. A) Not to _____ B) Not to _____	a. Not to make any influence on answers (0.5) b. Not to change orders or words of questions (0.5)
3	What is the main standard required for households to be enumerated in the "2015 census of TA Chimutu," a modified version of the "population and housing census"?	Enumeration of all people, all housing units, and all other structures in TA Chimutu, who have stayed in TA Chimutu for more than 3 months during the past 12 months (1)
4	What is the name of the document that proves your eligibility to conduct the census?	Endorsement letter (1)
5	As what kind of structure would you categorize the following? *"A structure with sun-dried brick walls and asbestos roof"*	Semi-permanent (1)
6	Choose one that is <u>not</u> counted as a collective household. A) Hospitals, including three staff houses sharing food B) Lodge, including staff dwelling and sharing food C) Prison with many inmates' dwelling D) Store with owner's dwelling E) Military barracks with soldiers' dwelling	D (1)
7	What is the name of the document you have to sign before you start enumeration?	Consent form (1)
8	What are the three things you have to check before you leave the household?	Questionnaire, outbuildings, and Household ID number. (1, 0.5 point for partially correct)
9	What number do you put when you cannot meet any respondent from the household?	a. Do not put any number and just note down the household. (0.5) b. Put the latest number on it if you arrange to meet later. (0.5)
10	Your distributed alphabet is "C" and this household is the third household you enumerated in the catchment area. How did you place an ID number on the wall of the household?	0003C (1)
11	<u>True or false questions</u> A) It is okay if the questionnaire gets wet when there is heavy rain. B) You should not come to the completion meeting if you did not finish enumeration of your area. C) If you complete enumeration in your area, you should report to your supervisors immediately. D) You should bring all your housing necessities to the kickoff meeting.	A) False (0.5) B) False (0.5) C) True (0.5) D) True (0.5)

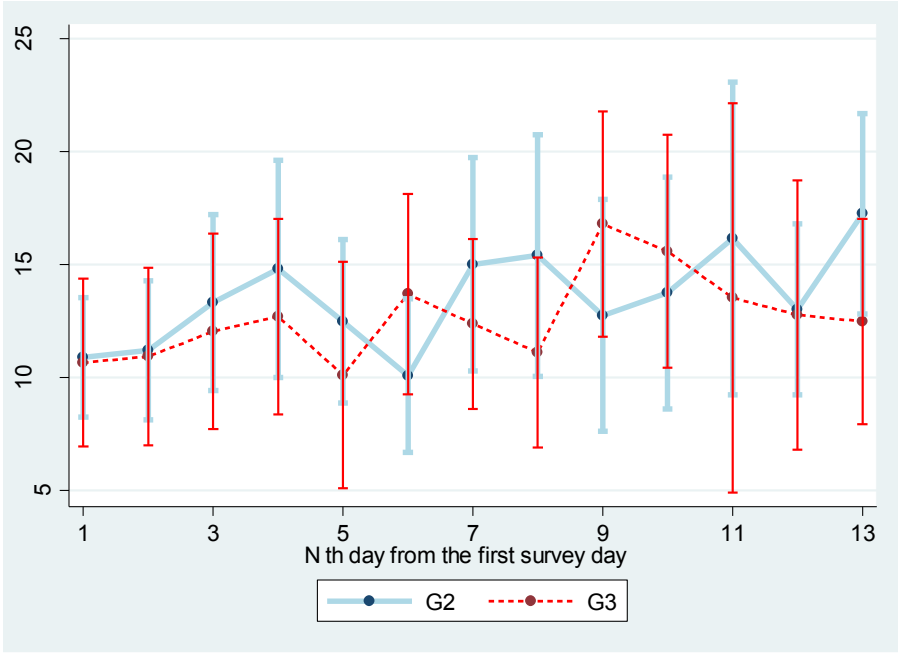
Note: The answers were not indicated in the actual training quiz questionnaire.

