Contemporaneous and Post-Program Impacts of a Public Works Program: Evidence from Côte d'Ivoire

Marianne Bertrand (University of Chicago, Booth School of Business)

Bruno Crépon (CREST)

Alicia Marguerie (CREST – ENSAE Paristech - Paris Saclay)

Patrick Premand (World Bank)

This version: 05/23/2017

Abstract

Public works are one of the most popular safety net and employment policy instruments in the developing world, despite limited evidence on their effectiveness and optimal design features. This paper presents results on contemporaneous and post-program impacts from a public works intervention in Côte d'Ivoire. The program provided 7 months of temporary employment in road maintenance to urban youths. Participants self-selected to apply for the public works jobs, which paid the formal minimum wage and were randomized among applicants. Randomized sub-sets of beneficiaries also received complementary training on basic entrepreneurship or job search skills. During the program, results show limited contemporaneous impacts of public works on the level of employment, but a shift in the composition of employment towards the better-paid public works wage jobs. A year after the end of the program, there are no lasting impacts on the level or composition of employment, although positive impacts are observed on earnings through higher productivity in non-agricultural self-employment. Large heterogeneity in impacts are found, particularly during the program. Results from machine learning techniques suggest potential trade-offs between maximizing contemporaneous and post-program impacts. Traditional heterogeneity analysis shows that a range of practical targeting mechanisms perform as well as the machine learning benchmark, leading to stronger contemporaneous and post-program benefits without sharp trade-offs. Overall, departing from self-targeting based on the formal minimum wage would lead to strong improvements in program cost-effectiveness.

Acknowledgements

The randomized control trial of the PEJEDEC public works program was implemented in close collaboration with the government of Côte d'Ivoire (in particular BCPE, Bureau de Coordination des Programmes d'Emploi, and AGEROUTE) and the World Bank. We are particularly grateful to Hamoud Wedoud Abdel Kamil at the World Bank; Adama Bamba, Hermann Toualy, Ismahel Abdoul Barry and Fabrice Konan at BCPE; as well as Marius Pokou, Martin Kouakou, Yves N'Cho and Kassi Ernest Bohoussou at AGEROUTE.

High-quality baseline and midline data were collected by ENSEA, and managed by Hugues Kouadio, Rosine Addy Mosso, Marie Judith Soro and Nathaniel Gbenro. High-quality endline data was collected by BCPE under the leadership of Ismahel Abdoul Barry and Alicia Marguerie, with support from Jean Awé. Sondoi Eloi Somtinda provided excellent field coordination during baseline, midline and endline data collection.

We are deeply indebted to Susan Athey and Stefan Wager for providing us access to and support with causalForest and causalTree packages, as well as Jonathan Davis and Jann Spiess for their suggestions on machine learning implementation. We thank Stefanie Brodmann, Deon Filmer, Emanuela Galasso, Jessica Goldberg, Dominique Van de Walle and seminar participants at CREST, U. Chicago, CSAE, ILO, IZA and the World Bank for useful comments and suggestions. All errors and omissions are our own.

In addition from funding from the PEJEDEC project, the study was also supported by the MESF and DIME i2i Trust Funds at the World Bank. A detailed impact evaluation policy report in French is available in Bertrand et al. (2016).

1 Introduction

Public works programs are an important component of the policy portfolio of decision makers trying to address the social challenges of underemployment and poverty. Under such programs, the government offers temporary employment, typically remunerated at the minimum wage or below, for the creation of public goods, such as road or infrastructure. Unlike welfare programs, such as cash transfers, public works programs transfer cash to their beneficiaries conditional on their meeting work requirements.

There are different types of public works programs. Some are employment guarantee schemes that offer participants a number of days of employment on demand each year and have as primary objective to provide social insurance for the poor. For example, the Mahatma Gandhi National Rural Employment Guarantee program (MGNREGA) in India guarantees 100 days of work per year per household to all rural households with members willing to do unskilled manual labor at the statutory minimum wage.

Other programs are implemented to address temporary shocks, such as those induced by an economic downturn, a climatic shock or period of violent conflict, and primarily aim to offer mass public employment as a stabilization instrument. They also include programs that provide temporary employment during the lean agricultural season, to address underemployment and seasonality in agriculture, as well as help households deal with the incidence of shocks and transient poverty. In Sub-Saharan Africa, our context in this paper, labor-intensive public works programs have often been adopted in response to transient negative shocks such as those induced by climatic shocks or episodes of violent conflicts. These programs typically offer temporary employment (for a few months), at the minimum wage or below.

While traditional welfare programs, such as unconditional cash transfers, could also be used to support the poor and most vulnerable, public work programs are often claimed to have a variety of advantages. First, while both “workfare” and welfare programs can transfer cash to the poor, workfare programs also contribute to the creation of public assets (e.g. better roads) which may benefit the broader community. This argument is particularly relevant in contexts where physical infrastructure was destroyed or damaged as an outcome of the crisis the programs are aiming to address (e.g. climatic shocks or violent conflict).

Also, “workfare” programs, through skill development or the signaling value of prior work experience, may increase the future employability or productivity of the participants. This benefit can be potentially further improved upon, even if at a higher cost, by adding some complementary productive interventions, such as savings facilitation or training, to the work experience.

Another advantage of workfare programs, as highlighted by Besley and Coate (1992), is that they can solve the difficult problem of efficiently targeting who should be the beneficiaries for the transfers. Indeed, the appropriate targeting of social protections programs is particularly complex in lower income countries because of a lack of robust data, challenges to identify beneficiaries at the bottom of the welfare distribution, as well as weak systems and institutions, leading to potential errors of inclusion or exclusion. Public works programs are appealing as they may in theory solve this problem through self-targeting, in that only the poor would be willing to supply labor in these programs at the stated (low) wage.

Also, and particularly relevant to post-conflict environments, engaging beneficiaries in work-for-cash rather than simply handing out cash may operate as a social stabilization tool. This might operate through an incapacitation effect: time spent working may displace socially disruptive activities such as crime. Moreover, work, even if unpleasant, may improve well-being, self-esteem, and overall mental health, a set of non-cognitive assets that may also help beneficiaries become more productively employed.

Finally, even though more mundane, another advantage of public work programs compared to traditional welfare programs is that they are often politically more acceptable and sustainable. Political preferences for workfare programs are often linked to (valid or not) concerns about welfare dependency (and how unconditional transfers may disincentive work) as well as a desire to generate immediate visible improvements to employment conditions.

In this paper, we present the results of a randomized control trial designed to assess both the contemporaneous (i.e. during the program) and post-program (i.e. a year after program completion) impacts of public work programs. The particular public works program we evaluate was implemented by the Côte d'Ivoire government in the aftermath of the post-electoral crisis that hit the country in 2010/2011, and was funded by an emergency loan from the World Bank. The stated objective of the program was to improve access to temporary employment opportunities among lower-skilled young (18-30) men and women in urban or semi-urban areas that were unemployed or underemployed, as well as to develop the skills of the program participants through work experience and complementary training.

In particular, participants in the public works program were employed for a period of 7 months to rehabilitate and clean road infrastructure and remunerated at the statutory minimum daily wage, corresponding to about $5 per day (FCFA 2500), approximately $110 per month (FCFA 55 000). Program participants were required to work 6 hours per day, 5 days per week. The road maintenance work, carried out in "brigades" of about 25 youths with a supervisor, was implemented by the National Roads Agency (AGEROUTE) and supervised by BCPE[1], under the Ministry of Labor and Social Affairs.

All young men and women in the required age range and residing in one of 16 urban localities in Côte d'Ivoire were eligible to apply to the program. Because the number of applicants outstripped supply in each locality, fair access was based on a public lottery, setting the stage for a robust causal evaluation of the impacts of the program. In addition, a randomized sub-set of beneficiaries were also offered (i) basic entrepreneurship training to facilitate set-up of new household enterprises and entry into self-employment, or (ii) training in job search skills and sensitization on wage employment opportunities to facilitate access to wage jobs (e.g. help in identifying wage job opportunities, CV production, interview skills, etc.).

In addition to a baseline survey of program applicants, we carried rich surveys of youth in the treatments and control groups both during the program (4 or 5 months after the program had started) as well 12 to 15 months after program completion to capture any post-program effects of participation.

Our analysis of contemporaneous effects demonstrates that the program had limited impacts on the level of employment, mostly inducing shifts in the composition of employment. Reflecting limited unemployment and a high concentration of the active population in low-productivity occupations, 86 percent of the control group was working 4 to 5 months after the lottery took place, compared to 98 percent of those assigned to the intervention. Moreover, the program did not substantially raise hours worked, with mean hours worked at about 41 hours in the control group compared to 44 hours for those assigned to the treatment. The value of the program for the modal applicant was therefore not as a way to escape unemployment but more as a way to escape under-employment in low-paying informal activities: monthly earning are about FCFA 20,000 higher in the treatment groups, from a base of FCFA 60,000 in the control group. So, while the program managed to lift earnings, foregone earnings are quantitatively important, with only about 40 percent of the transfer translating into earnings gain.

[1] Coordination Office for Employment Programs (« Bureau de Coordination des Programmes Emploi »)

These results strongly suggest that self-targeting based on the formal minimum wage did not succeed in this context in getting only the most vulnerable (e.g. those without outside employment opportunities) to benefit from the program. A couple of factors likely explain this failure of self-targeting. First, a job that pays the statutory minimum wage could still be of appeal to many in an environment where informal employment and self-employment are rampant. Second, because the work was only 6 hours per day, many applicants with outside employment opportunities, especially those that allow for more flexible hours, would still see value in applying for the public works program as they could combine it with other activities. Finally, while the unpleasant nature of the work may have discouraged some, it is unclear whether this work is more unpleasant than most informal activities. In fact, positive effects on overall well-being and behavioral skills are also found in the short term.

Twelve to 15 months after program completion, no impacts are observed either on the level (employment or hours worked) or on composition of employment (salaried work vs. self-employment). However, we do observe sustained positive impacts on earnings (FCFA 5,622 or $11 per month, a 11.6 percent increase compared to the control group), mainly stemming from non-agricultural self-employment activities. These sustained impacts are mostly driven by youths who were assigned to the public works and complementary basic entrepreneurship training.

Based on these estimates of direct earnings impacts on youths, and given rich data we have collected on program costs, we conclude the public works program under its current form is far from cost-effective, with cost per participant being about 3 times the estimated benefit.

While our results suggest that self-targeting based on the formal minimum wage failed in this context, it is possible that better targeting criteria may have resulted in improved cost-effectiveness. This is the issue we address in the remaining sections of the paper, where we study heterogeneity of program effects, both during the program and post-program.

First, using recent machine learning techniques (Athey and Imbens 2016, Wager and Athey 2016), and relying on the very rich data collected at baseline about each program applicant, we estimate the heterogeneity of program effects both during and after the program. This analysis confirms large differences in predicted impacts across various groups of program participants, especially during the program. In particular, the average predicted impact on earnings in the short-term in the upper quartile of the predicted impact distribution is over FCFA 28,000, compared to 9,900 in the lower quartile. Also, the average predicted post-program impact on earnings in the upper quartile of the predicted impact distribution is over FCFA 8,000, compared to 1,475 in the lower quartile.

While these results suggest that the program effectiveness, both in the short-term and the long-term, could be improved through better targeting, they also highlight the difficulty in improving contemporaneous and post-program impacts at the same time: program participants that benefit most during the program are not systematically those that benefit most after the program has ended.

However, an analysis of the distribution of predicted contemporaneous and post-program impacts does reveal the existence of some common targeting dimensions that may improve effectiveness without sharp trade-offs. Compared to the benchmark scenario with self-targeting based on the formal minimum wage, the cost-effectiveness ratio would improve from 3.2 to between 1.58 and 1.98 based on finer program targeting such as selecting youths with low predicted baseline earnings, self-selection based on a lower offered wage, targeting women only, or targeting based on self-declared baseline earnings. While the analysis cannot decisively indicate which targeting scenario would maximize cost-effectiveness given the confidence intervals around the impact estimates, it does highlight strong improvements in cost-effectiveness when departing from self-targeting solely based on the formal minimum wage.

2 Framework

The primary objective of workfare programs is to provide income support through temporary jobs. Beyond this short-term goal contemporaneous to the program, another objective is sometimes to facilitate transitions to more productive, higher-earnings occupations after the program. Several potential channels can contribute to such longer-term objectives, for instance the ability to save and invest in ongoing or new activities, or the opportunity to develop skills valued in the labor market.

