Robustness and External Validity: What do we Learn from Repeated Study Designs over Time?

Adedoyin Adesina
Modibbo Adama University of Technology, Yola

Jed Friedman
World Bank

Oladele Akogun
Modibbo Adama University of Technology, Yola

Sani Njobdi
Modibbo Adama University of Technology, Yola

Andrew Dillon
Michigan State University

Pieter Serneels
University of East Anglia

December 2017

Abstract: The replication of studies is foundational to the scientific method. But replication can have different meanings varying from results verification, to reproduction, to reanalysis with an alternative specification. Few studies are repeated in the same setting. Yet this may be key when environmental or other mediating factors change over time, as it may help establish validity over time or shed light on why impact changes over time. One example is the varied impact a health intervention can have, particularly when targeting infectious diseases, as disease incidence may vary across years. This paper presents the findings of a study that repeats the same health intervention in the same site in three different years, estimating the effect of malaria testing and treatment on agricultural worker earnings, labor supply and productivity. We find a significant impact on worker earnings across the years, but the impact size varies over time. The treatment on the treated estimates are lower in a year when the malaria prevalence rate is low. The 'treatment on the medically untreated', which captures an information and behavioral effect identified in an earlier study, is smaller in years when prevalence is lower and the possibility of substituting into and out of lower effort, lower return tasks is absent. These results underline the importance of changes in the prevalence rate as well as the worker's labor constraints. The results demonstrate that repetition, apart from providing a useful tool for validation, can also help shed light on reasons why effects may vary over time.

Acknowledgements: We would like to thank Bashir Aliyou and the Nigerian Ministry of Health, Malaria Control Program, especially Dr. Toulope Olayemi Sofola for helpful comments and close collaboration. The plantation staff and sugarcane cutters welcomed us and patiently assisted us throughout the study, without which this study would not have been possible. Sarah Kopper, Ashesh Prasann and Francis Smart provided excellent research assistance. Practical Sampling International undertook this work in the field and we appreciate the efforts of the supervisors, enumerators, and health workers who implemented this study. This study was ethically reviewed by the National Health Research Ethics Committee, Nigeria (NHREC 01/01/2007-28/10/10/2009c), Adamawa State Health Research Ethics Committee, Michigan State University's Institutional Review Board and IFPRI's Institutional Review Board.

INTRODUCTION

The replication of studies is foundational to the scientific method, but replication studies can have distinct objectives that vary from results verification, to reproduction, to re-analysis with different specification. Results replication studies have increased in recent years, with varied outcomes. Camerer et al. (2016) re-analyze 18 lab experimental studies published in two leading economics journals between 2011 and 2014 and found that they replicated quite well: for 61% of the studies they found a significant effect in the same direction and the overall replicated effect size was 66% of the original. Chang and Li (2015), on the other hand, replicate only 49% of the macro-economic US studies, drawn from 59 publications in prominent economics journals.[1] Both these studies are examples of pure replication of the authors' original specification using the authors' original data. However, the concept of replication is not always clearly defined, especially when external validity is one of the objectives. Clemens (2017) distinguishes between replication and robustness based on the similarity or difference of the sampling distribution of parameter estimates respectively. Robustness analysis, particularly when the study population and samples change either over time or place, rather than altering a specification with the same sample, provides evidence on the external validity of a study. Few studies are repeated to establish robustness or external validity, yet in health interventions, particularly those targeted at infectious diseases, environmental variability can affect disease incidence in any given year. Randomized control trials implemented in particular socio-economic contexts are often criticized as lacking external validity (Deaton (2009), Rodrik (2009)) due to varying effects across space or over time.

External validity is frequently defined as the ability to generalize study results to other sites or contexts. An emerging literature has investigated external validity across different implementers (Bold et al. 2013, Das et al. 2017) and difficulty in extrapolating study results across study sites (Allcott 2015, Flores and Mitnik 2013, Gechter 2016, Hotz et al. 2005 and Stuart et al. 2011). In health studies, external validity may further be limited due to changes over time that can affect the magnitude of impact estimates. Underlying all of these concerns is the possible existence of an unobserved mitigating factor(s) that partially determine the outcome of interest. In a one-off evaluation, every study unit is exposed to this factor – it essentially constitutes part of the context

[1] They focused on studies with a key empirical result produced by inclusion of US gross domestic product.

of the experiment – yet it may interact with treatment to yield an estimated effect size specific to that point in space or time. Ignoring this factor will yield biased estimates of impact when extrapolated to other settings.

This paper uses repeated implementation of the same study design to investigate the robustness of the results over time. We estimate the effect of malaria testing and treatment on agricultural worker earnings, labor supply and productivity over three different years. We find a strong impact on earnings across the three years. The effect size estimates also indicate a high degree of heterogeneity across rounds. First, the treatment on the treated estimates are smaller in a year where the malaria prevalence rate is low. Second, the treatment on the medically untreated estimate, which reflect an information and behavioral effect, is smaller in years when prevalence is higher and it is not possible to substitute into and out of lower effort, lower return occupations. The heterogeneity in treatment on the treated estimates demonstrate the importance of temporal variation in a key contextual factor in many evaluations of health programs – disease prevalence. The heterogeneity in treatment on the medically untreated estimates demonstrate that variation in disease prevalence as well as temporal variation in a key institutional arrangement governing worker occupational choice have a strong influence on estimates of program impact. All told, repeating the same intervention in the same site over time has shed light on key contextual factors influencing the estimated relationship between malaria and worker behavior.

The next section of the paper sketches a theoretical framework for replication when contextual factors mediate the effect of an intervention – it takes repeated implementation of the study in different sites or times to generate evidence on the influence of these mediating factors. Section three describes the study design. The fourth section outlines the econometric strategy, which closely follows the original analysis (Dillon et al. 2014). Section Five presents the results, while the last section concludes.

A FRAMEWORK FOR ROBUSTNESS AND EXTERNAL VALIDITY IN THE PRESENCE OF MEDIATING FACTORS

The presence of unobserved factors, fixed within the evaluation site but varying across sites or time, may generate context-dependent treatment effect heterogeneity if these factors interact with

treatment. Das et al. (2017) summarize an extension of the Rubin causal model which provides a framework for understanding how the influence of such spatial or temporal factors can mediate impact estimates and consequently confound extrapolation to other settings. Replicating the intervention in other sites or times can illuminate the influence of such factors or even, in situations with little a priori contextual knowledge, identify previously unknown influential factors.