Figure A.5: Daily job performance trend

Panel A: Survey error rate



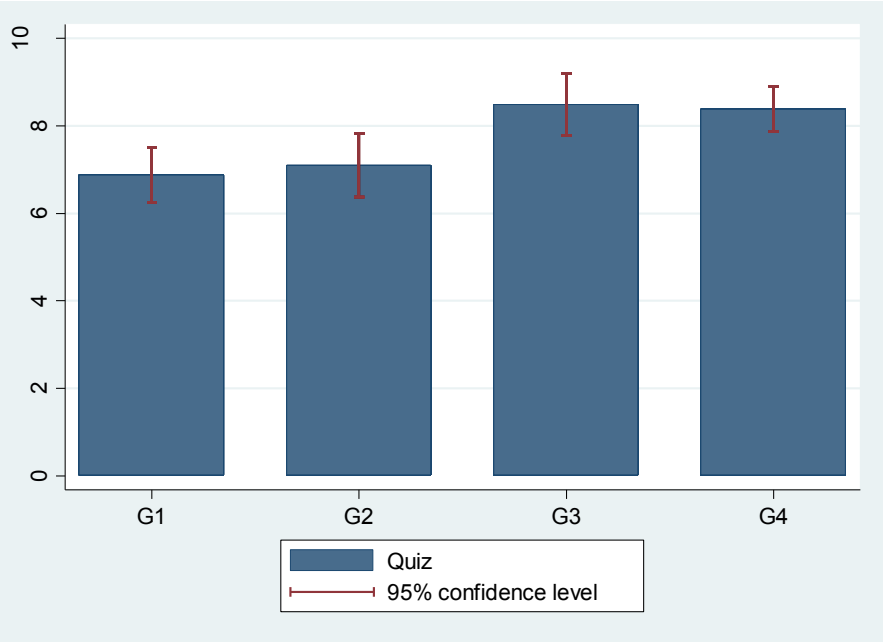
Panel B: Number of surveys per day



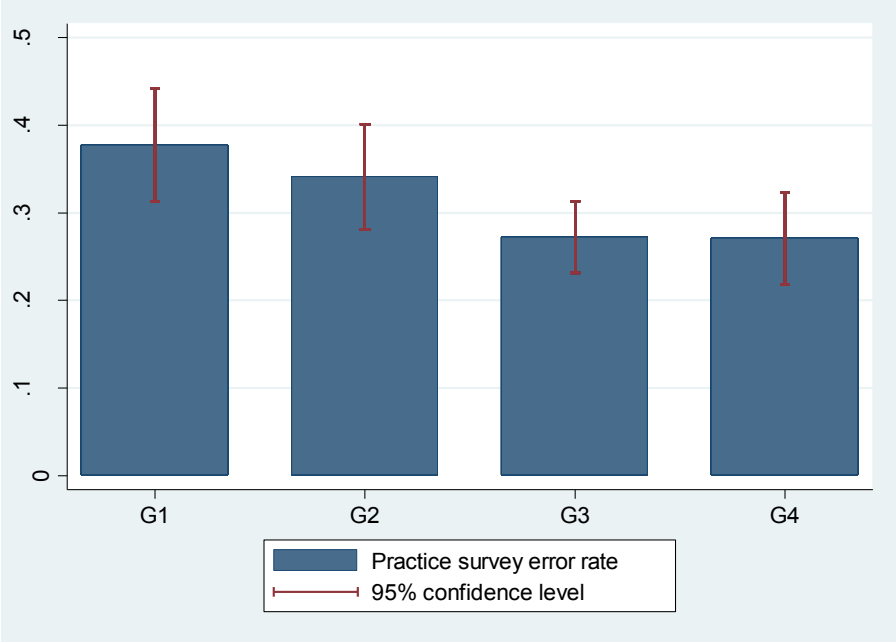
Notes: The light-blue solid and red dotted horizontal lines in each panel indicate the daily job performance of Group 2 and Group 3, respectively. The vertical lines indicate 95% confidence intervals.

Figure A.6: Training outcomes

Panel A: Quiz score



Panel B: Practice survey error rate



Notes: The maximum quiz score is 12. The vertical lines indicate 95% confidence intervals.

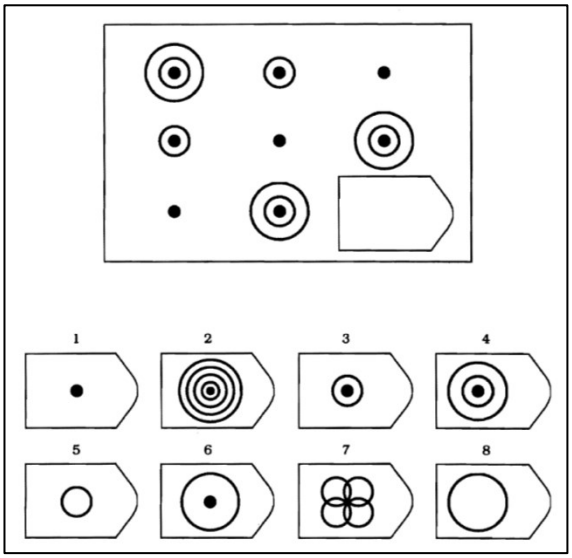
Data Appendix

Data A.1: Measurement of cognitive abilities and non-cognitive traits

A.1.1. Cognitive abilities

Raven's Progressive Matrices test

This is a widely used non-verbal test that evaluates "observation skills and clear-thinking ability" (Raven et al., 1998). Since it is independent of language skills, it is very easy to conduct in any setting including developing countries where the mother tongue is not English. The following figure is one example of the test questionnaire. In the test, a subject is required to choose one of eight options that match a missing pattern in the box. All questions follow similar visual patterns.



O*NET Ability Profiler

The O*NET Ability Profiler was originally developed by the United States Department of Labor as "a career exploration tool to help understand job seekers on their work skills (O*NET Resource Center, 2010, p. 1)". We use the verbal and clerical ability tests of the Ability Profiler, as these skills are the most relevant for the enumerator job.

a. The verbal ability test measures how well a test subject understands the definition of English words and properly uses them in conversation. The following is an example of the test questionnaire:

"Choose the two words that are either most closely the same or most closely opposite in meaning."

1. A. push
   B. dine
   C. nap
   D. eat

b. The clerical perception test measures an individual's "ability to see details in written materials quickly and correctly. It involves noticing if there are mistakes in the text and numbers, or if there are careless errors in working math problems.(O*NET Resource Center, 2010, p. 2)." The following is an example of the test questionnaire:

On the line in the middle, write S if the two names are exactly the same and write D if they are different.