2.1 Contemporaneous Impacts in the Short Term

Targeting is a key design feature of social safety nets programs seeking to provide income-support in the short term. It is a general issue that each transfer program faces. The specificity of workfare programs is that participants typically self-select into the program. This is a major difference compared to transfer programs, which often select participants based on a screening procedure such as a proxy-means test or a participatory community targeting approach. The mechanism through which workfare programs address the selection problem is to ask for hours of work in exchange for the transfer. This mechanism has several well-known implications. The first one is the self-selection mechanism: only potential participants for whom the utility level in the program, accounting for earnings and disutility of work, is larger than their current utility level should participate. The second important implication is that there are forgone earnings when participating in the program: the time participants spend in the program cannot be used for work on regular activities. Thus the contemporaneous program impact on earnings is the difference between transfers to the participants and forgone earnings. Another consequence is that the impact of participation on income in the short term is heterogeneous and depends on participants' alternative economic opportunities, as well as the extent to which they are able to keep operating these activities while in the program. A reasonable expectation is that the impact is almost zero for the 'marginal' participant. On the other hand, the impact should reach the amount of the transfer for those whose income would have been zero without the program. In this context, the average contemporaneous program impact on self-selected participants depends on the distribution of individual impacts over the population, and is typically smaller than the transfer from the program.

The following simple model helps formalizing these ideas. Assume for example that earnings for a program participant are $W_p = w_p h_p$. This is a lump sum. There are fixed hours to work in the program, with no part-time participation possible. Let $W_1(h)$ capture participants' earnings for h hours of work on other activities while in the program and let h_1 be the number of hours participants can spend on these activities while in the program. Let $W_0(h)$ capture earnings for h hours of work absent the program and h_0 be the number of hours participants would have spent on these activities absent the program. The impact of the program on earnings in the short term is therefore

$$Impact(Income) = W_p + E_{Part}\ \left(W_1(h_1) - W_0(h_0)\right) = W_p - E_{Part}\ (W_0(h_0) - W_1(h_1))$$

where E_{Part} means expectation on participants. It is thus different from the direct transfer received from the program by the amount of forgone earnings : $W_0(h_0) - W_1(h_1)$.
Similarly the impact on the number of hours of work is different from hours spent in the program

$$Impact(hours) = h_p + E_{Part}(h_1 - h_0) = h_p - E_{Part}(h_0 - h_1)$$

A standard ratio to assess the capacity of the program to increase earnings is the ratio of contemporaneous program impact on income to the average income of participants in the control group during the program.

$$\frac{W_p + E_{Part}\ \left(W_1(h_1) - W_0(h_0)\right)}{E_{Part}(W_0(h_0))}$$

One interesting parameter to add is the ratio of the contemporaneous program impact on income to the actual transfer:

$$\Lambda = \frac{W_p - E_{Part}\left(W_0(h_0) - W_1(h_1)\right)}{W_p}$$

In this context, one of the key empirical question for a given program is the extent to which hours of work and earnings change due to participation in the program. What are the changes in the portfolio of activities of program participants? Are there substantial changes in hours worked? Are there changes in income during the program?

2.2 Heterogeneity in contemporaneous program impacts

Assuming for a moment that only earnings during the program matter for participation, the basic idea of workfare as a selection device is that participants will self-select into the program. Hours of work are determined by participants seeking to maximize their utility. Assuming earning functions have the same form $W_1(h) = W_0(h) = w * h$ and that there is an increasing convex disutility of effort $c(h)$, there is an optimal number of hours of work $h_0(w)$, leading to a maximum utility level absent the program $V_0(w) = w * h_0(w) - c\big(h_0(w)\big)$. In such a simple setting, there is a threshold in earnings $\underline{w}$ (that is supposedly lower than w_p) such that for $w < \overline{w}$, $h_0(w) < h_p$. It can be shown that in such a case $h_1(w) = 0$, and that the utility level increases: for potential participants with very few outside employment opportunities, program participation would lead to an increase in the number of hours, in earnings and in the utility level. For example, if we consider people who have no employment opportunity: $w = 0$, hours of work and earnings absent the program would be zero. In such a case, the impact of program participation on hours and earnings would be an increase by respectively h_p and $w_p * h_p$. For individuals with intermediate opportunities, $\overline{w} < w < w_p$, the total number of hours does not change: $h_1(w) = h_0(w) - h_p$ but the utility level increases $V_1(w) = V_0(w) + \big(w_p - w\big)h_p$. Last, for $w > w_p$ the program does not improve the utility level and therefore individuals would not participate.

This simple setting could be adapted to account for several important practical aspects of public works program. For example, there could be a fixed cost of participation due to the cost of accommodation, or specific disutility for the type of work required by the program, or lost earnings due to transportation costs to the work sites, etc. Notice that, as is well-known, there might be cases when income effects imply that program participation would actually reduce the numbers of hours worked. A general result remains: contemporaneous program impacts on income and utility are heterogeneous in the population. There are ‘marginal’ participants for whom employment opportunities are large enough for them to be indifferent between participating or not. On the other hand, there are other participants with very few opportunities for whom contemporaneous program impacts are expected to be very large.

It is therefore critical to understand the heterogeneity of impacts in public works program. We address this issue in this paper by assessing whether there is evidence of heterogeneous program impacts on employment, hours worked and earnings during the program.

Notice that assuming a disutility of work $c_i(h) = c_i h^2/2$ for individual i, it is possible to show that the intervention is rank preserving for hours worked. This is also the case for earnings if we assume that w_i is the only source of heterogeneity, and thus that disutility of work is homogeneous ($c_i = c$). However, if there is also heterogeneity in c_i, there is no reason for the rank of individuals in the earning distribution to be the same with and without the program. This is important as rank preservation is a property that helps in the identification of the variance of the program impacts.

2.3 Targeting

Targeting is a question related to heterogeneity. Assume productivity (w) is the only source of heterogeneity and that the social planner seeks to maximize the average income of a target population.[2] Are there different assignment mechanisms that would improve the average impact on utility or earnings? A first idea is to change the characteristics of the workfare contract. Let $I_0(w)$ stands for $W_0(h_0(w))$ and $I_1(w) = h_p * w_p + W_1\big(h_1(w)\big) = h_p * w_p + \big(h_0(w) - h_p\big) * w * 1_{[w>\overline{w}]}$. Assuming the program is made available to everybody in the population, the average impact on income becomes

$$S\big(w_p\big) = \int_{w<w_p} (I_1(w) - I_0(w))f(w)dw$$

$$= \int_{w<w_p} [h_p * w_p + \big(h_0(w) - h_p\big) * w * 1_{[w>\overline{w}]} - w * h_0(w)]f(w)dw$$

[2] It is more common to consider that the social planner seeks to maximize well-being. We consider earnings here to make the link with the empirical section.

Given the apparent ‘waste’ due to participants for whom the marginal impact is zero, a natural idea would be to reduce the contract wage. However, this would be ineffective since for the marginal applicant the program effect would remain zero. A marginal reduction in the wage w_p would actually reduce the impact on average income, corresponding to the related reduction in income for each participant (there is no contribution of the newly selected out participants for whom the impact of the program is zero):

$$S'(w_p) = \ h_p * F(w_p).$$

However the program is generally not available to everybody in the population. Assume that there is oversubscription. One selection mechanism is to randomly assign applicants to the program until the budget constraint is reached. The surplus in such a case would write instead

$$S_r(w_p) = \lambda\, S(w_p)$$

with λ such that $\lambda\, F(w_p)w_p * h_p = M$, and M the budget available for the program.

In such a case, reducing the program wage and increasing the share of applicants to meet the budget requirement would have an ambiguous effect on the average impact on earnings. Straightforward computations show the change in the surplus due to a marginal change in w_p would be:

$$S_r'(w_p) = \lambda\ (\ h_p * F(w_p) - S(w_p)\frac{w_p f(w_p) + F(w_p)}{w_p * F(w_p)}\ )$$

A reduction in the wage w_p would not lead to a reduction as strong as in the former case. The expression also shows that there might actually be an optimum wage level. Notice that the formula can be rewritten as:

$$S_r'(w_p) = \lambda * h_p * F(w_p) * (\ 1 - \Lambda(1 + w_p f(w_p|application)))$$

where Λ is the ratio of the program impact on income to the actual transfer previously defined and $f(w_p|application)$ is the applicants’ density of the productivity level at the program wage rate, i.e. for the marginal applicant. This formula could be of some help to appreciate whether the derivative of the criterion with respect to the wage rate is positive or negative.[3]

Another potential assignment mechanism would be to randomly assign individuals to the program with different probabilities depending on some characteristic x, for instance proxies for their productivity or outside opportunities. The assignment probability would be a function $\lambda(x)$. Assume that the objective is to maximize the program impact on income

$$\mathrm{S}(w_p, \lambda) = F(w_p) \int \left(I_1(w) - I_0(w)\right) f\left(w, x | w < w_p\right) \lambda(x) dwdx$$

It is clear from the computation above that the individual gain $I_1(w) - I_0(w)$ is decreasing in w. If the productivity level can be observed, $\lambda(x)$ would only depend on w and an obvious choice for the function $\lambda(w)$ is to select individuals with the highest potential impact on income. For the same budget constraint, the best allocation would be to select participants with the least opportunities outside the program. Let $\widetilde{w}$ be the maximum productivity level consistent with full assignment and the budget constraint $F(\widetilde{w}) w_p h_p = M$. The assignment mechanism achieving the largest impact would be:

$$\lambda(w) = 1_{[w < \widetilde{w}]}.$$

It would allow to reach the largest contemporaneous program impacts:

$$\tilde{S} = \int_{(w < \widetilde{w})} (I_1(w) - I_0(w)) f(w) dw$$

It can easily be shown that $S_r(w_p) < \tilde{S}$. This leads to the conclusion that although changing contract features such as changing the wage w_p might lead to an improvement (keeping in mind it is hard to tell in which direction to adjust), the realized surplus will always be below the surplus obtained through a direct selection of those for whom the impact of the program is the largest.

[3] Similar formulas could also be derived for the number of hours h_p in the program.

Finally, if w is not fully observed but one can make a prediction of the impact at individual level (based on x), then we can consider assigning individuals based on that predicted impact:

$$S(w_p, \lambda) = F(w_p) \int E\big(I_1(w) - I_0(w)|x, w < w_p\big) f\big(x|w < w_p\big)\lambda(x)dx$$

Assignment can be based on $g = E\big(I_1(w) - I_0(w)|x, w < w_p\big)$. It is then possible to proceed as before and first assign individuals with the highest predicted impact on earnings until the budget constraint is saturated. The accuracy of the prediction is clearly a key element in the performance of the assignment mechanism.

When defining an assignment mechanism, it is important to keep in mind that the mechanism is known to potential participants, so that they can make participation decisions. Participation in the program requires that the expected gain from participation is positive. If we consider that applying costs an amount $\chi(w, x)$ (for example time to go to the registration office, time to enroll and to participate in the selection process and some related forgone earnings), the participation rule becomes:

$$\lambda(x)V_1(w) + \big(1 - \lambda(x)\big)V_0(w) - \chi(w, x) > V_0(w) \Leftrightarrow V_1(w) - V_0(w) > \chi(w, x)/\lambda(x)$$

Changing the assignment rule is likely to change the characteristics of the marginal applicant, unless there is no cost of application[4].

2.4 Other aspects of the analysis of contemporaneous program impacts

Another important aspect of the analysis of contemporaneous program impacts relates to use of the transferred income. We would expect an increase in consumption caused by program

[4] Programs frequently only advertise the number of slots available. For example, in the case of the public works program in Côte d'Ivoire we discuss in the rest of the paper, the number of slots available in each locality for men and women was known in advance. The program also initially introduced a fixed ratio of slots for women compared to men. Changing the total number of slots available or the ratio of slots reserved to women is likely to change expectation about λ for men and women. In such a case, the decision to participate might be affected, if applying is costly.

participation and higher earnings[5]. The allocation of income between consumption and savings also affects post-program impacts. Indeed, the additional income can be used to save or finance investments like training or capital for income-generating activities. Youths may encounter constraints to build up savings, and positive income shocks during the program may be associated with savings accumulation. Savings can have several potential post-program benefits, including precautionary savings to absorb future shocks, or savings to finance investments.

Finally, externalities are another important dimension of impacts during the program. Public works programs are likely to have an impact on participants themselves but also on people around them. While externalities are challenging to address empirically, it is nevertheless possible to examine the impact of participation of one member of the household on other members of the household. First, this might lead to an increase in the contribution of participants to expenses at the household level. Second, the participation of one member of the household in the program might have an impact on activities of other members of the household. On the one hand, traditional income effects would imply a reduction in other members' activity. On the other hand, underemployed members of the family could take on part of the forgone activities of the participant (especially if she was engaged in self-employment).

2.5 Post-program Impacts

A first-order question in the public works literature is about the existence and size of post-program impacts in the medium to long-term. A growing number of public works programs also have the objective to facilitate youth' transition towards more productive occupations after the program. There is little evidence in the literature on such long-term effects, although there are several potential channels through which they could unfold. First, the idea of return to capital. Several experiments have proved that returns to capital can be very large for poor households (for a review, see Blattman and Ralston, 2015). Common instruments to make capital available to youth have not proved very effective, for instance micro credit. Transfers

[5] Another question relates to the type of goods for which an increase in consumption should be expected. We could expect that program participation first increases the consumption of goods needed to meet basic needs. However, a common question when it comes to providing cash to young people is whether there are any increase in consumption of temptation goods such as alcoholic beverage, drug or gambling.

through public works programs are a way to help alleviate capital constraints for participants. A related empirical issue is whether participation in public works affects savings[6].