Within the general Rubin causal framework, let us define the treatment effect of an intervention, τ_i as the difference in outcome Y observed for individual i, both with and without treatment:

$$\tau_i = Y_i(T = 1) - Y_i(T = 0)$$

where $T\epsilon[0,1]$ indicates treatment status. Further, let $D(D\epsilon[0,1])$ indicate whether individual i resides in the evaluation site ($D_i = 1$), sometimes also termed the sample or trial site, or the target site ($D_i = 0$).[2] For our purposes, the estimated program impact is said to be externally valid if the mean causal effect of the program can be replicated in the target site without error. This unbiased extrapolation relies on an assumption of external unconfoundedness,

$$D_i \perp (Y_i(1) - Y_i(0))|F$$

which posits that site allocation is orthogonal to the program treatment effect, conditional on a set of factors F. This set of factors includes, in the most general formulation, all relevant characteristics both observed by researchers and those unobserved (X_O and X_U respectively) that mediate the treatment effect.

Following Imai et al. (2008), extrapolation error is defined as the degree of divergence between the sample treatment effect and the target site treatment effect. We denote this difference as Δ:

$$\Delta = E[\tau_i|D_i = 1] - E[\tau_i|D_i = 0]$$

If the impact estimate in the trial site has full external validity then $\Delta = 0$. However Δ may be non-zero due to differences across sites in the values of individual elements in F_i. When $\Delta \neq 0$, we can

[2] Site can refer to two points in space or two points in time at no loss of generality.

approximate Δ by linearly projecting the error into the constituent components of F_i in the following manner:

$$\Delta = f(X_o, X_u) \cong \Delta_{XO} + \Delta_{XU} + possible\ interaction\ terms$$

where Δ_{xo} and Δ_{xu} are the portion of total error due to differences in observed and unobserved characteristics between the trial population and the target population. A full projection also allows for all possible interaction terms between the two components.

Existing methods of site-specific imputation of the treatment effect, largely based on Hotz et al. (2005), attempt to minimize bias due to Δ_{xo} where data permits. This extrapolation is accomplished by balancing the observed characteristics of the target site (again, where data permits) in order to match the characteristics of the trial site. Typically this is realized through the application of propensity scores used as weights in the standard mean impact estimator.

The above formulation underscores the importance of measuring all relevant characteristics that (a) vary between the two sites and (b) mediate program impact. However certain characteristics are contextual – i.e. fixed for all units in the site – and often unobserved. Examples of such fixed characteristics can include the local institutional arrangements in the study site relevant for the generation of the outcome of interest. Where information on these fixed characteristics are lacking, even the propensity weighted estimator is likely to suffer from external confoundedness:

$$If\ \Delta_{XU} \neq 0 \rightarrow E[\tau_i|D_i = 1, X_O] - E[\tau_i|D_i = 0, X_O] = \Delta$$
$$\cong \Delta_{XU} + possible\ interaction\ terms$$

The inferential challenge of extrapolating program impacts in the presence of key unobservables is exemplified by the program analyzed here. The evaluation assesses an identical malaria program implemented in the same study site in three different years. However certain key mediating features found in the setting vary over time and likely account for at least some fraction of the cross-year difference in program estimates. This is a cautionary tale for case-study evaluations or

other relatively small-scale program evaluations. It is likely that missing (or non-varying) information on context limits the generalizability, and hence policy relevance, of any one study.

Given the general interest in evaluation research and the demand for such research among policy makers, it is important to highlight with this study: (1) the restrictiveness of the external unconfoundedness assumption, and (2) the likely benefit in evaluative research to assessing contextual factors F. Even if important elements of F are fixed in the study site, replication of the study in additional sites with contrasting values of F will yield a more generalizable estimate of program impact as well as how important content mediates such impact.

STUDY AND REPLICATION DESIGN

The experiment is situated on a single large (5,700 hectares) sugar cane plantation in rural Nigeria. The plantation employs sugarcane cutters who work for the entire harvest season that stretches from mid-November to April. Cane-cutters are paid a piece rate wage. While there are other activities on the plantation, including a sugar processing facility, this study focuses solely on the sugarcane cutter labor force.

Workers are hired from local villages surrounding the plantation and are transported daily to the assigned work site. This organized transportation also serves to standardize the number of hours of cane cutting work across all workers. The cane-cutters are organized into work groups, averaging between 80-100 workers per group depending on the year. Each group is managed by a supervisor. Every day the supervisor and his cutters are assigned a set of starting fields in the plantation and additional fields to cut when they have finished with their starting fields. Sugarcane cutters do not work in teams to complete the rows of cane but rather work individually along a row, allocated to them by the supervisor, until finished and are then assigned to another row to harvest. Rows of cane are typically of uniform density due to mechanized planting and the irrigated nature of sugarcane that requires fields to be encompassed with water canals. [3] Due to the size of

[3] This also means that the malaria season is not cyclical, as it is in rain fed production areas. The existence of the water canals throughout this large plantation creates larval breeding that is uncorrelated with rainfall patterns.

the plantation and the capacity of the processing factory on site, workers may cut as much cane as they can within a given work day. One worker's productivity does not impact other workers' earnings as additional cane rows, or even nearby fields, are always made available when workers have finished their row or field.[4]

Cane cutters are paid a piece rate for every measured "rod" of cane cut where a "rod" (approximately two meters in length) is a physical standard carried by every work group supervisor. At the end of each day, the worker's output for that day is entered on his personal 'blue card' and is signed off by both the supervisor and worker. The plantation thus keeps records of the daily output (quantity cut), the days worked, and the total earnings for each worker. Laborers are paid monthly and they often keep track of their daily output by maintaining their own separate ledger.

We supplement the worker productivity information with data from worker interviews covering socio-demographic, work history, and self-reported health information. We also collect blood samples during the interview to test for malaria. The experimental design randomizes the order in which workers are tested and treated over time with all workers receiving one test (and medical treatment if positive) over the survey period of six weeks. The study then exploits the exogenous variation in the timing of access to testing and treatment for malaria to identify the effects of the intervention.[5] To do this we construct a time-series of worker-week observations that permits us to compare the labor outcomes of treated and untreated workers for the same weeks of observation.

[4] In other words, this setting is not a zero-sum game, as there is no risk of shortage of cane to cut for any individual worker. Indeed, at the end of the season the firm typically brings in (old) machines to cut the remaining cane that workers have not been able to cut in time (i.e. during the window of optimal ripeness). Any worker's future earnings potential is therefore unaffected by another worker's productivity.

[5] A free health clinic to which workers have access already exists on the plantation but we do not expect the presence of this clinic to confound our impact estimates. The clinic is believed by the work force to be of poor quality. There is no patient follow up and the facility is far removed for most workers. Virtually no worker reported a visit to the clinic during the fieldwork period, and this was confirmed through inspection of the clinic's records. Malaria care outside the plantation is generally low quality with limited diagnostic capacity during the period of our study.