1.	Paramore & Co.	—	Paramore & Co.
2.	Bimler	—	Binler
3.	E-Z Neon	—	E-Z Neon
4.	Blackstone	—	Blackstone
5.	Chris Brasch	—	Chris Grasch

Math and English scores of Malawi School Leaving Certificate Exam in 2014

All secondary school students in Malawi are required to take the Malawi School Leaving Certificate Exam during the third semester in Form 4 of secondary school (Grade 12 in the U.S.) to achieve an official secondary school graduation status. The Malawi National Examination Board (MANEB) administers the whole process of the exam. Each student chooses 6–8 subjects out of approximately 20 subjects prepared by MANEB (MANEB, 2014). Math and English are mandatory subjects. The results of each subject are reported in terms of a scale from 1 to 9. We use English and math test scores because they are mandatory subjects and thus, there are no missing values in the exam transcripts. We obtained the administrative record of the MSCE exam transcripts for all study participants through the Malawi Ministry of Education.

A.1.2. Non-cognitive traits

Rosenberg self-esteem scale

This is a 10-item scale developed by Rosenberg (1965) and is widely used to measure self-esteem by measuring positive and negative feelings about the self. All items are answered using a 4-point Likert scale format ranging from *strongly agree* to *strongly disagree*.

Intrinsic motivation

Intrinsic motivation is an individual's trait that captures whether the individual is motivated to do things by intrinsic rewards such as his/her own desire to pursue goals or challenges. It is the opposite of extrinsic motivation described below. We measure intrinsic motivation using a 15-item scale (Amabile et al., 1994). All items are answered using a 4-point Likert scale format ranging from *strongly agree* to *strongly disagree*.

Extrinsic motivation

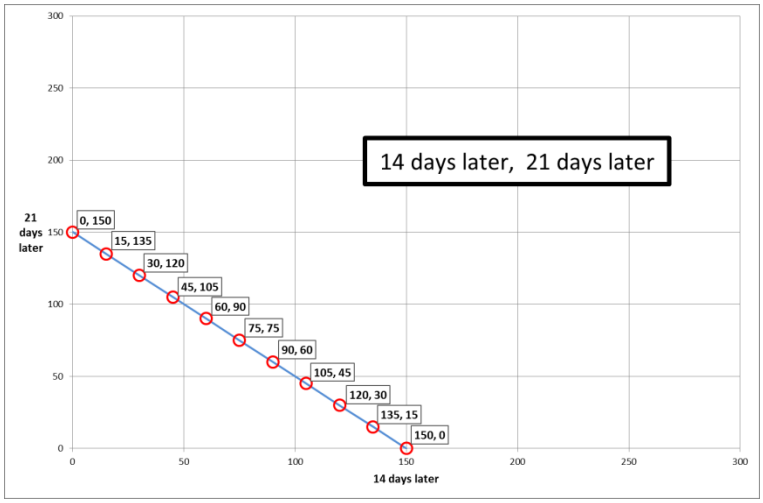
Extrinsic motivation is an individual's trait that captures whether the individual is motivated by external rewards, such as reputation, to do things. We use a 15-item scale to measure the level of motivation triggered by extrinsic values (Amabile et al., 1994). All items are answered using a 4-point Likert scale format ranging from *strongly agree* to *strongly disagree*.

Ten-item Big Five personality inventory (TIPI)

We measure an individual's personality types using a 10-item scale that assesses the respondent's characteristics based on traits commonly known as the Big 5 personality traits (openness to experience, conscientiousness, extroversion, agreeableness, and emotional stability) (Gosling et al., 2003). All items are answered using a 7-point Likert scale format (*Disagree strongly, Disagree moderately, Disagree a little, Neither agree nor disagree, Agree a little, Agree moderately, and Agree strongly*).

Time preference

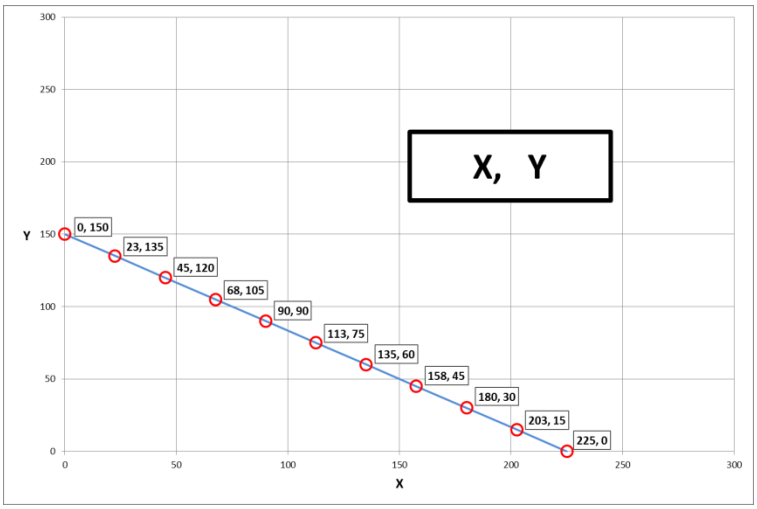
Participants were given 20 decision problems. In each, they were asked to choose 1 out of 11 options on the line. Each option [X, Y] is a payoff set indicating the amount of money (X) they would receive 14 days later and the amount of money (Y) they would receive 21 days later (see the figure below). Participants were informed that AFF would randomly choose 1 out of 20 problems and would provide the amount of payoff the participants selected in the chosen decision problem according to the payoff rule.