However, there are also other possible mechanisms for workfare interventions to have impacts in the longer run. Usually, subsidized jobs are seen as a way to improve experience, skills and productivity of participants. Raising their employability increases the likelihood that they find a wage job. This employment channel is another possibility for the program to have an impact in the long term. Lastly, we can think about behavioral aspects related to program participation. It might be the case that youth do not perceive financial constraints but that they have biased time preferences leading them to undervalue their situation in the future. It might be possible that a program requiring youths to form work habits, like waking up each morning to go to work, may induce lasting behavioral changes.

Finally, note that public works programs could also have negative long-term impacts on participants. Actually, this is a scenario that has been frequently considered. One channel for such negative long term impacts relates to the potential 'stigmatization' of participants, i.e. program participation sending negative signals to the labor market. Another possibility is that the experience provided to participants through the program is of little value or only enhances skills which are not demanded in the regular labor market. Participants may also directly forgo some activities, which may create a form of destruction of capital through program participation. Potential participants might struggle in day to day occupation requiring a lot of search and connection. They could be tempted by the 'easy' way to obtain earnings through a temporary workfare job. However, doing so can induce the loss of capital or connections which might take time to rebuild after exiting the program

3 Empirical literature on workfare programs

Despite the popularity of public work programs such as those implemented throughout much of Africa, experimental evidence on their overall effectiveness are limited (Subbarao et al.,

[6] Notice, also, that it is possible that the program has a temporary impact in the medium run. It would be the case if participants have been able to save part of the additional income due to the program but have been unable to use it to start or expand income generating activities. Savings in such a case is mainly used to cope with future income or expense shocks.

2013). Existing evidence mostly comes from quasi-experimental studies, and from a small number of programs such as those from India or Ethiopia.

As indicated in the introduction, a particularly important design feature of workfare programs is their traditional reliance on self-targeting mechanisms, dating back to Besley & Coates (1992). Early papers on workfare programs analyzed the profile of beneficiaries and benefit incidence patterns (Ravallion et al. 1993, Jalan and Ravallion, 2003). A related strand of the literature assessed the role of public works as a short-term safety net or insurance mechanism providing temporary employment and income to vulnerable populations during lean agricultural seasons or after economic shocks. Datt and Ravallion (1994) and Jalan and Ravallion (2003) estimate the net income gains from public works programs in India and Argentina, finding foregone income ranging between 30% and 50% on average. Datt and Ravallion (1994) point to differences in behavioral response across households, while Jalan and Ravallion (2003) also highlight variations in program effects along the welfare distribution. Galasso and Ravallion (2004) study how another program in Argentina affected employment outcomes and contributed to attenuate the welfare effects of an economic crisis. General findings on program impacts on welfare and food security remain mixed. Ravi and Engler (2015) find impacts of the India workfare scheme on consumption and food security, but Beegle, Galasso and Goldberg (2015) do not find significant effects on food security in Malawi in one of the few randomized control trial of a public works program so far. Gilligan et al. (2009) also find limited average welfare effects of the Ethiopia PNSP program, although households who received larger transfer amounts did see improvements in some measures of food security. Beyond welfare effects, a series of recent studies are attempting to estimate the impact of public works program on school enrolment and child labor (Li and Sekhri, 2013; Islam and Sivasankaran, 2015; Shah and Steinberg, 2015), also with mixed results.

In low-income or lower-middle-income countries, and in particular in Africa, most of the active population is engaged in low-productivity self-employment with average earnings lower than the formal minimum wage. In contrast, unemployment and formal wage employment is very limited, typically affecting 10%-15% of the active population (Filmer and Fox, 2014; Christiaensen and Premand, 2017). In this context, the extent to which public works programs trigger contemporaneous impacts on employment is unclear a priori. There are also questions on whether beneficiaries of public works programs can find pathways towards more productive post-program employment in wage jobs or in the informal sector. The evidence on mechanisms through which workfare programs affect employment in the medium to long run is particularly thin. Ravallion et al. (2005) analyze post-program impacts on earnings from a public works intervention in Argentina. Rosas and Sabarwal (2016) document investments from public works beneficiaries in assets and micro-enterprises in Sierra Leone. Deininger et al. (2016) find effects of the India public works program on agricultural productivity. A few studies assess the effectiveness of complementing public works programs with training or savings facilitation, including Galasso, Ravallion and Salvia (2004) and Almeida and Galasso (2010). Gilligan et al. (2009) report impacts of the Ethiopia public works program combined with agricultural support on adoption of agricultural technologies and off-farm small businesses.

The understanding of externalities generated from workfare programs is also particularly limited, even though such externalities are often part of the core rationale for these interventions. A series of recent studies from India analyze how the public works program affect labor markets and wages beyond program beneficiaries (Imbert and Papp (2015) and Zimmerman (2015)). Some papers have attempted to estimate the returns on the public goods created by public works programs (e.g. Deininger and Liu (2013) on land investments in India), although this remains one of the biggest gap in the literature (Alik-Lagrange and Ravallion, 2015). Finally, few studies have addressed the question on whether public works can generate social externalities by offering alternative occupations to populations in fragile or post-conflict settings and create a peace dividend. Recent exceptions include Fetzer (2014) and Amaral et al. (2015), who analyze the linkages between the Indian public works program and conflict, respectively gender-based violence.

To our knowledge, no study has analyzed whether there are trade-offs between maximizing contemporaneous and post-program benefits from public works, or between maximizing economic impacts on earnings or broader impacts on social outcomes. This paper addresses these questions.

4 Intervention and Data

4.1 The PEJEDEC public works program

Public works programs were introduced in Côte d'Ivoire in 2008 by a post-conflict assistance project following the 2003-2007 armed conflict.[7] Public works were later included as a component of an Emergency Youth Employment and Skills Development Project (*PEJEDEC*) set-up after the 2010/2011 post-electoral crisis[8]. The PEJEDEC public works program was managed by the Ministry of Social Affairs and Employment, through BCP-E[9], and implemented by the national roads management agency (AGEROUTE). A range of other institutions have been implementing public works interventions with similar features in Côte d'Ivoire[10], and more broadly across Sub-Saharan Africa

The PEJEDEC public works intervention aims to improve access to temporary employment in road maintenance for low-skilled youths in urban areas. The program targets youths aged 18-30 in 16 localities[11] throughout the country. A quota of 30% of program positions was initially reserved for women. Participants are offered temporary employment for 6 hours per day and 5 days a week for a total of six months[12]. Participants work in teams of 25 individuals (called "*brigades*"), under the supervision of a team leader and a local supervisor. They perform road maintenance activities such as sweeping roads or cleaning ditches. The jobs are paid FCFA 2,500 (approximately $5) per work day, a wage equal to the legal daily minimum wage in the formal sector. Wages are paid monthly on bank accounts that are set-up for all participants as they start working.

In addition to participating in the public works program, youths are offered various training activities. First, all participants receive a one-week basic life skills training covering issues related to HIV-AIDS, citizenship and hygiene. Second, some participants are offered a complementary *basic entrepreneurship training* to facilitate transition into more productive self-employment upon exit from the program. Third, other participants are offered a *training*

[7] *Projet d'Assistance Post Conflit* (PAPC) was implemented by the government of Côte d'Ivoire and supported by the World Bank. It was implemented between 2008 and 2014.

[8] *Projet Emploi Jeune et Développement des Compétences* (PEJEDEC) has been implemented by the government of Côte d'Ivoire (through BCP-E) and supported by the World Bank. It also included interventions for other target groups including internships, apprenticeships, professional training and entrepreneurship.

[9] Coordination Office for Employment Programs (« Bureau de Coordination des Programmes Emploi »)

[10] Among others, this includes the public works programs implemented by the government of Côte d'Ivoire as part of the C2D project with support from AFD, or as part of a program supported by the African Development Bank.

[11] 4 municipalities were covered in Abidjan (Abobo, Yopougon, Koumassi, Marcory) and 12 cities throughout the country (Yamoussoukro, Bouaké, San Pedro, Daloa, Korhogo, Abengourou, Man, Bondoukou, Gagnoa, Séguéla, Daoukro, Dimbokro).

[12] As explained further below, the program wave under evaluation lasted for 7 months, but the standard program lasts 6 months.

on wage jobs search skills and sensitization to wage jobs opportunities, with the objective to facilitate transition into wage jobs upon exist from the program.

The curricula for the complementary skills training are tailored for low-skill population that may not be able to read and write, in particular by relying on drawings and visuals. Each training lasts approximately 80-100 hours distributed over two two-week periods. They are accompanied with field exercises to be undertaken between the training periods, in parallel to the public works jobs (typically in the afternoons). The trainings are delivered by work brigades, i.e. in groups of 25 youths. Participants do not have to work during the trainings, but still receive their corresponding daily wage[13].

The *basic entrepreneurship training* aims to build skills to help youth set-up and manage a small non-agricultural micro-enterprise. The training lasts 100 hours and focuses on providing cross-cutting business skills and practical guidance to develop simple business plans for small-scale activities that can be set-up using savings from the public works program. A first phase (40 hours over two weeks) reviews themes related to basic entrepreneurship and business skills. A second phase includes field research for youths to gather information, undertake basic market research and sketch a business plan. A third phase (40 hours over two weeks) includes feedback on youths' basic business plans, and reviews of key related issues from the curriculum. The final phase (20 hours) is an individual post-training follow-up[14].

The *training on wage jobs search skills and sensitization to wage jobs opportunities* provides information on wage jobs opportunities, skills on jobs search techniques, as well as a more professional environment during the public works programs and skills certification to facilitate signaling upon exit from the program. The training itself lasts 80 hours. The first phase (40 hours over two weeks) reviews how to identify wage jobs opportunities (either locally or through migration), how to search for wage jobs, prepare a CV, apply for a job and participate in a job interview. The second phase includes field exercises to gather information on potential opportunities, identify and visit potential employers or professional networks, etc. The third phase (40 hours over two weeks) provides feedback on field exercises, reviews part of the curriculum and provides additional practical guidance. In addition, supervisors of the brigades who were offered the wage employment training were also trained on how to manage teams and provide feedback to workers, with the objective to mimic the professional experience one

[13] Some youths were offered the second half of the training after their exit from the public works program. While these youths were not paid during that time, they received a small stipend to cover transportation costs.
[14] The evaluation policy report has additional information on the scope of the training (Bertrand et al. (2016)).

would have in a more formal wage job. Youths were periodically rated on a range of skills, and these evaluations were later used to issue a work certificate that signaled between one and five competencies identified as strengths for each participant[15].

4.2 Experimental design: Enrollment and Randomization

The PEJEDEC public works program was implemented in 16 urban localities throughout Côte d'Ivoire, including Abidjan and cities in the interior[16]. Four waves were organized between 2012 and 2015, each covering all 16 localities, with a similar number of pre-determined places available for each locality in each wave. In total, 12,666 youths participated in the program. The randomized control trial focuses on the second wave of the program, which took place between July 2013 and February 2014[17]. The identification strategy relies on a two-step randomization process.

The first step involves individual randomization into the program. Before the start of the second wave, and as was the case for the other waves, an intense communication campaign was organized by the implementing agency (AGEROUTE) through local newspapers, local radios and public notice boards to invite interested youth to visit a registration office and apply to the program. Enrollment was open for two to three weeks in each locality, between June 2013 and July 2013. Only two eligibility criteria were applied during enrollment: applicants had to be between 18 and 30 years old, and could not have participated to the public works program before.

Once the enrollment period had closed, public lotteries were organized in each locality (separately for men and women, hence stratified by locality and gender) to randomly select beneficiaries among the registered applicants present at the lottery. Remarkably, the public lotteries were put in place at the time of the post-conflict assistance project. Since then, they have been used continuously as a transparent assignment mechanism to allocate limited public works jobs in a way that would be socially acceptable and limit potential tensions. As such, the

[15] The evaluation policy report has additional information on the scope of the training (Bertrand et al. (2016)).

[16] See footnote of section 4.1.

[17] Less than 5% of youths assigned to the public works program did not participate, or participated for less than 3 months. The second wave was extended from six months to seven months for all participants to ensure that the complementary training could be completed during that time for those who were assigned to them (see below).

first step of the randomization protocol was already implemented by the program in its routine operations.

In practice, during the enrollment phase for the second wave of the program, 12,188 individuals applied, of which 10,966 participated in public lotteries where 3,125 beneficiaries were selected and assigned to 125 brigades of 25 individuals each (17 men, 8 women)[18]. For the study wave, a waiting list was created to protect the control group, although in practice replacements were minimal[19].