Our study diagnoses malaria by measuring parasites in the worker from thick film blood smears read in a dedicated laboratory. Although expensive to implement as it requires trained personnel and appropriate instruments, thick blood film microscopy is considered the diagnostic gold standard. In practice, our study team takes a blood sample from each consenting worker and conducts microscopy analysis in a lab 2 hours away (by car) from the plantation. The microscopy analysis counts the number of parasites, with workers above a specified threshold considered to be malaria positive.[6] Over the three survey rounds, malaria diagnosis was carried out at the same laboratory with the same laboratory supervisor with the additional oversight of one of the study's authors. The clinical diagnostic standard was the same as recommended in the study area. As there is no universally adopted standard for both symptomatic and asymptomatic cases in a population, we rely on the clinical threshold in our study area as our objective measure following the recommendation in WHO (2010).[7]

The study design was the same in all three years in terms of sampling, randomized assignment to treatment, survey instrument, and clinical diagnosis which we describe in detail below. There were also some environmental differences across the years, outside the control of the study team. Table 1 describes differences across rounds. Three differences stand out. First, the prevalence of malaria varies across the years, from 36% in the first year, to 14% in year two and 21% in year three of the study. These are obtained from identical diagnostic protocol, and the same teams of health workers across the years. Second, work options were different in the first year, compared to year two and three. For all three rounds, sugarcane cutters were paid on a piece rate basis and that piece rate was consistent across rounds. Third, an unusual feature in the 2010 round is that, at the start of every day, workers have the choice of two daily tasks – sugarcane cutting or 'scrabbling'. Scrabbling is an occupation that includes the collection of cut sugarcane rods and then binding of them into bundles for loading on trucks destined for processing at the factory. Less physically intensive than cutting, and more difficult to observe individual effort, scrabbling pays a fixed wage

[6] A professional laboratory technician read all the slides to record the number of parasites in five viewing fields. After recording the parasite count, the laboratory supervisor selected random subsamples of slides to verify from each batch of 50 slides. If discrepancies between the primary laboratory technician and the supervisor were found, the whole batch of slides was re-validated.

[7] Several studies in the medical literature from different settings use distinct parasite density thresholds in classifying malaria infections as there is no unique medically established standard for population based malaria testing which includes asymptomatic malaria cases (see dalla Martha et al. (2007), Toure et al. (2006), and Rottmann et al. (2006)).

of 500 Naira per day (roughly half the expected earnings of a day spent cutting). In the first year, scrabbling work can be selected by a cane cutter at any day through a request to the supervisor at the start of the day. There is also a dedicated separate work force of scrabblers hired and managed by the plantation but these full-time scrabblers are not part of this study. While cane-cutters choose to scrabble only infrequently, the amount of time devoted to scrabbling in year 1 is not trivial – the average cane-cutter spends 3.5 days of the week cutting cane and 0.5 day scrabbling (with the other days of the week spent off the plantation, either inactive or in agriculture and household related activities).

Another across rounds is an unforeseen event that limited the number of cane cutting days in round 3. In that year, a sugar processing factory on the plantation had a mechanical failure that required sugarcane cutters to stop working until it was repaired. Sugarcane has to be processed within the day it is cut to maximize sugar yield. This interlinkage of factory and cane cutting labor meant that the factory failure stopped workers from cutting activity for 2 weeks.

There is also variation in the size of the workforce and distribution of workers across years. Due to management changes across the three survey rounds, the worker population differs by round as well as the number of returning workers who have experience on the plantation. In response to the changes in worker size, work group composition differs by round as well as the specific supervisor that worked with a given worker in each round. The sample composition, group size and supervisory structure could all result in differences in treatment effect estimation, but presumably a group by week fixed effects strategy would largely address concerns of work group composition and supervisory effects across rounds.

Data collection teams changed slightly in composition and supervisory structure across survey rounds, though our clinical diagnostic team was similar across rounds.

Table 2 presents selected mean individual and household characteristics of the workforce to illustrate where sample composition may or may not affect treatment effect estimation.[8] Workers are exclusively male and generally of prime age. The work force is getting older by round, which is expected as workers remain with the firm across rounds, but does potentially affect worker productivity as they age. The workforce's experience at the plantation varies between 4.3 years in 2010 and 6.2 years in 2013, suggesting that more experienced workers do remain with the plantation along with some worker turnover each year. Worker BMI decreases by one point between the 2010 and 2013 rounds from 23.8 to 22.5. Household characteristics of workers also vary between rounds with household size, asset holdings and imputed monthly expenditure all varying significantly, particularly between rounds. Average daily earnings range between 1,052 Naira (2010), 910 Naira (2011), and 1,064 Naira (2013) across seasons. The average harvest season comprises 66 workdays in 2010, 70 workdays in 2011, and 27 days in 2013. The average worker in round 1 elects to spend 17% of the work season as a scrabbler, with the remainder devoted to cane cutting. In subsequent seasons, cane cutters were no longer allowed to scrabble.

Workers health seeking behavior varies between rounds as we report in Table 3. Fever and morbidity self-reports are relatively low over the study rounds, reaching the highest level of reporting in the 2011 round, but decline to the lowest reporting level in 2013. These statistics are strongly correlated with malaria positive diagnoses from the study clinic and are statistically different across rounds.

[8] These characteristics include household expenditure which are not measured but rather predicted using the method suggested by Grosh and Baker (1995) and Ahmed and Bouis (2002). In our questionnaire we included questions on asset ownership drawn from the Nigerian Living Standard Survey 2009, a nationally representative survey, conducted by the National Bureau of Statistics (NBS), which collects detailed data on household consumption and expenditures. We run the weighted regression $Exp_i = \sum_{a=1}^{p}(\alpha^a D_i^a + u_i)$ on the NLSS 2010 data to obtain estimates of $\widehat{\alpha^a}$, the coefficient for each asset, which we then use to predict EXP_i for our own sample. Where D_i^a represents a dummy variable indicating whether asset a is present in the household. The regression uses population weights as calculated by the NBS. Since the estimates of the coefficients are relatively sensitive to outliers, we exclude the richest 10% of households in our weighted regression on the NLSS 2010 data.

ECONOMETRIC STRATEGY

As in the original analysis (Dillon et al. 2014), we estimate three types of treatment effects for the offer of malaria testing and treatment: an 'intent to treat effect' (ITT), a 'treatment on the treated' effect (TOT), and a 'treatment of the medically untreated' effect (TmUT). The first effect reflects the benefits of access to malaria testing and treatment, comparing outcomes of workers *with* access to testing and treatment to those of workers *yet without* access to testing and treatment (and who may or may not have fallen ill from malaria). The second effect compares outcomes of those who are ill and treated to those who are ill but not yet treated due to their randomly allocated later testing date. The third effect considers the sole effect of health information on labor outcomes (operating presumably through the mechanism of updated health perceptions) for those workers who test malaria negative. We do this by comparing labor outcomes for those workers who are tested and informed to be malaria negative with those workers not yet tested but assumed negative based on the results of subsequent tests. This effect can be thought of as a Treatment on the Treated estimate of the information component of the intervention where healthy workers learn about their actual good health, a potential 'good news' effect. However to distinguish these estimates from the Treatment on the Treated for those workers who are malaria positive, we adopt the TmUT appellation, following the original paper. As a robustness check, we present several different estimates of these effects using different durations for the observation reference period. We now discuss each of these estimates in more detail.