The choices that individuals made through this experiment were used to infer their time preference, measured between 0 and 1 following the methodology proposed by Choi et al. (2007). The closer the value is to 1, the more impatient a participant is, and the closer the value is to 0, the more patient the participant is.

Risk preference

Participants were given 20 decision problems. In each, they were asked to choose 1 option out of 11 options on the line. An option [X, Y] indicates the amount of money a participant would earn if the X-axis (the horizontal axis) was chosen and the amount of money a participant would earn if the Y-axis (the vertical axis) was chosen (see the figure below). Participants were informed that AFF would randomly choose one out of 20 problems, and again randomly choose either X or Y with equal probability, and that the chosen payoff would be provided to the participants.



The choices made by individuals through this experiment were used to infer their individual-level risk preference, measured between 0 and 1 following the methodology proposed by Choi et

al. (2007). The closer the value is to 1, the more risk-taking a participant is, and the closer the value is to 0.5, the more risk-averse the participant is. Values lower than 0.5 reflect a violation of stochastic dominance and are excluded from the analysis (Choi et al, 2007).

Rational decision-making ability

Using the Critical Cost Efficiency Index (CCEI; Afriat, 1972), we measured a level of consistency with the Generalized Axiom of Revealed Preference (GARP) based on the results from the time preference experiment. Considering all 20 decision problems in the time preference experiment, CCEI counts by how much the slope of the budget line in each problem should be adjusted to remove all violations of GARP. We took CCEI into account for the level of rational decision-making ability (Choi et al, 2014). CCEI is measured between 0 and 1. The closer CCEI is to 1, the more a participant satisfies GARP overall, and the more rational (from an economic prospective) are the decisions made.

Data A.2: Measurement of survey quality

AFF checked each questionnaire one by one and counted systematically inconsistent errors. First, census supervisors listed all possible systematic errors that could result from enumerators, not respondents. Second, data-entry clerks went through repeated training to catch those errors. Then, they started counting the number of systematic errors caused by enumerators for each sheet of the census survey.

Error collecting work was carried out in the following steps.

1. Two error-collecting data entry clerks checked one questionnaire separately.
2. They counted the total number of questions that must be answered.
3. Three types of errors from each page of the questionnaire were counted, as follows.
    1) The total number of questions that must be answered but are blank.
    2) The total number of questions that must be answered but are incorrectly answered.
    3) The total number of questions that must not be answered but are answered.
4. All the numbers on each page are added up and the total number of errors is recorded
5. The total number of errors independently counted by the two clerks is compared.
6. If the difference between the total errors counted by the two data entry clerks is larger than 5, a recount is undertaken.
7. The mean of the number of errors counted by the two data entry clerks is recorded.

The following table provides the basic statistics of each number counted.

Index	Measurement	Mean (SD)
A	The total number of all questions that must be answered	221.7 (61.8)
B	The total number of questions that must be answered but are blank	7.59 (10.3)
C	The total number of questions that must be answered but are incorrectly answered	3.90 (4.26)
D	The total number of questions that must not be answered but are answered	5.53 (9.28)

Note: *A* could be different across households due to differences in household-specific characteristics, such as family structure.

Finally, the final variable we use for survey quality (error rate) in the analysis is constructed as follows:

$$\text{error}_i = (\text{B}_i + \text{C}_i + \text{D}_i)/\text{A}_i$$

where error_i is the error rate of a specific census questionnaire i surveyed by an enumerator. A_i, B_i, C_i, and D_i are the corresponding numbers counted from the *i*-th census survey questionnaire by data clerks.

Data A.3: Imputation of missing survey beginning and end times

We find that there are significant missing values in the entries for the survey beginning time and end time of census interviews due to the enumerators' mistakes. To preserve the sample size, we impute either the survey beginning time or the end time when only one of them is missing. Specifically, we run the regression of the questionnaire-specific length of survey.

$$\text{Surveytime}_{ijklt}=\alpha+\gamma\cdot\text{H}_i+\phi\cdot\text{Z}_k+\text{V}_{lt}+\sigma_t+\psi_{ijklt} \quad (A1)$$

$\text{Surveytime}_{ijktl}$ is survey time of household *i* by enumerator *j* whose supervisor is *l*, in catchment area *k,* surveyed on the *t*-th work day. H_i is a vector of respondents' household characteristics and Z_k is a vector of catchment area characteristics. σ_t is the survey-date fixed effect. V_{lt} is the supervisor team-specific post-visit effect.

For the surveyed census questionnaire sheets with either missing start time or end time, we impute the missing time using the predicted length of a survey from the above regression. Note that we cannot use this method for an observation when both starting and ending times are

missing. In this case, we do not make any changes and thus the intermission time and survey length remain missing.

Data A.4: 2011 HIV/AIDS prevention programs of African Future Foundation

The HIV/AIDS prevention program of AFF covered 33 public schools in four districts in 2011: Traditional Authority (TA) Chimutu, TA Chitukula, TA Tsabango, and TA Kalumba. In Table A.6, the experimental design of the 2011 HIV/AIDS prevention program is summarized. The randomization process was implemented in two stages. Three types of interventions were randomly assigned to treatment groups independently. For the HIV/AIDS education and male circumcision programs, classrooms were randomly assigned to one of the three groups: 100% Treatment, 50% Treatment, and No Treatment classrooms. Treated students in the 50% Treatment classrooms were randomly selected at the individual level. The treatments were given to everybody in 100% Treatment classrooms. No one received the treatment in the No Treatment classrooms. For the girls' education support program, classrooms were randomly assigned either to the 100% Treatment or No Treatment group. AFF expected minimal spill-over between classes because there were limited cross-classroom activities and the majority (29 out of 33) of the schools had only one class per grade.