The second step involves the randomization of public works brigades into groups receiving different types of complementary training. Specifically, brigades were randomized into three groups: (i) 45 brigades (1,225 individuals) were assigned to receive the public works only; (ii) 40 brigades (1,000 individuals) were assigned to receive the public works plus the complementary basic entrepreneurship training, and (iii) 40 brigades (1,000 individuals) were assigned to receive the public works plus the wage jobs search skills training[20]. This second randomization was stratified by locality, and performed through a lottery held in the project office with implementing partners and a notary public in November 2013. Results remained confidential until two weeks before the start of the trainings.

4.3 Timeline and Data

4.3.1 Timeline and Surveys

The randomized control trial focuses on the second wave of the public works program. The public lotteries were held in each locality between the end of June and early July 2013, right after the end of the enrollment.

A baseline survey was conducted shortly after the public lotteries and before program implementation (between the end of June and mid-July 2013). The study sample comprised all

[18] Beneficiaries were assigned to brigades within localities based on the number they drew in the public lottery.

[19] Replacement of drop-outs was allowed during the first two-months of the program. Replacements were only possible based on the waiting list, and had to be stopped when the waiting list was exhausted. After two months, replacements were not allowed anymore. This ensured that individuals in the control group were not offered the program during its implementation.

[20] All brigades receive a one-week basic life skills training covering issues related to HIV-AIDS, citizenship and hygiene. This training is considered part of the basic public works program.

the individuals selected to participate in the program after the first randomization (3,125 individuals), as well as a control group obtained from a (random) sample of 1,035 individuals drawn among the non-beneficiaries not on the waiting list[21]. The data collected included measures of employment and earnings. It also captured a range of other characteristics such as preferences for risk and present, behavioral skills, as well as the results of practical tests measuring cognitive, manual and numeracy skills. Attrition at baseline was very small (1.5%).

The public works activities started between early and late July 2013, depending on the locality. In August, participants received the one-week life skills training that is considered part of the basic public works program. The second randomization took place in October 2013 in the main project office. Brigades' assignment to the various complementary training modalities were not made public until January 2014, in order to limit potential response bias during the midline survey.

In addition to the baseline survey, two surveys were conducted in order to study contemporaneous impacts *during* the program and medium-term *post*-program impacts.

To estimate contemporaneous program impacts, a midline survey was conducted on 3,036 individuals (2,001 beneficiaries[22] and the control group) between the end of November 2013 end early January 2014, i.e. 4 to 5 months after the start of the program. Both individuals and household heads were interviewed. A two-weeks tracking phase was implemented in February 2014 to limit attrition, mainly due to the mobility of control individuals[23]. Attrition at midline is limited (2.6%) and balanced across treatment and control groups. The midline questionnaire included very detailed sections on employment (up to three activities), specific information on

[21] Individuals randomly assigned to the waiting list are excluded from the sampling frame to prevent contamination of the control group. Sampling for the control was stratified by gender and locality (similar to the randomization procedure).

[22]The 2,001 treated individuals are a sub-sample of the 3,125 beneficiaries stratified by locality, brigade and gender. This sub-sample voluntarily excludes brigades which had been allocated to the wage employment training. Indeed, their supervisors were following a specific management training at the time of the survey, and we wanted to avoid potential anticipation effects or any behavioral changes that could potentially affect outcomes.

[23] The tracking helped reduce attrition rate from 5.4% (after main data collection) to 2.6%. Before tracking, a small attrition differential was observed between treatment and control groups due to larger mobility out of program localities in the control group. The survey firm had not planned tracking outside the localities. The sample for tracking was randomly selected among the treatment and control groups (stratified by locality and gender) among non-respondents who were alive, not outside Côte d'Ivoire, and excluding individuals that could not be reached since baseline. After tracking, remaining attritors were mainly people impossible to contact or highly mobile individuals, and attrition was balanced between treatment and control groups.

characteristics of independent activities, a time use module and measures of behavior and well-being.

The public works program was originally expected to end in January 2014. However, as the complementary trainings were starting in January, participants were given a one-month extension on their contracts, which exceptionally extended the public works duration from 6 to 7 months. This ensured that all 'brigades' selected to participate in one of the trainings could do so while being paid by the program (at the same wage) for the first half of the training, which reduced the opportunity cost of time during the trainings. Complementary trainings were organized between January and mid-March and the second wave of the program ended between early and mid-February 2014 (depending on the locality). Some beneficiaries attended the second half of complementary skills trainings after the end of their contracts, and were given a transport allowance[24].

To evaluate post-program impacts, an endline survey was conducted between March and July 2015, i.e. between 12 to 15 months after the end of the program. The sample included 4,360 individuals. It was comprised of the whole baseline sample of 4,160 individuals in the treatment and control groups, plus 200 individuals randomly selected to be added to the control group[25]. Again, both individuals and household heads were interviewed. A one-and-a-half month (random) tracking phase took place in September 2015. The final attrition rate was 6.2%, and was balanced between treatment and control groups. The endline questionnaire was based on the midline survey and enriched with 'historic' information on job search, independent activities (including past projects) and an employment calendar.

4.3.2 Key outcomes and descriptive statistics

Descriptive statistics

[24] Half of the brigades assigned to complementary skills trainings (50% of each type of training) had 25% of their training hours after the end of the public works contract, and received a transportation allowance of FCFA 1500 (the program wage was FCFA 2500). 25% of the brigades had 25% of their training hours (i.e. the second phase of the training) after the end of the public works contract and received the same transportation allowance. The remaining 25% were fully under contract during their trainings. Transportation allowance was paid ex-post in one transfer, based on the actual number of days attended.

[25] The replenishment of the control group is explained in section 4.4.1.

As in many countries in Sub-Saharan Africa, Côte d'Ivoire faces a relatively low unemployment rate, but also a small share of individuals working in wage jobs. A large part of the population is concentrated in informal occupations, mainly in agricultural and non-agricultural self-employment (Filmer et al, 2014; Christiaensen and Premand, 2017). In addition, most of wage employment takes place in casual and informal jobs without contracts. Overall, many earn less than the legal minimum wage, as the regulations are only binding for formal private companies and public administration (INS and AGEPE, 2014). Inactivity and unemployment tend to be more frequent for individuals in households in the top of the wealth distribution, especially those holding a higher education degree. This reflects the fact that the poor and vulnerable often cannot afford not to work. Moreover, gender disparities are strong. Women are more likely to be inactive and unemployed compared to men. They are also more likely to be self-employed rather than in wage jobs.

Public works applicants are on average 25 years-old, and mostly live in urban areas (93%). They live in households with an average of 6 individuals (with 4 adults). 25% of applicants are head of the household and most of applicants have no more than one child (50% of applicants have no children, 25% have one child). Three quarters of them attended (at least partially) primary school, but around half of the applicants (47%) have no degree[26]. This reflects the fact the program was designed to attract low-skilled youths, although 11% of the applicants have completed secondary school. Less than half of the applicants have attended some form of vocational training, mostly informal apprenticeships. 80% of applicants were already working before the program, in line with the national employment situation marked more by underemployment in low-earning occupations rather than unemployment. Also, although most applicants report searching for wage jobs, most of them declare aspiring to be self-employed in the future. Finally, the data also points to limited financial resources, as only half of the applicants have saved money over the last three months and nearly 75% of them report facing constraints for basic needs expenditures.

We compare our evaluation sample to a national sample of individuals who are 18 to 30 years old and live in urban areas[27] to provide insights on public works participants' profile (Table 1). Overall, the program attracts a lower share of inactive and unemployed individuals compared to all youth aged 18-30 in urban Côte d'Ivoire. Among the employed population, public works

[26] At the end of primary school students pass a certifying exam (CEPE).

[27] We use the 2013 national employment survey (ENSETE) which data was collected in February 2014 and compare it to our closest dataset, the control group of the midline survey which occurred from November 2013 to January 2014.

applicants are more likely to hold wage jobs (in their main occupation) rather than be self-employed. The educational attainment is quite similar among the general population and public works applicants. Individuals with a relatively high level of education were also attracted by the program, even though it had been originally been conceived for low-skilled youth.

Key outcomes

This section describes the main outcomes measured both at midline and endline surveys.

Total monthly earnings are expressed in CFA francs. They are aggregated over up to three (parallel) activities undertook by an individual in the 30 days preceding the survey. They include payments received in cash and the monetary equivalent for in-kind payments. The variable is winsorized at 99%. Total monthly earnings are decomposed in total (monthly) earnings from wage employment and self-employment (as well as earnings from other occupations, which are not displayed separately).

Has an Activity is a dummy taking a value of 1 if the individual has worked at least one hour over the 7 days preceding the survey, consistent with the official employment indicators used in Côte d'Ivoire. It takes a value of 0 for inactive and unemployed individuals. We also report having at least one wage-job (*Wage employed*) and at least one self-employment activity (*Self-employed*), which is a decomposition of this outcome.

Weekly hours worked capture the total number of hours worked per week. It is aggregated from up to three (parallel) activities undertook by an individual across all occupations (wage employment, self-employment or other type of activity). The variable is winsorized at 99%. Weekly hours worked are decomposed in hours worked in wage employment and self-employment (as well as hours worked in other occupations, which are not displayed separately).

Savings stock is the total amount of savings in CFA francs at the time of the survey. It aggregates savings from formal or informal mechanisms. The variable is winsorized at 99%.

Total expenditures aggregates several types of expenditures made by the youth whether the expense would benefit himself of another member of his/her household. It includes basic expenses (health, clothing, sanitation, and accommodation), communication (mobile, internet, and medias), investment type (education, training, maintenance of assets), transportation, temptation goods (alcohol, tobacco, gambling, and luxury goods) and social expenses (celebrations and charity). The variable is winsorized at 99%.

A well-being index aggregates 6 measures: measures of happiness and pride in daily activities taken from a time-use module, the Rosenberg[28] self-esteem scale, a 'positive affect' sub-scale (from the CESD scale[29]), a sub-scale of (positive) attitude towards the future (from the ZTPI scale[30]), and a sub-scale of (internal) locus of control (the inverted 'fatalist present' sub-scale from the ZTPI scale).

A behavior index aggregates 6 measures: an inverted measure of anger or frustration in daily activities taken from the time-use module, an inverted measure of impulsiveness, an inverted 'conduct problem' sub-scale (from the SDQ scale[31]) and a (positive) 'pro-social behavior' sub-scale (from the SDQ scale).

The well-being and behavior indices are z-score, with a mean set to zero and a standard deviation to one for the control group, so that estimated coefficients can be interpreted in standard deviations. A positive impact on the well-being index is interpreted as an overall increase in well-being and a positive impact on the behavior index as an overall improvement in attitudes.

4.4 Empirical Methodology

4.4.1 Main specifications

We estimate intent-to-treat (ITT) effects for contemporaneous and post-program impacts for the pooled treatment via an ordinary least squares (OLS) regression:

(1)

$$Y_i = \alpha \ + \beta \, W_i \ + \delta \, X_{i,l} + \epsilon_i$$

[28] The Rosenberg Self Esteem scale includes 10 items and measures self-esteem. We use the validated French version.

[29] The Center for Epidemiologic Studies Depression (CESD) scale was specifically developed to measure depression. It also includes an inverted scale that measure positive feelings ("Positive Affects").

[30] The Zimbardo Time Perspective Inventory (ZTPI) is an instrument measuring time perspective for individuals in different dimensions. In particular, we use the two dimensions of "future" (to have a positive attitude towards future) and "fatalist present" which is very close to the concept of internal locus of control. We use the validated French version.

[31] The Strength and Difficulties Questionnaire was initially created to measure behavioral issues for young children and teenagers (3 to 16 years old). We use two sub scales (out of five): "conduct problems" and "pro-social behavior".

where Y is an outcome for individual i, W is an indicator for treatment (being assigned to the public works program at first randomization), and X is a vector of stratification variables (locality, gender). Robust standard errors are clustered at the level of "large" brigades for treated individuals[32].

To estimate post-program ITT effects across treatment arms, we use the following specification:

(2)

$$Y_i = \alpha \ + \beta_1 W_i \ + \beta_2 \ (W_i * T1_i) \ \ + \beta_3 \ (W_i * T2_i) \ \ + \delta_1 \ X_{i,l} \ + \epsilon_i$$

where $T1$ (respectively $T2$) is an indicator for being assigned to the complementary self-employment training (respectively wage employment training). Coefficient β_1 estimates the impact of the 'pure' public works while the coefficient β_2 estimates the additional effect of the self-employment training and β_3 the additional effect of the wage employment training. $\beta_1 + \beta_2$ (respectively $\beta_1 \ + \beta_3$) capture the total effect of the program for individuals assigned to public works and complementary self-employment training (respectively wage employment training). In the results table, we also provide the p-value for the test that this sum is equal to zero.