The ITT effect is estimated by comparing labor outcomes over some observation period at the weekly level, *t*, for those workers who were tested at time *t*-, a period before the observation period *t*, with the labor outcomes for workers who are tested at *t*+, after the observation period *t*. The sets of workers assessed at *t*- and *t*+ are denoted as W_{t-} and W_{t+}. The difference in outcomes over period *t* represents the combined effect of testing and treating for malaria, as it compares the output of a randomly selected subsample of workers who are tested with a randomly selected subsample of worker yet to be tested. Specifying the weekly level of aggregation provides two fixed effects approaches and different potential outcomes on which the ITT could be estimated, varying the interpretation of the treatment effect.

We estimate the ITT at the weekly level which permits identification of a productivity effect in addition to earnings and labor supply estimates of malaria testing and treatment. To control for the potential non-random placement of workers across work groups, as well as the natural weekly variation in work outcomes both across and within workgroups, a full set of workgroup-workweek fixed effects, F_{gt}, are included in the specification. Specifically we estimate:

$$L_{igt,r} = \alpha + \beta^1 T_{igt-,r} + \beta^2 R_r + \beta^3 (R_r \times T_{igt-,r}) + F_{gt,r} + \varepsilon_{it,r}, \forall i \in W_{t-} \cup W_{t+} \quad (1)$$

where $L_{igt,r}$ measures the three labor outcomes of interest in log form: weekly earnings, weekly labor supply and weekly productivity for worker *i* in work group *g* at period *t*, and $\varepsilon_{it,r}$ is the worker specific error term. β^1, in equation 2, captures the effect of a change in perceived health in parasitemic negative workers and the combined effect of a change in actual health (as a result of treatment) and the provision of more accurate information about the worker's health status in parasitemic positive workers. Note that the content of information is distinct for the two groups. The ITT thus reflects a combined effect of good news for the parasitemic negatives and bad news and medical treatment for the parasitemic positives.

Following a similar approach, the TOT on malaria positives is estimated by comparing labor outcomes at time *t* for those workers who had access to treatment at time *t*- and were treated if ill (and are therefore healthy over the period *t*) with the labor outcomes for workers who were not tested until time *t*+ but at that point found to be malaria positive (and thus assumed sick over the period *t*). To estimate the TOT, Equation 1 and 2 are re-estimated but now for the subset of workers P who have tested positive, as given in Equation 3 for the daily, worker fixed effects specification and Equation 11 for the weekly, work group by week fixed effects specification:

$$L_{igt,r} = \alpha + \beta^1 T_{igt-,r} + \beta^2 R_r + \beta^3 (R_r \times T_{igt-,r}) + F_{gt,r} + \varepsilon_{it,r}, \forall i \in P_{t-} \cup P_{t+} \quad (2)$$

as before, $L_{it,r}$ reflects the log labor outcomes of interest: daily labor supply, and daily productivity where $L_{igt,r}$ reflects the log labor outcomes of interest: weekly earnings, weekly labor supply, and

weekly productivity. The TOT reflects the combined effect of receiving an illness diagnosis and medical treatment for such a diagnosis.

Finally, we estimate a possible 'good news' effect by comparing labor outcomes at time *t* for those workers who were tested and found negative at time *t-* with the labor outcomes for workers who were not tested until time *t+*, but found to be negative at that point. This is estimated for the subset of workers N who have tested negative, as given in Equation 5 for the daily, worker fixed effects specification and Equation 6 for the weekly, work group by week fixed effects specification:

$$L_{igt,r} = \alpha + \beta^1 T_{igt-} + \beta^2 R_r + \beta^3 (R_r \times T_{it-,r}) + F_{gt} + \varepsilon_{it}, \forall i \in N_{t-} \cup N_{t+} \quad (3)$$

These specifications are only adjusted from our previous 2011 round analysis (Dillon et al. 2014) through the inclusion of round and interaction variables. This permits us to test whether treatment effect heterogeneity is significant and reject the null hypothesis that the coefficient is zero.

Table 4 presents the observable characteristic balancing tests by worker-group weeks. Panels A, B, and C correspond to each survey round from 2010, 2011, and 2013. Balancing tests are conducted by worker-group weeks because this was the randomization strategy described above that allocated worker interviews for testing and treatment. Worker characteristics such as age, years of experience, years of schooling, and BMI are balanced across worker group weeks by round. Of the 92 tests for these observable worker characteristics, we reject the null hypothesis of mean equality in 8 tests at the 10% level of statistical significance. We also conduct balancing tests on household characteristics of workers including household size, number of rooms in the worker's house, number of cattle, number of poultry and imputed monthly per capita expenditures. These demographic and wealth proxies help demonstrate that workers who were treated either earlier or later in our study are not systematically different with respect to these measures of wealth and household size. Of the 115 balancing tests, the null hypothesis is rejected in 15 of the tests at the 10% level of statistical significance.

We also conduct balancing tests on worker health and health behavior variables by worker group week. The p-value of the equality of means test is presented for each variable in Table 5. We find balance within round and in the pooled sample. In the first round, six tests out of 48 tests are rejected at the 10% level of statistical significance, with 4 rejected in the second round and 8 rejected in the third round. These statistics are consistent with the statistically expected number of rejections at the 10% level of statistical significance. The results demonstrate that the malaria rate across study weeks and the likelihood of fever and morbidity self-reporting and health seeking behavior does not vary systematically within worker groups by week.

RESULTS

Tables 6, 7, and 8 present the intent to treat, treatment on the treated, and the treatment on the medically untreated respectively. All tables present the weekly group by week fixed effects specifications described above for the two and three week reference periods. We have the most confidence in these reference periods, because the one week results might be contaminated by variation in the return of test results after interview which may mean that workers are only partially treated in week one. The four week reference period, on the other hand, results in a truncated sample and requires stronger assumptions about the counterfactual's malaria status which may be less plausible as the reference period increases. The appendix provides the results for the one and four week reference periods.

The intent to treat results in Table 6 estimate a robust treatment effect for both the two and three week reference periods. The weekly earnings estimates range from 10.8% and 13.4% for the two and three week reference respectively, while the weekly labor supply estimates vary between 4.5% and 4.9%. The productivity estimates vary between 6% and 9% for the two and three week reference periods. The intent to treat estimates establish a robust treatment effect across specification choice and reference period for the pooled sample of the three rounds. Turning to the round and round by treatment interactions, they provide estimates of the variation in treatment effects on earnings, labor supply and productivity by year. In both the two and three week references, the round effects are uniformly statistically significant which demonstrate round effects

do affect the levels of the outcome variables. Tests of the equality of the round 2 and 3 effects are rejected at the 1% level of statistical significance with the exception of the 3 week reference period's earnings effect which is rejected at the 10% level of statistical significance. As documented above, differences across rounds could be due to differences across rounds with variations in malaria rates, labor arrangements and interruption of the work.