The HIV/AIDS education intervention was designed to provide the most comprehensive HIV/AIDS education. In addition to the existing HIV/AIDS education curriculum, AFF provided information on the medical benefits of male circumcision and the relative risk of cross-generational sexual relationships. The education was provided to both male and female students by trained staff members with a government certificate. The HIV/AIDS education was comprised of a 45-minute lecture and a 15-minute follow-up discussion. Study participants were assigned to one of four research groups: 100% Treatment (E1), Treated in 50% Treatment (E2), Untreated in 50% Treatment (E3), and No Treatment (E4).

The male circumcision offer consisted of free surgery at the assigned hospital, two complication check-ups (3-days and 1-week after surgery) at students' schools, and transportation support. Free surgery and complication check-ups were available for all study participants, but transportation support was randomly given. Selected students could either choose a direct pick-up service or use a transportation voucher that is reimbursed after the circumcision surgery at the assigned hospital. The value of the transportation voucher varied according to the distance between the hospital and a student's school. Study participants were also assigned to one of four research groups: 100% Treatment (C1), Treated in 50% Treatment

(C2), Untreated in 50% Treatment (C3), and No Treatment (C4). Transportation support was given to groups C1 and C2 during the study period, and the remaining temporarily untreated group (groups C3 and C4) received the same treatment one year later.

Table A.6: Experimental Design

1) HIV/AIDS Education				
	Group	Assignment	Classrooms	Students
100% Treatment	E1	Treatment	41	2480
50% Treatment	E2	Treatment	41	1303
	E3	No Treatment		1263
No Treatment	E4	No Treatment (Control)	42	2925
Total			124	7971
2) Male Circumcision				
100% Treatment	C1	Treatment	41	1293
50% Treatment	C2	Treatment	41	679
	C3	No Treatment		679
No Treatment	C4	No Treatment (Control)	42	1323
Total			124	3974
3) Girls' Education Support				
100% Treatment	S1	Treatment	62	2102
No Treatment	S2	No Treatment (Control)	62	1895
Total			124	3997

Notes: For the HIV/AIDS education and Male circumcision interventions, the randomization was done in two stages. First, classrooms for each grade across 33 schools were randomly assigned to one of three groups: 100% treatment, 50% treatment, and no treatment. Then, within the 50% treatment group, only half of the students were randomly assigned to receive the treatment.

The girls' education support program provided a one-year school tuition and monthly cash stipends to female students in randomly selected classrooms (S1). School tuition and fees per semester (on average US$7.5, 3,500 MWK) were directly deposited to each school's account and monthly cash stipends of 0.6 USD (300 MWK) were distributed directly to treated students. The total amount of scholarship was approximately US$24 per student during the study period.

Reference

Afriat, Sidney N., 1972. Efficiency Estimation of Production Function. *International Economic Review* 13 (3): 568–98.

Amabile, T.M., K.G. Hill, B.A. Hennessey, and E.M. Tighe, 1994. The Work Preference Inventory: Assessing Intrinsic and Extrinsic Motivational Orientations. *Journal of Personality and Social Psychology*, 66(5): 950.

Barrick, Murray R., Greg L. Stewart, and Mike Piotrowski, 2002. Personality and Job Performance: Test of the Mediating Effects of Motivation Among Sales Representatives, *Journal of Applied Psychology*, 87 (1), 43–51.

Choi, S., Fisman, R., Gale, D. and Kariv, S., 2007. Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97(5), pp.1921-1938.

Choi, S., Kariv, S., Müller, W. and Silverman, D.. 2014. Who is (more) Rational?. *American Economic Review*, 104(6), pp.1518-1550.

Claes, Rita, Colin Beheydt, and Björn Lemmens, 2005. Unidimensionality of Abbreviated Proactive Personality Scales Across Cultures, *Applied Psychology*, 54 (4), 476–489.

Duckworth, A.L, & Quinn, P.D,, 2009. Development and validation of the Short Grit Scale (GritS). *Journal of Personality Assessment*, 91, 166-174

Edmondson, A., 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly*, 44(2), pp.350-383.

Gosling, S.D., P.J. Rentfrow, and W.B. Swann. 2003. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, 37(6): 504–528.

Malawi National Examination Board (MANEB), 2014. Malawi School Certificate of Education Examination—Grades and Awards for Candidates, Malawi.

O*NET Resource Center, 2010. *O*NET Ability Profiler™ Score Report*, pp. 1–2

Pearlin, Leonard I.; Lieberman, Morton A.; Menaghan, Elizabeth G.; and Joseph T.

Mullan, 1981. The Stress Process. *Journal of Health and Social Behavior*, V.22, No. 4 (December): 337-356.

Radloff, L. S., 1977. The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurements*, 1, 385-401.

Raven, J., J.C. Raven, and J.H. Court., 1998. Manual for Raven's Progressive Matrices and Vocabulary Scales. The Standard Progressive Matrices, Section 3. Oxford Psychologists Press: Oxford, England/The Psychological Corporation: San Antonio, TX.