We analyze heterogeneity in treatment effects by groups G determined by a set of baseline characteristics Z (see discussion in section 6). The specification used in that case is the following:

(3)

$$Y_i = \ \alpha + \ \gamma_1 (W_i * \ G_i) + \ \gamma_2 (W_i * (1 - G_i)) + \gamma_3 * G_i + \delta_2 \ \mathrm{X}_{i,l} + \epsilon_{\mathrm{i}} \ ,$$

where G is an indicator for belonging to the group determined by Z. We are interested in coefficient γ_1, which estimates the impact of the pooled treatment for a specific group G. We

[32] We suspect within-brigade error correlation due to the interactions between treated individuals who worked together in the same brigade for several months. Brigades are sometimes aggregated to account for the fact that some individuals have been moved across brigades during public works implementation for various reasons: when such movement occurred, we group the different brigades together.

provide standard errors for $\widehat{\gamma_1}$and the p-value for the test that $\gamma_1 + \gamma_2$ is equal to zero at the bottom of the tables.

For specifications (1) to (3), we use probability weights. They are composed by up to five[33] multiplicative weights to account for (i) public lotteries specificities in the first randomization, (ii) locality specificities in the second randomization, (iii) the sub-sampling of non-respondents during tracking surveys, (iv) the sub-sampling of the treated group for midline survey and (v) cases of control individuals who enrolled and eventually participated in later waves of the program[34].

Potential threats to internal validity

We check for balance across treatment and control groups in Table 2. Column (3) contains the p-values for the test of difference between treatment and control groups, and column (4) for the test whether all treatment arms are jointly equal to zero. Overall, there are no meaningful differences across groups[35]. We note that collecting the baseline survey after assignment to the program affected a few variables for which control or treatment groups may have had incentives to over or under report. Based on that, we do not add baseline controls on top of stratification controls to our specification.

At midline, compliance to program assignment was high. Only 6 control individuals in the sample succeeded in 'cheating' the public lottery (by registering for the program in different locations) and only 2 generated contamination after being selected for the program. Among youth assigned to the public works, take-up was high as 93% of them participated for more than five months out of the seven months of implementation. In total, youth worked an average of 141 days, out of a maximum of the 154 days expected in case of full attendance.

An unforeseen issue emerged at endine. Between the end of the phase of the program being evaluated and the time of the endline survey, a few individuals from the control group were

[33] With midline data, we use weights related to (i), (iii) and (iv). With endline data, we use weights related to (i), (ii), (iii) and (v). More details on weight construction provided in Appendix A.

[34] This specific point is detailed in the next paragraphs and in Appendix A.

[35] We also checked the balance across groups for both midline and endline respondents.

able to enroll (and eventually participate) in the third or fourth wave of the program[36]. However, administrative data about enrollment and public lotteries enabled us to observe control individuals' behavior for the third and fourth wave (i.e. whether control individuals applied or not to the program, and whether they were selected and participated or not). For post-program impact analysis, control individuals who (later) participated to the program are not included. To adjust for this, control individuals who also applied in future waves but were not selected through the public lotteries are assigned relatively larger weights[37]. 200 individuals were also (randomly) added[38] to increase the total size of the control group.

Finally, the take-up of complementary training is lower than the take-up for the 'pure' public works: 72% of individuals assigned to self-employment training and 67,2% of those assigned to wage-employment training attended at least 75% of the training (i.e. at least 60 hours out of the 80 total hours). This level of take up is aligned with take-up observed in other skill training programs. For both trainings, only 10% of individuals never attended.

4.4.2 Heterogeneity analysis using machine learning

Machine learning and heterogeneity of treatment effects

The use of machine learning methods in this paper comes from the desire to understand the heterogeneity of treatment effects, in an experimental setting with rich and multi-dimensional baseline data. In particular, we would like to assess how much the treatment effect would vary under alternative assignment rules (based on observable characteristics), and especially if larger treatment effects could be obtained by targeting specific sub-groups compared to average treatment effects estimated across all participants. Understanding how much different types of individuals benefit from the program could help policy makers define a more efficient selection rule and maximize the impact of the program. In addition, analyzing heterogeneity can improve

[36] We identified 140 individuals from our baseline control group (i.e. 13,5%) among next waves beneficiaries (91 for third wave, 49 for fourth wave). Among the 200 individuals randomly added, 30 were also identified to be among third or fourth wave beneficiaries.

[37] Specifically, we use specific weights to account for the behavior of enrollment to the waves and put a zero-weight on control individuals who ever received the program (see Appendix A).

[38] To keep the same stratification as the initial control group sample, we selected the 20% next ones of each (baseline) lotteries' list.

our understanding of the mechanisms: across outcomes, some subgroups can highly benefit in terms of earnings but not in other dimensions and vice versa.

When looking at heterogeneity in treatment effects, we are traditionally interested in estimating the coefficient of the interaction between treatment and a binary indicator for a group, in a linear specification. However, looking at subgroups defined by one or two interaction terms is rather limited, and searching intensively for subgroups may lead to spurious conclusions that will probably not hold once one accounts for multiple hypothesis testing. A different approach would be to consider the identification of heterogeneous treatment effects as a *prediction* problem: mapping individual characteristics to treatment effect estimates (Mullainathan and Spiess, 2017). In fact, a recent literature has been reconsidering the importance of prediction in policy issues (while the focus is on causal inference) and how machine learning could contribute to solve these “prediction policy problems” (Kleinberg et al., 2015). These range from police hiring and teacher tenure decisions (in which predicting productivity can inform decisions) to crime risk prediction to help judges’ decisions in bail-type problems (Chalfin et al., 2016; Kleinberg et al., 2017).

Targeting is an area where machine learning techniques are potentially extremely useful. If policymakers have a clear assignment rule to implement (e.g. target the poorest or those at higher risk), the challenge is to accurately predict the corresponding outcome (wealth or risk, to follow the same examples) in order to improve the targeting of a given program. Machine learning tools typically outperform standard regression tools at this task[39]. Another way machine learning can inform targeting is to understand if among eligible people some subgroups could benefit more than others. This is the approach followed by Davis and Heller (2017a and 2017b), using machine learning to look at heterogeneous treatment effects across different outcomes of interest, in a large-scale employment program for disadvantaged youth. Based on such analysis, policy makers could adapt the selection rule to target individuals predicted to be those who benefit the most, and therefore maximize program impacts.

Using causal forests to predict (conditional) treatment effects

[39] For example, McBride and Nichols (2016) shows how machine learning algorithms can improve poverty targeting compared to Proxy Means Tests.

In supervised machine learning methods[40] (e.g. regression trees, random forests), a model of the relationship between a set of features (X) and an observed outcome (Y) is first built by training the model on a dataset where (Y, X) are observed. That model is subsequently used to predict Y on a population for which only the characteristics X are observed. Many applications can be found in the fields of medicine or more recently online marketing. Such an approach could be applied to the prediction of treatment effects (instead of the outcome Y) using for X a set of covariates measured at baseline (before treatment). In particular, algorithms such as regression trees or random forests can help us predict treatment effects conditional on a set of characteristics X without making assumptions on which Xs or combinations of Xs might be relevant, or on the functional form of this heterogeneity (i.e. whether it is linear or not[41]). Given the prediction model built, we would be able to predict treatment effects for any set of characteristics included in X and simulate the treatment effects of the program under alternative selection rules (based on X)[42]. Using predictions for different outcomes, we could also look at the variation of treatment effects across outcomes.

Actually, adapting the prediction of outcomes to treatment effects is not straightforward. The main challenge is that machine learning algorithms require to assess the quality of predictions: this is relatively easy when predictions can be directly compared to realizations of the outcome for each individual; however this is 'infeasible' when only one potential outcome is observed (fundamental problem of causal inference). Athey and Imbens (2016) build a framework and propose a new algorithm - causal tree - which adapts classification and regression trees (CART) to the specific case of predicting treatment effects conditional on a set of X. Causal trees differ from CART in two main aspects[43]: (i) the splitting criteria has been adapted to maximize the variation in treatment effects across leaves (rather than the variance of the outcome), (ii) an independent sample (different from the sample used to partition the data) is used to estimate treatment effects within each leaf ("honesty" property) which should eliminate some overfitting. This new sample splitting ("honesty") also makes inference on treatment effects valid (conditional on a tree). Wager and Athey (2016) extend this framework to random forests based on causal trees and the possibility to make causal inference with the obtained results[44].

[40] See Hastie, Tibshirani and Friedman (2009) for their detailed review of these methods.

[41] In particular, random forests can detect nonlinear heterogeneity that could not be identified using LASSO.

[42] More details on machine learning procedure in Appendix B.

[43] Appendix B.2.3 provides more details on the differences between causal trees and CART.

[44] Random forests were introduced by Breiman (2001). See Wager, Hastie and Efron, 2014 for the computation of confidence intervals for random forests.

We rely here on the framework developed by Athey and Imbens (2016). We want to estimate the conditional average treatment effect (CATE): $\tau(x) = E[\, Y_i\,(1) - \; Y_i\,(0) |\, X_i^K = x]$ with X^K a vector of K baseline covariates (features) and Y the outcome of interest. This requires a dataset $\left(Y_i{}^{obs}, W_i, X_i^K\right), i = 1 \ldots N$, considered as an i.i.d. sample drawn from an infinite population, where W is a binary indicator for treatment and Y^{obs} the realized and observed outcome, that is to say the potential outcome corresponding to the treatment received. We also need to make the assumption of 'unconfoundedness' (e.g. treatment is randomized conditional on observable covariates), which is reasonable given our randomized experiment.

In this paper, we implement the causal forests algorithm[45] to recover a prediction of $\tau(x)$. In a nutshell, *causal* forests are random forests of *causal* trees, the latter being standard trees adapted to the case of treatment effect estimation as in Athey and Imbens (2016).

First of all, as in most machine learning methods, our data is split in two separate samples: a training sample used to build the model, and a test sample. This is fundamental as all subsequent analysis will be performed on the test sample, a sample that is not involved at any step in the construction of the prediction model. Given our dataset (small[46] N compared to K), this will be a constraint in terms of sample size.

Then, the model giving a prediction of $\tau(x)$ is built using a training sample set to 50% of the full sample N (standard). In a second step, $\tau(x)$ is estimated on the test sample (set to 50% of N). This is applied to both midline and endline surveys, for our main outcome of interest total monthly earnings (see section 4.3.2), and with $x \in X^K$ a rich set of 101 baseline covariates comprising individual and household characteristics, education, employment and savings dimensions, assets held, measures of personality, preferences, ability and cognitive skills. Appendix B provides extensive details on the construction of the model and the covariates used.

To discuss heterogeneity of the impacts, we report the mean of $\hat{\tau}(x)$ over each quartile on the test sample. To document heterogeneity we will focus on the upper and lower quartile (respectively top and bottom 25% groups). In particular, impacts at the top quartile of the distribution indicate the potential gains in treatment effects when selecting individuals based on a large set of characteristics.

[45] We thank Susan Athey and Stefan Wager for graciously sharing their code with us. We use causalForest function from causalTree package in R, with small modifications. See Appendix B.

[46] The sample size varies depending on the survey used for the algorithm and ranges from 2,884 to 3,910 units. Appendix B.1 provides details on the sample used for causal forests.

5 Results: ITT Estimates for Public Works Impacts

This section presents ITT estimates for public works impacts on main outcomes for youths, based on the pooled treatment specification in equation (1). The ITT estimates are discussed separately for contemporaneous impacts (3-4 months after the start of the program, while youths are still participating) and post-program impacts (12-15 months after youths have exited from the program). The next section discusses mechanisms for post-program impacts and impacts by treatment arms (using specification in equation (2)).

5.1 ITT Estimates for Main Outcomes

Table 3 (Panel A, columns 1-6) presents contemporaneous ITT estimates on employment and hours worked. 86% of youths in the control group have an activity at midline, with 53% of youths holding a wage job, and 33% self-employed. This is consistent with the employment situation in Côte d'Ivoire and much of Sub-Saharan Africa, where a large numbers of youths are underemployed by working in low-productivity occupations, often in self-employment or informal wage jobs paying less than the minimum wage in the formal sector, and few are formally unemployed. In this context, impacts during the program on the overall level of employment are limited for youths in the sample (+12 pp). The stronger employment effects stem from a change in the composition of employment, with strong impacts on the share of youths holding wage jobs (+44 pp) driven by the public works jobs, and a smaller decrease in self-employment (-9 pp). Similar patterns are observed for contemporaneous program impacts on hours worked per week, with a small overall increase in total hours worked (by 3.5 hours from on an average of 41 hours per week in the control group). This is driven by a large increase in hours worked in wage employment (+14 hours) and smaller decrease in hours worked in self-employment (-6.7 hours). Employment in the public works program accounts for approximately 30 hours a week for individuals in the treatment group, so that the small increase in overall hours worked in fact hides a large decrease in hours worked in other activities. Youths in the treatment group also become more likely to cumulate various activities[47]. Overall, the observed

[47] It is estimated than the share of youths holding multiple activities increase in the treatment group, see Bertrand et al. (2016).

contemporaneous program impacts on employment raise questions on the effectiveness of the self-targeting mechanism in a context where most youths are working and have to rearrange their activities to make time to access the better public wage jobs offered by the program.