These round variations do affect the level of the outcome variable, but we also estimate statistically significant treatment effect interactions for round 2 and 3. In round 2, the earnings effect is half the pooled estimate for the two week estimate (5% in round 2 compared to 10.8% for the pooled estimate) and two-thirds less for the three week estimate (8.7% in round 2 compared to 13.4% for the pooled estimate), though the interaction is not statistically significant. The round 2 interaction for the productivity estimate suggests that the decline in the earnings effect is primarily due to a decline in productivity in round three, as the statistically significant treatment effect interaction in round 2 is similar and significant at the 10% level of statistical significance. There are no labor supply treatment effect-round interactions for round 2. Differences in the estimated intent to treat effect on earnings are due to changes in productivity between rounds 1 and 2.

In round 3, the interaction suggests there is no earnings in round 3, again primarily though not entirely due to differences in productivity across rounds. The marginal ln earnings effect for round three in the two and three week reference period is -0.002 and 0.001, respectively, estimated at the 1% level of statistical significance. This is consistent with both reference periods estimates of the marginal effect of round on productivity. Productivity declines by 6.1 percentage points and 8.8 percentage points in the two and three week reference estimates, explaining 55% of the earnings effect decrease for the two week reference and 66% of the earnings effect decrease for the three week reference in round 3. None of the labor supply treatment effect round-interactions were statistically significant.

In Table 7, we estimate round-specific treatment on the treated estimates. The pooled treatment on the treated estimate for the two week reference on earnings is 9.1% and 11.2% for the three week reference. We also estimate a pooled labor supply treatment on the treated of 6.9% in the two week reference period. In both the two week and three week reference period, the round specific effects are statistically different from each other at the 5% level of statistical significance in both the earnings and labor supply specifications. Only the round 2 treatment effect interaction is statistically significant in the two week reference specification. The round 2 treatment effect interaction coefficients for the earnings and labor supply specifications are quite large resulting in marginal effects that are negative. However, the interaction coefficient estimates are not statistically different between rounds 2 and 3. The confidence interval around the marginal effect is not statistically different from zero for round 2, so it is difficult to conclude that the treatment on the treated is negative in round 2 when in the other rounds it is significantly positive. One reason for the difference in magnitude of effects across rounds is likely the malaria positivity rate. In round 2, the malaria positive rate is only 14% of workers whereas in rounds 1 and 3 the positivity rate is 36% and 21% respectively. In the three week reference specification for the treatment on the treated, none of the interactions are statistically significant nor statistically different from each other.

Table 8 provides the treatment on the medically untreated estimates. The pooled earnings estimates for both the two and three week reference periods are similar to those reported for the two and three week reference periods. For the malaria positives, this earning effect was driven by a labor supply response, while for the malaria negatives the earning effect is driven by a productivity effect. This productivity effect is where there is the most heterogeneity between rounds particularly because the mechanism in round 1, the scrabbling contract was eliminated in rounds 2 and 3. The ability for well workers to shift first into and then out of lower return, lower effort tasks which was demonstrated in Dillon et al. (2014) increased productivity of well workers. Without the ability to shift between lower and higher effort tasks, workers may have smoothed effort differently in rounds 2 and 3 in comparison to round 1. The productivity-round interactions are not statistically different from each other with p-values of 0.363 and 0.179 for the two and three week reference period respectively. There was no strong productivity effect for the treatment of the medically

treated in either round 2 or 3. As such the earnings effect for round 3 is very close to zero, while the treatment effect round 2 interaction coefficient is not statistically significant, but is negative suggesting a smaller, though imprecisely estimated, earnings effect for round 2.

The treatment on the medically treated results illustrate an important insight about external validity which also has implications for policy. The malaria testing and treatment program demonstrated a strong effect on workers who had not shifted into a low return task in year 1. Without this possibility to select into fixed wage, lower effort work, the program's impact was much smaller. This has potentially important implications for program design and policy. Labor constraints may limit worker behavioral responses. In much of the economics of malaria literature (Sauerborn et al. (1991), Shephard et al. (1991), Ettling et al. (1994), Guiguemde et al. (1994), Attanayake et al. (2000), Chima et al. (2003), Akazili et al. (2007), or Ayieko et al. (2009)), it is presumed that malaria affects earnings and labor supply, but the task selection and productivity channels remain typically unmeasured in these studies. This treatment on the medically untreated result shows that with labor constraints, such that selection first into and the out of lower effort tasks is prohibited, the effect of malaria information on worker behavior is minimal. Yet the first round study results demonstrate the importance of this mechanism, when it is available. The combined results highlight an important benefit of replication, showing that what much of the literature assumes does not exist, because of the assumed failure of task allocation in developing countries, plays a potential key role, precisely because it demonstrates the existence of a behaviorial mechanism that is not frequently observed.

CONCLUSIONS

An emerging literature has investigated the validity of economic studies from the perspective of either replication or robustness. This study estimates the effects of malaria on earnings, labor supply and productivity across three years, using the same study design including randomization protocol, questionnaire, and malaria diagnostic procedure over three different round-years to assess the robustness of the results over time. We find that estimated treatment effects vary in size over time. These differences seem best explained by variation in malaria rates and labor constraints over time – key contextual factors that only replication are able to identify as they do not vary within site and round. The malaria prevalence rate in our study years ranges from 14% to 37%, and the estimated earnings and labor supply treatment effects on the treated are small in the years that had the lowest prevalence rate. In the round where workers could substitute into and out of lower effort, lower return tasks, we estimate a significant treatment effect on the medically untreated (the malaria negatives), which reflects the impact of information about one own malaria status. In rounds when this task substitution is not possible, we estimate no earnings or productivity effect for workers, presumably because the mechanism is not available.

These results demonstrates the importance of variation over time on the magnitude of treatment effect estimates. The treatment on the medically untreated result suggests that labor constraints inhibit behavioral responses that would otherwise arise. This bring to the fore a valuable contribution of replication studies, beyond being a useful tool for validation, namely to shed light on the potential contextual factors that determine the variation in treatment effects over space or time.

BIBLIOGRAPHY

Ahmed, A. U., and Bouis, H. E. 2002. “Weighing what’s practical: proxy means tests for targeting food subsidies in Egypt.” Food Policy, 27(5): 519-540.

Akazili, J., Aikins, M., and Binka, F. N. 2008. "Malaria treatment in Northern Ghana: What is the treatment cost per case to households?" African Journal of Health Sciences, 14(1): 70-79.

Allcott, H. 2015. "Site selection bias in program evaluation." Quarterly Journal of Economics, 130(3): 1117-1165.

Attanayake, N., Fox, M.,Rushby, J., and Mills, A. 2000. “Household costs of malaria morbidity: a study in Matale district, Sri Lanka.” Tropical Medicine & International Health, 5(9): 595-606.

Ayieko, P., Akumu, A. O., Griffiths, U. K., and English, M. 2009. “The economic burden of inpatient pediatric care in Kenya: Household and provider costs for treatment of pneumonia, malaria and meningitis.” Cost Effective Resource Allocation, 7(3).