Rosenberg, Morris, 1965. Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Extra Appendix (for Reviewers)

Tables EA1-EA4 present the additional regression results of Table 5 in the main text with a variety of specifications with different sets of control variables. Specifically, Tables EA1, EA2, EA3, and EA4 report the regression results of survey quality, survey quantity, SPEs by survey respondents and supervisors, respectively. Table EA5 presents the results of Table A.3 without the imputation of survey beginning or end time.

Table EA1. Survey quality

Columns (1), (12), and (15) correspond to columns (1), (2), and (3) in Table 5 of the main text.

VARIABLES	Survey quality (error rate)														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Selection effect (G2 vs G3)															
G2	-0.020*	-0.020*	-0.022**	-0.020*	-0.020*	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**	-0.021**
	(0.011)	(0.011)	(0.009)	(0.011)	(0.011)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Observations	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130	11,130
R-squared	0.156	0.208	0.248	0.204	0.156	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302
Mean of G3 (SD)								.077 (.078)							
Panel B: Causal effect of career incentives (G3 vs. G4)															
G3	0.011	0.008	0.012	0.015	0.011	0.006	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Observations	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775	11,775
R-squared	0.181	0.202	0.210	0.233	0.181	0.265	0.273	0.273	0.273	0.273	0.273	0.273	0.273	0.273	0.273
Mean of G4 (SD)								.082 (074)							
Panel C: Causal effect of wage incentives (G1 vs. G2)															
G2	-0.028*	-0.019	-0.023	-0.033*	-0.028*	-0.019*	-0.017	-0.017	-0.017	-0.017	-0.017	-0.017	-0.017	-0.017	-0.017
	(0.017)	(0.011)	(0.015)	(0.017)	(0.017)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Observations	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779	9,779
R-squared	0.167	0.287	0.245	0.182	0.167	0.354	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357
Mean of G1 (SD)								.075 (.068)							
Panel D: Combined effect (G1 vs. G4)															
G1	-0.002	-0.006	-0.004	0.001	-0.002	-0.003	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	(0.013)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Observations	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424	10,424
R-squared	0.194	0.248	0.218	0.204	0.194	0.276	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277
Mean of G4 (SD)								.082 (074)							
Demographics	NO	YES	NO	NO	NO	YES	YES	YES	NO	NO	NO	YES	YES	NO	YES
Cognitive Ability Index	NO	NO	YES	NO	NO	YES	NO	NO	YES	YES	NO	YES	YES	YES	YES
Non-cognitive Traits	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES	YES

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings, catchment area characteristics, supervisor team-specific post visit variables, survey date-fixed effect, and binary indicator variables for previous AFF programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Training performances include the quiz score and practice survey error rate. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

Table EA2. Survey quantity

Columns (1), (12), and (15) correspond to columns (4), (5), and (6) in Table 5 of the main text.

VARIABLES	Survey quantity (number of surveys per day)														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Selection effect (G2 vs G3)															
G2	1.478***	1.396***	1.473***	1.394***	1.478***	1.407***	1.305**	1.305**	1.305**	1.305**	1.305**	1.305**	1.305**	1.305**	1.305**
	(0.516)	(0.521)	(0.515)	(0.481)	(0.516)	(0.486)	(0.546)	(0.546)	(0.546)	(0.546)	(0.546)	(0.546)	(0.546)	(0.546)	(0.546)
Observations	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003	1,003
R-squared	0.144	0.157	0.144	0.151	0.144	0.166	0.173	0.173	0.173	0.173	0.173	0.173	0.173	0.173	0.173
Mean of G3 (SD)								10.7 (5.45)							
Panel B: Causal effect of career incentives (G3 vs. G4)															
G3	-0.594	-0.604	-0.602	-0.809	-0.594	-0.867	-0.894	-0.894	-0.894	-0.894	-0.894	-0.894	-0.894	-0.894	-0.894
	(0.602)	(0.626)	(0.601)	(0.598)	(0.602)	(0.623)	(0.612)	(0.612)	(0.612)	(0.612)	(0.612)	(0.612)	(0.612)	(0.612)	(0.612)
Observations	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063	1,063
R-squared	0.149	0.167	0.149	0.162	0.149	0.185	0.189	0.189	0.189	0.189	0.189	0.189	0.189	0.189	0.189
Mean of G4 (SD)								11.5 (6.36)							
Panel C: Causal effect of wage incentives (G1 vs. G2)															
G2	1.188*	1.506**	1.150*	1.026	1.188*	1.184	1.182*	1.182*	1.182*	1.182*	1.182*	1.182*	1.182*	1.182*	1.182*
	(0.619)	(0.714)	(0.619)	(0.688)	(0.619)	(0.735)	(0.679)	(0.679)	(0.679)	(0.679)	(0.679)	(0.679)	(0.679)	(0.679)	(0.679)
Observations	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914
R-squared	0.203	0.215	0.203	0.208	0.203	0.229	0.238	0.238	0.238	0.238	0.238	0.238	0.238	0.238	0.238
Mean of G1 (SD)								9.84 (5.19)							
Panel D: Combined effect (G1 vs. G4)															
G1	-1.446	-1.554	-1.368	-1.240	-1.446	-1.351	-0.876	-0.876	-0.876	-0.876	-0.876	-0.876	-0.876	-0.876	-0.876
	(0.984)	(0.939)	(0.964)	(1.114)	(0.984)	(1.052)	(1.054)	(1.054)	(1.054)	(1.054)	(1.054)	(1.054)	(1.054)	(1.054)	(1.054)
Observations	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974
R-squared	0.157	0.193	0.162	0.181	0.157	0.221	0.225	0.225	0.225	0.225	0.225	0.225	0.225	0.225	0.225
Mean of G4 (SD)								11.5 (6.36)							
Demographics	NO	YES	NO	NO	NO	YES	YES	YES	NO	NO	NO	YES	YES	NO	YES
Cognitive Ability Index	NO	NO	YES	NO	NO	YES	NO	NO	YES	YES	NO	YES	YES	YES	YES
Non-cognitive Traits	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES	YES