Table 3 (Panel B, columns 1-6) presents post-program ITT estimates on employment and hours worked. Despite strong shifts in youths' employment portfolios during the program, no post-program impacts on employment level, employment composition or hours worked are observed. Youths in the treatment group display a similar employment profile than youths in the control group, with no statistical difference in the main indicators for employment and hours worked. On the one hand, these results show that the public works program did not lead to "stigmatization" or "scarring" effects for youths. Past participants are not less likely to be employed (including self-employed or holding wage jobs) a year after their exit from the program. This suggests a relatively rapid adjustment back to the pre-program occupations. On the other hand, results also show that the public works does not bring longer-term benefits to youths in terms of employment types or hours worked. An important exception, however, relates to the earnings and productivity in these occupations, to which we now turn.

Table 3 (Panel A, columns 7-9) presents contemporaneous ITT estimates of program impacts on earnings. The public works leads to a net increase in earnings of FCFA 20,885 per month (or approximately $42) during the program. The net earnings gains represent a 35% increase from the level of earnings in the control group (FCFA 60,052, or $120), or approximately 42% of the average net monthly transfer amount (FCFA 50,600, or $101[48]). As such, the estimated effects point to substantial foregone earnings from activities that youths left or scaled down in order to participate in the program. As for employment patterns, contemporaneous impacts on earnings stem from the strong increase in earnings from the wage jobs offered by the program (+FCFA 35,385, or $71), with a significant decrease in earnings from self-employment (- FCFA 12,625, or $25).

Table 3 (Panel B, columns 7-9) presents post-program ITT estimates of program impacts on earnings. The public works leads to a small but significant increase in earnings a year after the end of the program (+FCFA 5,622, or $11 per month), a 11.6% increase from the level of earnings in the control group. The increase in earnings is concentrated in self-employment, where earnings increase by FCFA 6,223 ($12.4) per month, or substantial relative increase of

[48] This figure is the average amount transferred over all individuals assigned to the public works (independently of non-compliance and days not worked).

32%. On the other hand, post-program earnings in wage jobs are not statistically different between the treatment and control groups. These results show that, while participants were not more likely to be employed, employed in different occupations, or working longer hours, they were on average running more profitable self-employment activities one year after exiting from the program.

To go beyond traditional economic indicators, we also discuss the impacts of public works on indices of well-being and behavior. The consideration of broader well-being indicators are important as the temporary jobs offered by the program may have indirect benefits or costs beyond economic dimensions. On the one hand, the public works activities are hard manual labor activities, which some may consider depreciating. On the other hand, there can be a certain status associated with holding a public wage job in the community, in particular a predictable and secured formal wage job. Changes in youths' well-being and behavior are particularly relevant in a post-conflict setting such as Côte d'Ivoire, including as they point to potential program externalities on social cohesion, an issue of strong interest for policymakers.

Table 3 (Panel A, columns 12-13) presents ITT estimates of contemporaneous program impacts on indices of well-being and behavior (see section 4.3.2 for definition). Results show significant contemporaneous program impacts on the well-being index (+0.2 standard deviations), as well as the behavior index (+0.13 standard deviations). Improvements in well-being while youth participate in the program come from a larger share of youths reporting feeling happy and proud, scoring higher on sub-scales for self-esteem, positive affect and positive attitude towards the future, as well as reporting higher present and future life satisfaction[49]. Improvements in behavioral dimensions are more limited, but point to less anger and frustration in daily life, and less impulsiveness, although no related changes in other domains such as pro-social behavior and conduct problems are observed. These results may be associated with the economic gains mentioned earlier. They also raise the possibility that some youths who do not benefit substantially in economic dimensions may nevertheless benefit from the program in psychological or behavioral dimensions.

Table 3 (Panel B, columns 12-13) presents post-program ITT estimates on indices of well-being and behavior. Some lasting improvements are observed on psychological well-being for youths in the medium-term (0.09 standard deviations), although they are more muted than during the program. They are also concentrated in a narrower set of domains such as happiness, self-

[49] In contrast, present fatalism is unaffected.

esteem and present life satisfaction. In contrast, there are no lasting impacts on sub-scales for pride, positive affect, positive attitude towards the future and future life satisfaction. A year after the end of the program, no lasting impacts are observed in the behavior index or any of its subcomponent, suggesting that short-term behavioral gains have faded.

5.2 Mechanisms for Post-Program Impacts

Intermediary Outcomes: Short-term Expenditures and Savings

Table 3 (Panel A, columns 10-11) presents contemporaneous ITT estimates on expenditures and savings, to complement the estimated impacts on earnings documented earlier. During the program, the observed increase in earnings (+FCFA 20,885 per month, or $42 as mentioned above) translate into an increase in both expenditures and savings. Total monthly expenditures are estimated to increase by FCFA 15,085 ($32), constituting approximately 70% of the nets earnings gains. The overall increase in expenditures can be decomposed in roughly equal shares between youths' own expenditures and their contribution to household expenditures. The additional expenditures are mostly for basic necessities (food, clothes, …), as well as education and training.

Yet, beyond consumption support, youths are able to save a significant share of their net earnings gains. On average, after about 4 months in the program, youths in the treatment group have increased their stock of savings by approximately FCFA 39,633 ($79). This impact is of large magnitude as it corresponds to a 182% increase from the average stock of savings in the control group (FCFA 21,752, or $43). The order of magnitude is also consistent with youths saving approximately 30% of their earnings gains. Importantly, youths are not only more likely to save and to save larger amounts, but most of these savings are kept in formal bank accounts. These include accounts in which youths are paid their public works wages. Overall, these substantial contemporaneous increases in savings are likely to contribute to the post-program effects on earnings in self-employment. Indeed, post-program results show that youths are not more likely to be self-employed, but are more likely to operate micro-enterprises with a relatively larger asset stock and scale of operations, pointing to higher productivity. Youths in the treatment group also report higher investments in household enterprises, which likely have been facilitated by savings from the program and are consistent with observed higher earnings in self-employment.

Impacts by Treatment Arms

Table 3 (Panel C) further illustrates the contribution of the various treatment arms to disentangle the causal impacts of complementary skills training from the average post-program impacts documented so far. Overall, little variation in impacts across treatment arms is observed, suggesting limited value-added of the complementary skills training. Specifically, post-program impacts on employment level, employment composition and hours worked are very consistent across the different treatment arms[50]. One noteworthy distinction relates to the post-program impacts on earnings. Post-program impacts on total earnings are mostly observed for the groups of individuals that were assigned to the basic entrepreneurship or jobs search skills training. In particular, post-program impacts on self-employment earnings mostly come from youths in the treatment group that was assigned to the basic entrepreneurship training. Still, the impacts on earnings are not statistically different between arms, so that we cannot reject equality of the impacts on earnings across groups. The only difference in earnings that is significant is when comparing self-employment earnings across individuals assigned to the basic entrepreneurship training and individuals affected to the basic public works only. Ultimately, since there are no statistical differences in impacts on overall earnings across treatment arms, we pool treatment to conduct finer heterogeneity analysis in the rest of the paper.

The limited value-added of the complementary training suggests that skills acquisition through these trainings is not the main mechanism that explains the post-program impacts. In fact, the trainings were effective in improving knowledge in basic entrepreneurship, respectively jobs search skills, as they intended to do. They also led to youths applying these skills in practice, either by intensifying their search for wage jobs (e.g. using a CV for a job search, searching using adds or by applying independently) or their efforts to set-up a new activity (e.g. by undertaking a market study or a preparing a business plan)[51]. However, these changes in skills and practices were not sufficient to generate earnings beyond those generated by the basic public works program.

[50] The only exception is that hours in self-employment are significantly larger in the jobs search training arms compared to the public works only arm. Still, the coefficient is not statistically different from 0 or from the estimate from the basic entrepreneurship training arms.

[51] We document these two aspects of trainings impact (learning during the training and trying to put this intro practice) in the policy report, Bertrand et al. (2016).

6 Heterogeneity Analysis

The public works program was oversubscribed, with the number of applicants exceeding the number of available program slots by a ratio of 4 to 1. While participation in the program was the result of a random assignment process[52], which has the advantage of being fair and transparent, the performance of the program might have been improved with a better targeting of the 25 percent of program beneficiaries among applicants.

Whether and by how much alternative targeting might have improved program effectiveness depends on the magnitude of heterogeneity in program impacts. In theory, given the self-selection mechanism, we would expect heterogeneity in impacts among program applicants, with marginal applicants experiencing very limited gains in earnings compared to infra-marginal applicants with more limited employment opportunities outside the program[53].

6.1 Quantile Treatment Effects

In practice, detecting heterogeneity of treatment effects is complicated. Indeed, some parameters of the treatment effects distribution, such as its variance, are not identified. Their identification would require knowledge of the joint distribution of potential outcomes whereas only one potential outcome can be observed at a time for each individual.

To study the potential heterogeneity of program impacts in an experimental setting, researchers traditionally look at quantile regressions. Indeed, even if not identified, it is possible to obtain a bound for the variance of impacts (Heckman Smith and Clements, 1997, Djebbari and Smith 2008). Results from quantile regressions provide information about the lower bound of the variance. When quantile treatment effects are homogeneous, the lower bound of the variance is zero: a constant treatment effect is consistent with homogeneous quantile treatment effects. Bitler et al. (2006, 2014) provide a well-known application of quantile treatment effects to analyze heterogeneity of impacts. As highlighted in this prior work, quantile treatment effects

[52] Recall that the only criteria enforced at enrollment are age (18 to 30) and not being a previous beneficiary of the first wave of public works.

[53] These marginal applicants could still benefit from the program in dimensions others than earnings.

will help to detect heterogeneity of impacts only if the intervention preserves ranks, or at least does not lead to too much churning in the distribution of outcomes. In fact, the rank preservation assumption allows one to interpret quantile treatment effects as the "effect *at* quantile", hence making it directly informative about heterogeneity of impacts. Under this assumption, observing non homogeneous quantile treatment effects indicates that the lower bound of the variance is strictly greater than zero. However, if the intervention does not preserve ranks, quantile treatment effects are of little help to detect heterogeneity.

Figure 1 presents quantile treatment effects on earnings during the program (Panel (a)) and after the program (Panel (b)). The horizontal axis in each panel reports the quantile and the vertical axis the estimate of the treatment effect at the corresponding quantile. The shaded area around the estimate provides the 95% confidence interval. As the figure clearly illustrates, the quantile analysis shows substantial heterogeneity of impacts on earnings during the program. The quantile treatment effect is as large as FCFA 45,000 at the 15% quantile, but only FCFA 15,000 at the 85% quantile. Moreover, the estimated quantile treatment effects are quite precise, strongly suggesting the existence of true heterogeneity rather than just sampling variation.

The quantile analysis of treatment effects on income after the program (Panel (b)) offers a rather different picture. Post-program quantile treatment effects are uniformly small and the small dispersion is within confidence bounds, consistent with sampling variation.

In summary, there appears to be large heterogeneity of quantile treatment effects during the program, but less heterogeneity after the program. The model derived in the framework (section 2) shows that the intervention possibly preserves ranks, or at least might not induce too much churning in the distribution of contemporaneous effects. Thus, the true variance of impacts might not be too far from the lower bound and the heterogeneity seen in quantile treatment effects points to true underlying heterogeneity during the program.

However, this is not necessarily the case for post-program earnings. There might be individual latent factors, such as return to capital, that would not contribute to the ranking of individual earnings in absence of the program but contribute to the ranking of individuals' post-program earnings. For example, if there are individuals trapped at the bottom of the earnings distribution without the program, but with high returns to capital (e.g. through setting-up a highly profitable activity), these individuals, thanks to their participation in the program, might end up further towards the top of the earnings distribution. This is because the program allows them to save and implement their latent project after program completion.

6.2. Machine learning to study heterogeneity in impacts

Given the possible limitations of the quantile regression approach, in particular when it comes to assessing heterogeneity in post-program effects, we turn to a different empirical approach. The technique is based on the identification of underlying baseline variables contributing to heterogeneity in the treatment effect. In a standard regression framework, this would be done by interacting the treatment variable with covariates, before recovering predicted impacts conditional on these covariates. Djebarri and Smith (2008) apply this method and show that it captures a substantial share of the variability of impacts. The machine learning methods developed by Athey and Imbens (2016) and Wager and Athey (2016) go one step further in identifying the expectation of impacts conditional on a set of covariates. The main innovation of these machine learning methods is that they can uncover the underlying heterogeneity in impacts without making any assumptions about the source or form of this heterogeneity[54]. Although these methods can miss some determinants of heterogeneity (especially if the set of covariates is not rich enough) [55], they offer an alternative and systematic way to explore heterogeneity of impacts. These methods are particularly relevant to our context given the large number of covariates we have captured in our rich baseline survey. Moreover, these methods permit to detect heterogeneity even if the intervention is not rank preserving.