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., and Sandefur, J. 2013. "Scaling up what works: Experimental evidence on external validity in Kenyan education." Center for Global Development Working Paper No. 321.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. 2016. "Evaluating replicability of laboratory experiments in economics." Science, 351(6280).

Chang, A.C. and Li, P. 2015. "Is economics research replicable? Sixty published papers from thirteen journals say 'usually not.'" FEDS Working Paper No. 2015-083.

Chima, R. I., Goodman, C. A. and Mills, A. 2003. “The economic impact of malaria in Africa: A critical view of the evidence.” Health Policy and Planning, 63:17–36.

Clemens, M.A. 2017. "The meaning of failed replications: A review and proposal." Journal of Economic Surveys, 31(1): 326-342.

dalla Martha RC, Tada MS, Ferreira RG, da Silva LH, and Wunderlich G. 2007. “Microsatellite characterization of Plasmodium falciparum from symptomatic and non-symptomatic infections from the Western Amazon reveals the existence of non-symptomatic infection-associated genotypes." Memórias do Instituto Oswaldo Cruz, 102:293-298.

Das, A., Friedman, J., and Kandpal, E. 2017. "Does involvement of local NGOs enhance public service delivery? Cautionary evidence from a malaria-prevention program in India." Health Economics, 0:1-17.

Deaton, Angus S. 2009. “Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development.” National Bureau of Economic Research Working Paper 14690.

Dillon, A., Friedman, J., and Serneels, P. 2014. “Health Information, treatment, and worker productivity: Experimental evidence from malaria testing and treatment among Nigerian sugarcane cutters.” World Bank Policy Research Working Paper No. 7120.

Ettling, M., Mcfarland, D.A., Schulz, L.J., and Chitsulo, L. 1994. “Economic impact of malaria in Malawian households.” Tropical Medical and Parasitology. 45. Supplement 1: 74-79.

Flores, C.A. and Mitnik, O.A. 2013. "Comparing treatments across labor markets: An assessment of nonexperimental multiple-treatment strategies." Review of Economics and Statistics, 95(5): 1691-1707.

Gechter, M. 2016. "Generalizing the results from social experiments: Theory and evidence from Mexico and India." Working paper, http://www.personal.psu.edu/mdg5396/Gechter_Generalizing_Social_Experiments.pdf.

Grosh, M and Baker, J. 1995. “Proxy means tests for targeting social programs: Simulations and speculation.” LSMS Working Paper No. 118. Washington, DC: World Bank.

Guiguemde, T., 1994. “Household expenditure on malaria prevention and treatment for families in the town Of Bobo-Dioulasso, Burkina Faso.” Transactions of the Royal Society of Tropical Medicine and Hygiene, 88(3), 285–292.

Hotz, V., Imbens, G., and Mortimer, J. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." Journal of Econometrics, 125(102), 241–270.

Rodrik, Dani. 2009. “The new development economics: We shall experiment, but how shall we learn?” In What Works in Development? Thinking Big and Thinking Small, ed. Jessica Cohen and William Easterly, 24–47. Washington, D.C.: Brookings Institution Press.

Rottman, M., Lavstsen, T., Mugasa, J.P., Kaestli, M., Jensen, A., T. R., Müller, D., Theander, T., and Bech, H-P. 2006. "Differential expression of var gene groups is associated with morbidity caused by Plasmodium falciparum infection in Tanzanian children." Infection and immunity, 74 (7): 3904-3911.

Sauerborn, R., D. S. Shepard, M. B. Ettling, U. Brinkmann, A. Nougtara, and Diesfeld, H. J. 1991. "Estimating the direct and indirect economic costs of malaria in a rural district of Burkina Faso." Tropical Medicine And Parasitology, 42(3): 219-223.

Shepard, D. S., Ettling, M. B., Brinkmann, U. and Sauerborn, R. 1991. "The economic cost of malaria in Africa." Tropical Medicine and Parasitology, 42 (3): 199.

Stuart, E., Cole, S., Bradshaw, C., and Leaf, P. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2), 369–386.

Toure F.S., Bisseye C., and Mavoungou E. 2006. "Imbalanced distribution of Plasmodium falciparum EBA-175 genotypes related to clinical status in children from Bakoumba, Gabon." Clinical Medicine and Research, 4:7-11.

WHO. 2010. Guidelines for the Treatment of Malaria (2e). 2nd Edition. Geneva: World Health Organisation.

TABLES

Table 1: Study Design Differences by Round

	2010	2011	2013
Implementation period	Last week of January-March	February-March	February-April
Sampling protocol	Between group week randomization	Between group week randomization	Between group week randomization
Data collection			
Worker survey	National survey firm	National survey firm	University based enumerators
Health survey	University based enumerators	University based enumerators	University based enumerators
Clinical team	University based staff	University based staff	University based staff
Malaria Prevalence	35.7%	14.1%	20.5%
Worker sample size	803	871	682
Plantation management change		New CEO Same Plantation Manager	New CEO Same Plantation Manager
Worker contract	Fixed daily wage 'scrabbling' option	No Fixed daily wage 'scrabbling' option	No Fixed daily wage 'scrabbling' option
Unforseen events			Work stoppage due to processing plant failure

Table 2: Worker Characteristics by Round

	Round 1			Round 2			Round 3			p-values of tests for equality of means		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	R1 vs R2	R1 vs R3	R2 vs R3
Individual characteristics												
Age	803	29.99	8.16	871	30.76	7.72	682	32.24	8.21	0.03	0.00	0.00
Experience with firm in years	803	4.30	4.11	871	4.83	5.19	682	6.18	5.90	0.01	0.00	0.00
Years of education	803	8.24	4.31	871	8.45	4.30	682	8.70	5.17	0.38	0.05	0.21
Body mass index (BMI)	799	23.77	2.60	871	22.86	2.24	682	22.48	2.38	0.00	0.00	0.00
Household characteristics												
Household size	803	5.39	6.30	871	6.30	4.04	682	5.78	4.24	0.00	0.14	0.01
Number of rooms in the house	803	2.78	2.15	871	3.25	2.05	678	2.92	2.40	0.00	0.25	0.00
Number of cattle	803	1.27	4.21	871	3.46	6.25	682	2.33	4.93	0.00	0.00	0.00
Number of poultry	803	7.39	12.00	871	10.34	12.02	682	9.70	11.53	0.00	0.00	0.29
Imputed per capita monthly exp.	803	12,542.50	6,251.46	871	16,743.82	5,780.44	677	13,495.86	7,549.26	0.00	0.00	0.00
Worker characteristics												
Average daily earnings (Naira)	803	1,052.07	263.35	871	909.87	259.52	682	1,161.04	305.02	0.00	0.00	0.00
Total days worked	802	66.49	15.59	868	70.04	12.44	679	27.21	4.85	0.00	0.00	0.00
Proportion of time spent scrabbling	803	0.17	0.22	871	0.00	0.00	682	0.00	0.00	0.00	0.00	-

Note: These descriptive statistics are calculated for all workers in the sample. The p-value results from a t-test of mean-differences. Differences in sample sizes within a round due to missing values.