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings, catchment area characteristics, supervisor team-specific post visit variables, survey date-fixed effect, and binary indicator variables for previous AFF programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Training performances include the quiz score and practice survey error rate. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

Table EA3. Subjective performance evaluation by census respondents

Columns (1), (12), and (15) correspond to columns (7), (8), and (9) in Table 5 of the main text.

VARIABLES	Subjective performance evaluation (by survey respondents)														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Selection effect (G2 vs G3)															
G2	0.783**	0.897**	0.763**	0.615	0.783**	0.691*	0.682*	0.682*	0.682*	0.682*	0.682*	0.682*	0.682*	0.682*	0.682*
	(0.387)	(0.351)	(0.378)	(0.389)	(0.387)	(0.364)	(0.382)	(0.382)	(0.382)	(0.382)	(0.382)	(0.382)	(0.382)	(0.382)	(0.382)
Observations	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473
R-squared	0.443	0.526	0.444	0.500	0.443	0.592	0.594	0.594	0.594	0.594	0.594	0.594	0.594	0.594	0.594
Mean of G3 (SD)								2.09 (1.65)							
Panel B: Causal effect of career incentives (G3 vs. G4)															
G3	0.095	0.232	0.096	0.259	0.095	0.391	0.327	0.327	0.327	0.327	0.327	0.327	0.327	0.327	0.327
	(0.368)	(0.353)	(0.368)	(0.376)	(0.368)	(0.351)	(0.346)	(0.346)	(0.346)	(0.346)	(0.346)	(0.346)	(0.346)	(0.346)	(0.346)
Observations	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233	7,233
R-squared	0.379	0.425	0.380	0.437	0.379	0.492	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499
Mean of G4 (SD)								2.08 (1.59)							
Panel C: Causal effect of wage incentives (G1 vs. G2)															
G2	0.276	0.205	0.247	0.293	0.276	0.237	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021
	(0.546)	(0.548)	(0.537)	(0.592)	(0.546)	(0.608)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)
Observations	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516
R-squared	0.389	0.465	0.398	0.504	0.389	0.607	0.656	0.656	0.656	0.656	0.656	0.656	0.656	0.656	0.656
Mean of G1 (SD)								2.67 (1.66)							
Panel D: Combined effect (G1 vs. G4)															
G1	-0.269	-0.327	-0.267	0.048	-0.269	-0.042	-0.076	-0.076	-0.076	-0.076	-0.076	-0.076	-0.076	-0.076	-0.076
	(0.474)	(0.443)	(0.476)	(0.473)	(0.474)	(0.472)	(0.552)	(0.552)	(0.552)	(0.552)	(0.552)	(0.552)	(0.552)	(0.552)	(0.552)
Observations	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276	5,276
R-squared	0.517	0.556	0.517	0.566	0.517	0.623	0.628	0.628	0.628	0.628	0.628	0.628	0.628	0.628	0.628
Mean of G4 (SD)								2.08 (1.59)							
Demographics	NO	YES	NO	NO	NO	YES	YES	YES	NO	NO	NO	YES	YES	NO	YES
Cognitive Ability Index	NO	NO	YES	NO	NO	YES	NO	NO	YES	YES	NO	YES	YES	YES	YES
Non-cognitive Traits	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES	YES

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings, catchment area characteristics, supervisor team-specific post visit variables, survey date-fixed effect, and binary indicator variables for previous AFF programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Training performances include the quiz score and practice survey error rate. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

Table EA4. Subjective evaluation of work attitude by supervisors

Columns (1), (12), and (15) correspond to columns (10), (11), and (12) in Table 5 of the main text.