We apply the causal forest algorithm developed by Wager and Athey (2016) to study the existence and magnitude of heterogeneity in treatment effects. As already discussed above, these machine learning methods are implemented in a two-step procedure. First, the relevant set of covariates is identified on a “training sample” comprised of 50% of the total sample (random subsample). Second, impacts conditional on these identified covariates (and interactions of covariates) are estimated on a “test sample,” the remaining 50% random subsample of the total sample. After having generated several causal forests, we can recover a

[54] Specifically, (i) it doesn’t require assuming linearity and looks for more complex relationships across covariates; (ii) it can search for heterogeneity across high-dimensional sets of covariates rather than restricting to a few covariates (in standard approach).

[55] Causal forests could fail to detect heterogeneity (when there is true underlying heterogeneity) if heterogeneity depends on unobservable characteristics (missing in the “input” features of the model). They would have hard time finding heterogeneity if underlying heterogeneity is highly linear or if some features strongly affect the level of the outcome but not the treatment effect (leading to spurious splits in the algorithm). Lastly, measurement errors in covariates could also decrease the efficiency of the algorithm to determine the right splits of the data.

predicted treatment effect conditional on baseline covariates for each unit in the sample. In the following section, we will simply refer to them as “predicted treatment effects”, but these predictions are obtained from test samples only and are *conditional* on the set of baseline covariates used. The full procedure, including the implementation of the causal forest algorithm, is detailed in Appendix B. This approach can be used to estimate the heterogeneity of predicted impacts both *during* and *after* the program; and it can be applied to study heterogeneity in the impacts on earnings, but also on other outcomes, such as well-being.

Assessing the magnitude of heterogeneity in impacts on earnings

Figure 2 shows the distribution of predicted impacts on earnings derived from the causal forest algorithm during the program (darker shade) and after the program (lighter shade). It reveals substantive heterogeneity of impact during the program, consistently with the quantile regression results reported above. The distribution of predicted impacts on earnings during the program is skewed to the left, which implies that more than half of program participants benefit more than the average treatment effect. Conversely, the extended lower tail suggests that some individuals benefit far less from the program. In comparison, the distribution of predicted impacts after program reveals less heterogeneity; both the standard deviation and the interquartile range are smaller than for the distribution during the program.

Table 4 complements Figure 2 by reporting predicted impacts by quartile, both during and after the program. Contemporaneous program effects on total earnings are reported in column (1), and post-program effects in columns (2). Although confidence intervals cannot yet be constructed for the predictions recovered from the causal forests, the average treatment effects per quartile highlight how much impacts vary along the distribution. We focus in particular on the difference between the average impact in the top quartile and in the bottom quartile, as well as with the estimated average treatment effect.

Panel (A) of Table 4 presents the means and standard errors of the distribution of estimated conditional impacts on earnings during and after the program. Summarizing the evidence from Figure 2, we find that the mean predicted impact on earnings during the program is FCFA 20,230, with a standard deviation of 7,614. The mean predicted impact on earnings after the program is FCFA 4,914, with a standard deviation of 2,713.

Column (1) of Panel (B) offers further confirmation that there is a large amount of heterogeneity in impacts on earnings during the program. The average predicted impact on earnings is FCFA 9,934 in the lower quartile of the distribution compared to FCFA 28,622 in the upper quartile. In contrast, column (2) of Panel C suggests more modest heterogeneity after the program. The average predicted impact on earnings is FCFA 1,475 in the lower quartile of the distribution compared to FCFA 8,329 in the upper quartile.

In other words, finer targeting could in theory strongly improve program impacts. For the bottom 25% of applicants, the contemporaneous impacts on earnings are less than half of those of the mean applicant, and represent only approximately 20% of the transfer made by the program. Replacing randomized assignment of the program by a targeting rule that "maximizes" contemporaneous program impacts on earnings would lead to a 37% increase in impacts during the program (FCFA 28,622 compared to FCFA 20,885 achieved under random selection). Replacing randomized assignment of the program with a targeting rule that "maximizes" post-program earnings would roughly double post-program impacts (FCFA 8,329 compared to FCFA 5,621 achieved under random selection). This last result is an important one, as we now further discuss.

It does not come as a surprise that workfare programs have large and heterogeneous contemporaneous impacts. This is actually what we expect from the self-selection mechanism, with a fraction of marginal participants almost indifferent between being enrolled or not (in terms of earnings), and others being infra-marginal. One major question with workfare programs is whether they have impacts after program completion. The potential mechanisms to explain heterogeneity in post-program impacts are less straightforward than for the contemporaneous impacts. They might have to do with experience in the labor market and ability to signal skills. They could be related to return to capital and the impact of investments on income generating activities, through accumulated savings during the program.

While the results in Table 4 so far suggest the possibility of improving program effectiveness both during and after the program through better targeting, an important question is whether the targeting that would maximize contemporaneous program impacts would also maximize post-program impacts. In other words, are the covariates that are associated with large predicted impacts during the program associated with large predicted impacts after the program as well (and vice versa)? The scatter plot in Figure 3 suggests a negative answer. Figure 3 shows predicted impacts on earnings during the program (x-axis) against predicted impacts on earnings after the program (y-axis), all derived from applying the causal forest algorithm. The solid vertical and horizontal lines on the scatter plot respectively correspond to the average predicted impacts during and after the program. Similarly, the horizontal (respectively vertical) dashed lines represent the 1st and 3rd quartile of the distribution of predicted impacts during (respectively post) program. A high correlation between impacts on earnings during and after program would lead to a concentration of predictions along the diagonal from the top right corner – those who have the largest impacts during and after the program, to the bottom left corner. On the contrary, the scatter plot shows that even within the top quartile of impacts during the program, the post program impacts are dispersed on the opposite axis (and similarly for the top quartile of post program impacts). The remaining columns in Panel (B) and (C) of Table 4 illustrate further.

Column (2) of Panel (B) and column (1) of Panel (C) confirm the visual evidence in Figure 3 in that there is no systematic relationship between those who benefit the most during the program and those who benefit the most after the program. Specifically, targeting the program to the 25% of applicants that benefit the most during the program would result in average predicted impacts after the program of FCFA 5,119, marginally below the average treatment effect. Similarly, targeting the program to the 25% of applicants that benefit the most after program completion would result in average predicted impacts during the program of FCFA 20,706, close to the average treatment effect. In other words, these results suggest tradeoffs when trying to improve program effectiveness through better targeting. They highlight the challenge in isolating a single targeting rule that can "maximize" impacts both during and after the program. This relates to a broader trade-off between the two objectives of the public works program, namely its safety net role "during" the program and its productive dimension related to post-program employment and earnings.

Analyzing patterns of heterogeneity

Given the important heterogeneity of impacts on earnings during the program, we want to learn more about the differences between individuals in the bottom and top quartile of the distribution of predicted impacts. On the one hand, one could argue that individuals in the bottom quartile should be explicitly excluded from the program. On the other hand, these individuals decided to participate in the program despite small earning gains so they might benefit from the program in other dimensions.

It is not straightforward how one can use causal forest predictions to summarize the differences in characteristics associated with high versus low treatment effects. The function $\widehat{\tau(x)}$ is derived from non-parametric procedures, which makes it difficult to display differences in the characteristics that matter most. Moreover, interpreting the structure of the covariate splits, say for a given tree of the causal forest, could be misleading[56] (Mullainathan and Spiess, 2017). By looking at the summary statistics for quartiles of the distribution of treatment effects, Table 5 offers a way to grasp some of the most important differences in terms of baseline covariates between the two groups. The table investigates the differences between the groups with lowest (column 2) and highest (column 3) estimated conditional impacts on earnings during the program by presenting the average of several baseline characteristics over the two groups as well as for the sample of participants (column (1)).We present in column (4) the p-value for the test of the hypothesis of no difference across these groups.

The table clearly shows differences in profiles between individuals who benefit the least and the most during the program. The share of women is significantly higher among the upper quartile at midline (43%) than among the bottom quartile (22%). A large set of characteristics related to the financial status of the individuals suggest that the lower quartile was "better-off" at baseline, whereas the upper quartile was poorer. This holds for assets, expenditures, savings and earnings. There is a large difference in the stock of savings among individuals in the bottom quartile of the distribution of predicted impacts during the program (FCFA 48,115) and those in the top of the distribution (FCFA 17,535). Similarly, the share of participants in the bottom quartile saying they face credit constraints (44%) is much lower than for applicants in the top quartile (59%). Finally, the share of people working is substantially higher among the bottom quartile (90%), compared to the top quartile (66%). Baseline earnings are four times higher for the group in the bottom quartile compared to the top quartile. Twice as many individuals in the bottom quartile are engaged in independent activities at baseline compared to the upper quartile (respectively 49% and 23%).

Those who benefit the least during the program appear to be less financially constrained and have a higher stock of savings. While they have larger foregone earnings than other participants,

[56] Mullainathan and Spiess (2017) consider that the underlying structure (covariate splits) is part of what "we do not learn" from machine learning. Although machine learning algorithms are stable in terms of prediction quality, they are not in terms of model selection. It means that across trees within a forest, one can get similar predictions based on very different partitions determined by covariates on which splits are performed. One reason for that is the positive correlation across variables and with treatment effects, for example. It makes the interpretation of the covariate splits highly challenging.

they may be able to save a greater share of their income during the program, or they may be able to better invest these savings into income-generating activities[57]. Table 4 and Figure 3 suggested earlier that those who benefit the least during the program do not necessarily benefit more than others on average after program completion. The observed differences in baseline characteristics suggest it is worthwhile exploring more how treatment effects on other outcomes differ between the two groups, especially outcomes related to self-employment and productive investments. By doing this, we highlight how machine learning techniques can help tease out mechanisms explaining program impacts[58].

Tables 6 and 7 present estimates of treatment effects on a range of outcomes, first for individuals in the upper quartile of (predicted) impacts on earnings during the program (designated as "top 25%" group in Panel A), second for individuals in the lower quartile (designated as "bottom 25%" group in Panel B). Table 6 presents estimates of treatment effects (ITT) for the main outcomes during the program, while Table 7 contains post-program estimates. In both tables, we report the coefficient and standard errors of the interaction between treatment and dummies for the relevant quartile[59], which gives us the total treatment effect (ITT) for the group of interest. We also report the p-value of the test assessing if the estimated treatment effect for the relevant quartile is statistically different from the estimates for the rest of the sample.

Columns (1) to (6) in Table 6 show that the top 25% group have the highest impact on wage employment and related earnings. This is consistent with results in Table 5, which suggested they were more vulnerable people less likely to have an activity prior to the program. Panel (B) also confirms that the self-targeting mechanisms based on the formal minimum wage failed in the sense that it brought to the program a subgroup –the bottom 25%- for which impacts on total earnings are close to zero and significantly lower than for the other applicants (Column (5))[60]. Column (7) reveals interesting differences in terms of savings. Panel B suggests that the

[57] It is also possible that these individuals may particularly benefit from the opportunity to save through the bank accounts set-up by the program.

[58] We follow here Davis and Heller (2017b), who explicitly use machine learning to test underlying mechanisms relying on differential treatment response from disadvantaged youth who benefitted from summer jobs. Indeed, if some subgroups are strongly affected in one dimension but negatively on another one, it will hard to identify that while looking at average treatment effects. Compared to them, with look at mechanisms across time (during the program versus post program) rather than across outcomes measured at the same moment.

[59] Quartile dummies are based on the predicted treatment effect obtained for each individual using causal forest algorithms. Appendix B details how assignment to quartile groups is made when using simulations of several causal forests.

[60] Note that for this table, column (5) is the outcome which was used in the causal forest algorithm to predict treatment effects on earnings and therefore used to define the top 25% and bottom 25% groups. This can be interpreted as an indirect test for detecting heterogeneity using causal forests, as implemented in Davis and Heller (2017b). However, note that it is limited by the fact that (i) it tests for a linear interaction whereas the

savings stock of the bottom 25% group increases significantly more than for the other applicants, while total earnings only marginally increase, which represents an increase in the marginal propensity to save. For this group, the impact on savings is 25% higher than the average impact. Individuals in the upper quartile experience smaller increases in savings than average, although the difference between this group and the rest of the sample is not statistically significant. Finally, positive impacts on well-being are observed consistently across the top and bottom quartiles, without statistical differences with the rest of the sample. This suggests that gains in this other dimension are shared broadly irrespective of the magnitude of impacts on earnings.

Overall, Table 6 confirms that, in the short run, the program benefits a more vulnerable group by offering a wage employment opportunity and raising earnings and attract less vulnerable people for which impacts are much more limited. At the same time, impacts on savings open the possibility that these better-off individuals might have opportunities to make productive investments and improve their employment prospects after the program.