Table 3: Worker health and health-seeking behavior by round

	Round 1-2010			Round 2-2011			Round 3-2013			p-value for test of equality of means		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	R1 vs R2	R1 vs R3	R2 vs R3
Fever self-report	803	0.08	0.27	871	0.12	0.32	682	0.02	0.15	0.01	0.00	0.00
Formal health care seeking	803	0.07	0.25	871	0.10	0.30	682	0.02	0.13	0.01	0.00	0.00
Informal health care seeking	803	0.00	0.00	871	0.00	0.00	682	0.00	0.05	-	0.11	0.10
Any self-reported morbidity	803	0.07	0.25	871	0.08	0.28	682	0.04	0.19	0.16	0.01	0.00
Other illness self-report	803	0.12	0.33	871	0.57	0.50	682	0.10	0.31	0.00	0.58	0.00

Note: These descriptive statistics are calculated for all workers in the sample. The p-value results from a t-test of mean-differences.

Table 4: Balancing Test p-values within round

Work group	Number of workers	Age	Years of experience in cane cutting	Years of schooling	BMI	Household size	Number of rooms in house	Number of cattle	Number of poultry	Imputed monthly PC expenditures
Round 1 within work-group balance of worker characteristics across survey week										
1	96	0.411	0.831	0.748	0.148	0.722	0.342	0.728	0.914	0.977
2	93	0.391	0.567	0.511	0.149	0.664	0.614	0.532	0.248	0.924
3	111	0.911	0.830	0.526	0.523	0.383	0.719	0.481	0.540	0.662
4	106	0.018	0.438	0.126	0.088	0.271	0.793	0.947	0.579	0.711
5	102	0.712	0.232	0.928	0.434	0.878	0.029	0.668	0.136	0.534
6	96	0.220	0.070	0.871	0.232	0.754	0.241	0.680	0.669	0.892
7	99	0.528	0.593	0.061	0.162	0.819	0.697	0.095	0.636	0.585
8	89	0.517	0.690	0.756	0.270	0.042	0.257	0.589	0.694	0.519
Round 2 within work-group balance of worker characteristics across survey week										
1	84	0.228	0.246	0.152	0.004	0.058	0.289	0.509	0.122	0.538
2	68	0.868	0.420	0.152	0.003	0.734	0.303	0.707	0.688	0.142
3	88	0.242	0.193	0.160	0.116	0.019	0.148	0.719	0.618	0.549
4	75	0.522	0.559	0.701	0.278	0.609	0.219	0.273	0.930	0.839
5	87	0.754	0.166	0.032	0.105	0.452	0.808	0.380	0.711	0.037
6	60	0.837	0.695	0.127	0.502	0.894	0.021	0.351	0.280	0.061
7	72	0.952	0.676	0.911	0.012	0.966	0.496	0.091	0.707	0.832
8	76	0.316	0.250	0.273	0.710	0.363	0.172	0.619	0.743	0.528
Round 3 within work-group balance of worker characteristics across survey week										
1	92	0.107	0.709	0.794	0.985	0.230	0.620	0.859	0.832	0.129
2	91	0.067	0.497	0.184	0.880	0.050	0.142	0.741	0.389	0.004
3	94	0.692	0.999	0.663	0.527	0.884	0.894	0.117	0.077	0.486
4	92	0.679	0.006	0.297	0.544	0.007	0.751	0.907	0.921	0.005
5	92	0.729	0.410	0.163	0.141	0.635	0.155	0.802	0.050	0.732
6	91	0.384	0.727	0.965	0.052	0.120	0.299	0.030	0.228	0.269
7	86	0.796	0.040	0.527	0.798	0.300	0.000	0.731	0.030	0.226

Note: The p-value reported is from an F-test that tests the equality of coefficients across survey weeks in a regression of each of the individual characteristics on week indicators for each work group.

Table 5

Work group	Number of workers	Malaria positive	Fever self-report	Formal health care seeking	Informal health care seeking	Any self-reported morbidity	Other illness self-report
Round 1 within work-group balance of worker health and health behaviors across survey week							
1	96	0.327	0.074	.	0.949	0.477	0.724
2	93	0.676	0.877	0.484	0.908	0.115	0.643
3	111	0.610	0.790	0.227	0.546	0.221	0.855
4	106	0.353	0.746	.	.	0.959	0.848
5	102	0.727	0.247	0.937	0.044	0.076	0.552
6	96	0.750	0.037	0.805	0.805	0.605	0.001
7	99	0.376	0.578	.	0.843	0.615	0.111
8	89	0.016	0.564	0.147	0.859	0.932	0.002
Round 2 within work-group balance of worker health and health behaviors across survey week							
1	84	0.349	0.553	0.546	0.111	0.310	0.026
2	68	0.830	0.201	0.268	0.615	0.229	0.213
3	88	0.589	0.116	0.294	0.165	0.098	0.464
4	75	0.143	0.566	0.387	0.755	0.810	0.846
5	87	0.613	0.504	0.836	0.200	0.178	0.225
6	60	0.000	0.094	0.229	0.551	0.690	0.926
7	72	0.350	0.255	0.839	0.898	0.443	0.842
8	76	0.292	0.589	0.809	0.532	0.086	0.529
Round 3 within work-group balance of worker health and health behaviors across survey week							
1	92	0.009	0.069	0.824	0.553	0.345	0.877
2	91	0.508	0.478	.	0.001	0.004	0.893
3	94	0.004	0.000	0.050	0.921	0.790	0.955
4	92	0.469	0.973	0.740	0.161	0.100	0.358
5	92	0.762	0.850	0.279	0.291	0.209	0.922
6	91	0.119	0.277	.	0.463	0.255	0.463
7	86	0.410	0.088	0.946	0.679	0.062	0.350

Note: The p-value reported is from an F-test that tests the equality of coefficients across survey weeks in a regression of each of the individual characteristics on week indicators for each work group. Missing p-values indicate there was no worker in the work group that reported seeking formal health care.