VARIABLES	Subjective evaluation of work attitude (by supervisors)														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Selection effect (G2 vs G3)															
G2	-0.114**	-0.110**	-0.114**	-0.133***	-0.114**	-0.122**	-0.138**	-0.138**	-0.138**	-0.138**	-0.138**	-0.138**	-0.138**	-0.138**	-0.138**
	(0.047)	(0.055)	(0.048)	(0.045)	(0.047)	(0.057)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)
Observations	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65
R-squared	0.196	0.286	0.205	0.320	0.196	0.390	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401
Mean of G3 (SD)								.850 (.163)							
Panel B: Causal effect of career incentives (G3 vs. G4)															
G3	0.254***	0.240***	0.254***	0.248***	0.254***	0.236***	0.235***	0.235***	0.235***	0.235***	0.235***	0.235***	0.235***	0.235***	0.235***
	(0.040)	(0.042)	(0.040)	(0.042)	(0.040)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Observations	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74
R-squared	0.536	0.587	0.536	0.562	0.536	0.617	0.629	0.629	0.629	0.629	0.629	0.629	0.629	0.629	0.629
Mean of G4 (SD)								.583 (.119)							
Panel C: Causal effect of wage incentives (G1 vs. G2)															
G2	-0.133**	-0.122*	-0.133**	-0.121**	-0.133**	-0.097	-0.110	-0.110	-0.110	-0.110	-0.110	-0.110	-0.110	-0.110	-0.110
	(0.055)	(0.067)	(0.054)	(0.058)	(0.055)	(0.068)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Observations	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63
R-squared	0.261	0.353	0.261	0.310	0.261	0.422	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.444
Mean of G1 (SD)								.803 (.162)							
Panel D: Combined effect (G1 vs. G4)															
G1	0.236***	0.244***	0.236***	0.225***	0.236***	0.222***	0.180***	0.180***	0.180***	0.180***	0.180***	0.180***	0.180***	0.180***	0.180***
	(0.041)	(0.040)	(0.041)	(0.049)	(0.041)	(0.052)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
Observations	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72
R-squared	0.457	0.490	0.457	0.482	0.457	0.523	0.544	0.544	0.544	0.544	0.544	0.544	0.544	0.544	0.544
Mean of G4 (SD)								.583 (.119)							
Demographics	NO	YES	NO	NO	NO	YES	YES	YES	NO	NO	NO	YES	YES	NO	YES
Cognitive Ability Index	NO	NO	YES	NO	NO	YES	NO	NO	YES	YES	NO	YES	YES	YES	YES
Non-cognitive Traits	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	YES	NO	YES	YES
Training performance	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES	YES

Notes: Robust standard errors clustered at the enumerator level are reported in parentheses. ***, **, and * denote the significance level at 1%, 5%, and 10%, respectively. All specifications include the number of siblings, catchment area characteristics, supervisor team-specific post visit variables, survey date-fixed effect, and binary indicator variables for previous AFF programs. Individual characteristics include age, asset score, cognitive ability index, and a set of non-cognitive traits (self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items). Training performances include the quiz score and practice survey error rate. Catchment area characteristics include the total number of households, family size, asset score, number of births in the last 3 years, incidence of malaria among children under 3, and deaths in the last 12 months.

Table EA5. Analysis in Table A.3 without the imputation of survey beginning or end time

VARIABLES	Survey quantity					
	Work hours (in mins)		Survey time per household (in mins)		Intermission time between surveys (in mins)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Selection effect (G2 vs G3)						
G2	1.309	-3.185	-1.189	-0.900	-4.155	-3.754
	(18.006)	(17.153)	(1.166)	(0.991)	(2.554)	(2.322)
Observations	987	987	9,966	9,966	8,057	8,057
R-squared	0.151	0.180	0.268	0.305	0.019	0.029
Mean (SD) of G3	419.4 (196.5)		25.3 (12.2)		22.7 (49.3)	
Panel B: Causal effect of career incentives (G3 vs. G4)						
G3	43.399**	35.039*	0.917	1.264	6.337***	5.590***
	(16.915)	(18.480)	(1.224)	(1.173)	(1.919)	(2.049)
Observations	1,053	1,053	10,908	10,908	8,937	8,937
R-squared	0.147	0.170	0.237	0.253	0.019	0.026
Mean (SD) of G4	386.4 (193.1)		23.9 (11.4)		17.1 (43.7)	
Panel C: Causal effect of wage (G1 vs. G2)						
G2	21.996	27.377	-2.752*	-1.854	1.535	0.613
	(24.345)	(23.318)	(1.562)	(1.574)	(3.141)	(3.495)
Observations	887	887	8,789	8,789	7,056	7,056
R-squared	0.187	0.219	0.293	0.327	0.022	0.033
Mean (SD) of G1	382.3 (187.1)		27.2 (12.4)		19.0 (41.7)	
Panel D: Combined effect (G1 vs. G4)						
G1	-12.771	-24.533	2.307	0.258	2.420	1.688
	(25.596)	(27.664)	(1.462)	(1.504)	(2.306)	(2.211)
Observations	953	953	9,731	9,731	7,936	7,936
R-squared	0.157	0.188	0.269	0.318	0.015	0.024
Mean (SD) of G4	386.4 (193.1)		23.9 (11.4)		17.1 (43.7)	
Individual characteristics and training performance	NO	YES	NO	YES	NO	YES

Notes: ***, **, * denote the significance level at 1%, 5%, and 10% respectively. All specifications include number of siblings, number of family members, catchment area control variables, supervisor fixed effect variables, survey date fixed effect variables, and binary indicator variables of whether an individual received a male circumcision offer, HIV/AIDS education, and/or he belonged to the same class of female scholarship beneficiaries, respectively. Demographics include age and asset score. Cognitive ability index is the average z-score of the Raven's matrices test score, the math and English scores of the 2014 Malawi School Certificate of Education (MSCE) test, and the verbal and clerical ability test scores of the O*NET ability test. Non-cognitive traits consist of self-esteem, intrinsic and extrinsic motivation, and Big 5 personality items. Catchment area control variables include the total number of households, asset score (whether to own refrigerator, bicycle, and improved toilet), birth rate in the last 3 years, incidence of malaria among under 3, and deaths in the last 12 months. Supervisor fixed effect variable is multiplication of dummy variables for each supervision team and a dummy variable for surveys after the supervisor visit. Survey date fixed effect variable is a dummy variable of each day from the beginning.