Table 6: Intent to treat effects

	2 week reference			3 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.108***	0.049**	0.059***	0.134***	0.045*	0.089***
	(0.031)	(0.021)	(0.022)	(0.040)	(0.027)	(0.030)
Round 2	-0.877***	-0.690***	-0.186***	0.149***	-0.019	0.168***
	(0.047)	(0.032)	(0.031)	(0.054)	(0.036)	(0.038)
Round 3	0.566***	-0.133***	0.364***	0.068	-0.189***	-0.313***
	(0.031)	(0.021)	(0.022)	(0.047)	(0.033)	(0.039)
Treat x round 2	-0.058	-0.003	-0.055*	-0.047	0.019	-0.067*
	(0.047)	(0.032)	(0.031)	(0.054)	(0.036)	(0.038)
Treat x round 3	-0.110***	-0.005	-0.061*	-0.133***	-0.035	-0.088**
	(0.041)	(0.029)	(0.035)	(0.047)	(0.033)	(0.039)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.000	0.000	0.000	0.065	0.000	0.000
Round 2 x treat = Round 3 x treat	0.249	0.968	0.851	0.049	0.062	0.510
N	5,625	5,625	5,625	7,382	7,382	7,382

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1

Table 7: Treatment on the treated

	2 week reference			3 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.091**	0.069**	0.022	0.112*	0.070	0.042
	(0.045)	(0.034)	(0.037)	(0.057)	(0.045)	(0.051)
Round 2	0.328***	0.292***	0.036	-0.234	0.295***	-0.529***
	(0.045)	(0.034)	(0.037)	(0.153)	(0.088)	(0.101)
Round 3	0.641***	0.034	0.012	0.247***	0.475***	-0.383***
	(0.079)	(0.065)	(0.076)	(0.086)	(0.068)	(0.082)
Treat x round 2	-0.260*	-0.182**	-0.078	-0.114	-0.072	-0.043
	(0.146)	(0.077)	(0.100)	(0.153)	(0.088)	(0.101)
Treat x round 3	-0.031	-0.034	0.038	-0.032	-0.069	0.038
	(0.079)	(0.065)	(0.076)	(0.086)	(0.068)	(0.082)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.000	0.000	0.717	0.002	0.048	0.177
Round 2 x treat = Round 3 x treat	0.135	0.095	0.304	0.599	0.980	0.457
N	1,394	1,394	1,394	1,823	1,823	1,823

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1

Table 8: Treatment on the medically untreated

	2 week reference			3 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.117***	0.043	0.074***	0.142***	0.031	0.112***
	(0.042)	(0.028)	(0.028)	(0.052)	(0.034)	(0.037)
Round 2	-0.895***	-0.716***	-0.179***	-1.957***	-1.643***	-0.314***
	(0.057)	(0.038)	(0.037)	(0.064)	(0.042)	(0.044)
Round 3	0.576***	-0.139***	0.379***	0.653***	0.031	0.285***
	(0.042)	(0.028)	(0.028)	(0.052)	(0.034)	(0.037)
Treat x round 2	-0.040	0.023	-0.063*	-0.049	0.033	-0.082*
	(0.057)	(0.038)	(0.037)	(0.064)	(0.042)	(0.044)
Treat x round 3	-0.141***	0.004	-0.098**	-0.162***	-0.016	-0.131***
	(0.052)	(0.035)	(0.042)	(0.059)	(0.040)	(0.046)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.000	0.000	0.000	0.000	0.000	0.000
Round 2 x treat = Round 3 x treat	0.040	0.595	0.363	0.016	0.118	0.179
N	4,231	4,231	4,231	5,559	5,559	5,559

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1

Appendix Tables

Table 9: Intent to treat effects (1 and 4 week reference)

	1 week reference			4 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.043	0.041*	0.001	0.065	0.029	0.036
	(0.033)	(0.024)	(0.022)	(0.101)	(0.062)	(0.062)
Round 2	1.788***	1.556***	0.231***	-2.053***	-1.136***	-0.917*
	(0.067)	(0.050)	(0.040)	(0.505)	(0.066)	(0.499)
Round 3	2.668***	0.005	0.680***	1.117***	0.541***	0.241**
	(0.046)	(0.041)	(0.039)	(0.125)	(0.065)	(0.097)
Treat x round 2	0.044	0.053	-0.008	0.038	0.037	0.000
	(0.067)	(0.050)	(0.040)	(0.107)	(0.066)	(0.066)
Treat x round 3	-0.022	-0.005	0.020	-0.071	-0.030	-0.042
	(0.046)	(0.041)	(0.039)	(0.104)	(0.065)	(0.068)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.000	0.000	0.000	0.000	0.000	0.021
Round 2 x treat = Round 3 x treat	0.311	0.286	0.534	0.020	0.028	0.239
N	3,219	3,219	3,219	8,300	8,300	8,300

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1

Table 10: Treatment on the treated (1 and 4 week reference)

	1 week reference			4 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.007	0.023	-0.015	-0.160	-0.111***	-0.048
	(0.049)	(0.038)	(0.038)	(0.099)	(0.042)	(0.103)
Round 2	0.457***	0.428***	0.029	-2.188***	-1.475***	-0.713***
	(0.049)	(0.038)	(0.038)	(0.193)	(0.113)	(0.139)
Round 3	0.132	-1.433***	-0.451***	0.728***	0.203**	0.279
	(0.111)	(0.105)	(0.107)	(0.258)	(0.079)	(0.259)
Treat x round 2	0.037	0.138	-0.101	0.057	0.089	-0.032
	(0.231)	(0.190)	(0.125)	(0.193)	(0.113)	(0.139)
Treat x round 3	0.174	0.047	0.197*	0.258**	0.127*	0.147
	(0.111)	(0.105)	(0.107)	(0.121)	(0.067)	(0.123)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.001	0.000	0.000	0.000	0.000	0.000
Round 2 x treat = Round 3 x treat	0.578	0.664	0.058	0.262	0.743	0.123
N	802	802	802	2,028	2,028	2,028

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1

Table 11: Treatment on the medically untreated (1 and 4 week reference)

	1 week reference			4 week reference		
	ln(earnings)	ln(days)	ln(wage)	ln(earnings)	ln(days)	ln(wage)
Treat	0.056	0.050	0.004	0.162	0.102	0.061
	(0.046)	(0.032)	(0.029)	(0.128)	(0.076)	(0.081)
Round 2	0.335***	-0.037	0.371***	-1.760***	-1.067***	-0.693***
	(0.076)	(0.055)	(0.045)	(0.134)	(0.081)	(0.085)
Round 3	1.635***	1.660***	-0.363***	1.235***	0.618***	0.286**
	(0.046)	(0.032)	(0.029)	(0.149)	(0.080)	(0.111)
Treat x round 2	0.026	0.037	-0.009	-0.046	-0.032	-0.015
	(0.076)	(0.055)	(0.045)	(0.134)	(0.081)	(0.085)
Treat x round 3	-0.074	-0.023	-0.022	-0.189	-0.107	-0.087
	(0.058)	(0.047)	(0.046)	(0.132)	(0.080)	(0.086)
P-values for tests of equality of coefficients						
Round 2=Round 3	0.000	0.000	0.000	0.000	0.000	0.000
Round 2 x treat = Round 3 x treat	0.155	0.301	0.789	0.004	0.024	0.061
N	2,417	2,417	2,417	6,272	6,272	6,272

Note: Round 1 is the omitted category. Robust standard errors clustered at the worker level. Regressions include workgroup by week fixed effects. ***p<0.01, **p<0.05, *p<0.1