Table 7 investigates post-program impacts on a range of outcomes, including employment, earnings, well-being and productive investments. Table 7 displays how post-program impacts vary in the top quartile of the predicted short-term impacts compared to the rest of the population (Panel A), and similarly for the bottom quartile (Panel B). Overall, little heterogeneity in employment and earnings is observed. One exception is that individuals in the bottom quartile – who had higher baseline level of savings, and saved a higher fraction of their income during the program – are marginally more likely to be self-employed in the long run (+ 7 pp, column (3)). Results also suggest that they made larger investments in independent activities either by starting new activities requiring a higher amount of capital to start (column (9)), or by investing in productive assets (column (10)). Although the differences are not statistically significant, partly due to large standard errors, the point estimates suggest that the program nearly doubled the value of productive assets (in independent business) for this specific group, and in comparison other applicants invested twice less. One caveat is that some of the differences between this group and the rest of the sample are statistically weak. The last column indicates that a substantial share of these independent activities are "new" businesses

forest searches for more complex relationships, (ii) quartile indicators are defined on an estimator, which is not taken into account in the standard errors displayed. The results confirm the visual we had in Figure 2: strong heterogeneity is detected for the bottom of the distribution, with smaller heterogeneity at the top.

launched after the end of the program, so that impacts on self-employment are not only driven by pre-existing activities.

Overall, results on post-program impacts reveal limited heterogeneity. Vulnerable individuals who saw the largest gains in earnings and employment during the program do not enjoy relatively larger post-program gains. Similarly, better-off individuals who saw smaller gains in earnings during the program but nevertheless managed to mobilize substantial savings do not see post-program impacts on earnings, despite signs that they are more likely to be self-employed in slightly expanded activities. Little heterogeneity on impacts on well-being are observed either.

6.3. Alternative Targeting Rules

The analysis in this section so far suggests the possibility of targeting the program in alternative ways that would improve contemporaneous program impacts or post-program impacts on income, but also highlights some trade-offs between these two objectives. It is also clear that the targeting rules that would emerge from machine learning approaches would be too complicated and expensive to implement, relying on information that is not easily observable or verifiable and involving complex functions based on this information, and hence of limited practical relevance. An important follow-up policy question is therefore whether there are simple targeting rules, or simple changes to the self-targeting procedure, that could be implemented and come close to achieving the predicted impacts in the upper quartiles of the distribution in Table 4 column (1) of Panel (B) (contemporaneous program impacts) and column (2) of Panel C (post-program impacts). We address this question in Table 8. In particular, Table 8 presents average contemporaneous program impacts on income (upper panel) and post-program impacts on income (lower panel) for specific sub-populations. For reference, the first column displays averages of predicted impact on income over the whole sample of participants. Similarly, columns (1A) and (1B) report averages of the causal forest predictions of impacts on income over the top quartile and bottom quartile of the distribution.

One of the findings from Table 5 was the greater relative representation of women in the subset of the population that benefit the most during the program. What if only women had been able to apply for the program? Column (3) in Table 8 shows that such a change to the program participation rule would improve impacts both during and after the program. In particular, average impacts on income during (after) the program would be FCFA 37,150 (FCFA 8,258) if the randomized assignment had been restricted to female applicants. This corresponds to a 77% improvement in average estimated impacts on during-program income and a 46% improvement in post-program impacts compared to randomized assignment of both men and women. The point estimates are not statistically different from the machine learning benchmark.

As we discussed above, one of the reasons why self-targeting might have led many non-poor to apply to the program is that the wage was set at the statutory minimum wage, well above hourly rate for many in the informal sector. What if access to the program had been randomized among those, men and women, willing to participate at a lower wage rate? We can, albeit imperfectly, explore this question. Indeed, our endline survey asks participants whether they would have participated in the workfare program for a FCFA 1,500 (daily) salary instead of the offered FCFA 2,500. While it is clearly suboptimal to study heterogeneity of effects based on a variable measured at endline, we note that identical proportions in treatment and control groups agreed to the idea of participating at the lower wage. So, while it is somewhat speculative to run heterogeneity analysis based on such a characteristic, it remains interesting to consider because self-selection is central to workfare programs[61]. Results are displayed in column (2) of the table. We see that here again there is a substantial improvements in both contemporaneous and post-program impacts. Average impacts would have reached FCFA 32,076 at midline and FCFA 8,121 at endline if a random sample of only those willing to work for FCFA 1,500 had been selected. Here again, the point estimates are not statistically different from the machine learning benchmark.

Importantly, we should note though that this is not what the effect of a workfare program paying FCFA 1,500 would have given. Indeed, these are the earnings impacts, during and after the program, among those willing to participate at FCFA 1,500 but actually being paid FCFA 2,500. Hence, while self-selection would have been better (with lower foregone earnings) had the wage paid to program participants been set below the minimum wage, the value of the program for the participants, at this lower wage, would have been reduced. In other words, while lowering the offered wage improves the self-targeting, it also reduces the transfers, and hence program benefits, to this better targeted group. This is one of the key tradeoff embedded in the self-

[61] We implemented some robustness checks by examining results not based on the endline binary variable but on a prediction of this variable, using baseline covariates. The (linear) prediction model is estimated on the control group (expecting that their answer to this question is not affected by their treatment status) and applied to the treatment group. The results are similar.

targeting logic behind workfare programs. If this were a one-shot game, policymakers might have been able to surprise program applicants expecting a FCFA 1,500 wage with a higher actual wage and hence mimic the results above. Clearly though, in a realistic repeated game setting, such a practice would certainly not be sustainable.

While self-targeting may offer the cheapest targeting approach, what about simple targeting rules based on a few predictors of baseline income? If there is limited churning in the distribution of income over a period of time as short as the length of this program (6-7 months), those with the lowest baseline income are likely to have the lowest income at midline absent program participation. In columns (4) and (5) of Table 8, we experiment with two approaches to directly target the poor. First, we assess impacts during and after the program among those that report baseline earnings in the bottom quartile of the distribution (column (4)). Second, we use the machine learning methods outlined above to predict baseline earnings among program applicants using a limited set of covariates that are both easily observable and not easily manipulated, including gender, age, household characteristics and assets. We then assess program impacts among participants in the lowest quartile of this distribution of predicted baseline income. This second approach is meant to mimic the proxy means tests that are often used when targeting safety nets to the poor, and more robust to misreporting than self-reported income. Columns (4) and (5) show that the midline impacts would match predicted impacts in the upper quartile of the predicted impacts distribution at midline under either of these targeting rules. Targeting the 25% with the smallest baseline self-reported income leads to an average expected impact on income during the program of FCFA 32,695; targeting the poor based on the 25% lowest predicted income leads to an average expected impact on income during the program of FCFA 36,822. Targeting based on a proxy test for income leads to estimated post-program impacts that are roughly comparable to restricting program participation to women, and not statistically different from the machine learning benchmark.

The previous analysis has shown that the allocation of program slots can be substantially improved with better self-targeting or targeting rules. The machine learning heterogeneity analysis suggests that optimal targeting would differ based on the outcome of interest (contemporaneous or post-program) and that trade-offs will appear. Yet the consideration of alternative targeting rules suggests no sharp trade-offs. In other words, targeting those that will benefit the most in one dimension does not lead to targeting those that will benefit the least in another dimension. Our results in Table 8, with the exception of the result in column (3), are consistent with this. We can substantially improve impacts during the program through better targeting, at no large costs, in terms of post-program impacts. Furthermore, any of the alternative targeting rules performs as well as the machine learning benchmark.

7. Cost-effectiveness

The estimated direct economic impacts of the program on youths' earnings can be used to perform cost-effectiveness calculations. The total costs per beneficiary for the basic public works program amounts to FCFA 660,478 ($1321), of which FCFA 354,166 ($708) are direct transfer to beneficiaries, FCFA 255,189 ($510) are other direct costs (material, team leaders and supervisors, basic life skills training), and FCFA 51,123 ($102) are indirect management costs[62]. For the public works only, the contemporaneous impacts on earnings are estimated at approximately FCFA 20,900 ($42) and the post-program impacts on earnings at FCFA 4,100 (8.2$) per month.

Table 9 presents the main results from cost-effectiveness calculations. Assuming that the contemporaneous impacts are constant during the 7 months of the program, and that the estimated post-program impacts are constant for the 13 months after the end of the program[63], the discounted total impacts on earnings (over 20 months) is estimated at FCFA 206,695 ($42). The cost-effectiveness ratio for the intervention is 3.2, meaning that the average cost per beneficiary is 3.2 times higher than the average discounted direct impacts on earnings. This poor cost-effectiveness ratio is not surprising given that the net earnings gains are only 42% of the average monthly transfer amounts during the program, and that the cost of transfers only

[62] Cost-effectiveness calculations presented in that section are performed for the public works only (without complementary basic entrepreneurship training or jobs search skills training). The comparisons of relative cost-effectiveness between scenarios remain similar if considering cost and impacts from the pooled treatment instead.

[63] We compute a discounted sum of impacts on total earnings from program starts (month 1) up to 13 months after exit from the program (month 20) which corresponds to the moment when post-program impacts are measured. This is conservative in the sense that we make no assumptions for what happen after month 20 (or equivalent to zero-impact assumption). The following assumptions are used for the calculations:
H1 : contemporaneous program impacts β^{During} correspond to the impact on earnings at the end of month 4 and this impact is constant (up to a monthly discount factor) over the program period (month 1 to month 7)
H2 : post-program impacts β^{Post} correspond to the impact on earnings at the end of month 13 and this impact is constant (up to a monthly discount factor) from the end of the program (month 8) to the endline survey (month 20).
H3 : the annual discount rate is equal to $1/\ (1+\delta)$, with $\delta = 10\%$. Using monthly discount factors, $\rho = 1/(1+\delta)^{1/12}$.
Finally, the total discounted impact flow (DIF) is : $DIF = \sum_{k=1}^{7}(\rho^{k-4} * \beta^{During}) + \sum_{k=1}^{13}(\rho^{k-13} * \beta^{Post})$, with β^{During} (respectively β^{Post}) the contemporaneous (respectively post-program) ITT estimates of monthly total earnings impact.

represents 54% of overall program costs. To highlight the poor average cost-effectiveness of the program in another way, the post-program impacts observed after 13 months would need to be sustained for 22.9 years for the program to be cost-effective based on net earnings gains for youths.

This cost-effectiveness analysis should be considered a lower-bound as it does not account for non-economic benefits such as those on psychological well-being or behavior mentioned above, or other externalities from the program, including the indirect benefits of roads rehabilitation. Nevertheless, they provide a benchmark to assess the cost-effectiveness of potential program improvements such as the implementation of alternative targeting mechanisms, in particular if we consider in a first-order approximation that non-economic benefits and externalities are similar across these potential improvements.

Table 9 indicates how adjustments in program targeting would strongly improve cost-effectiveness. In light of the strong observed heterogeneity in impacts (at least during the program), cost-effectiveness improves strongly across the various targeting approaches considered. Compared to the benchmark scenario with self-targeting based on the formal minimum wage, the total discounted total impact on earnings more than double (from FCFA 206,695 to FCFA 418,961, or $413 to $838) under the scenario of selecting youths with low predicted baseline earnings. Large improvements are also observed under other alternative selection criteria, including those that proxy self-selection based on a lower wage, target women only, or target based on self-declared baseline earnings. Overall, the cost-effectiveness ratio would improve from 3.2 to between 1.6 and 2 based on finer program targeting. The years needed to sustain post-program impacts for impacts on youths' earnings to justify investments in the program would go down from 22.9 to between 3 and 5.5. While the analysis cannot decisively indicate which targeting scenario would maximize cost-effectiveness given the confidence intervals around the impact estimates, it does highlight strong improvements in cost-effectiveness when departing from self-targeting based on the formal minimum wage.

8. Conclusion

The Côte d'Ivoire public works program we have evaluated shares many of the features of other public works program that have been adopted throughout the developing world in response to transient negative shocks such as those induced by climatic shocks or episodes of violent conflicts. It provided a few months of employment in road rehabilitation to those willing to work at the stated wage. Our analysis, relying on a randomized control trial as well as the collection of rich data before, during and after the program, has allowed us to assess the effectiveness of these programs in lifting participating youth, both economically and psychologically.

Results show that program impacts on employment are limited to contemporaneous shifts in the composition of employment towards the public works wage jobs, with no lasting post-program impacts on the level or composition of employment. However, participation in the program does raise earnings and psychological well-being, both during the program but also, and maybe most importantly, after program completion. Post-program earnings gains are mainly achieved through more vibrant small-scale entrepreneurial activities, likely boosted by the additional savings participants were able to secure during the program, but also possibly by other skills that were developed through the workfare and related complementary training.

However, the program as currently implemented is far from cost-effective when benefits are measured solely based on the earnings gains of those participating. This is primarily due to the fairly high indirect cost of implementing public works programs compared to more traditional welfare programs, but also due to the use of self-selection mechanisms based on the formal minimum wage. Many of the individuals who apply to participate in the program are already employed, consistent with the widespread underemployment challenge and limited unemployment in Sub-Saharan Africa. In an environment where informal employment is rampant, many of those who already have some form of occupation self-select into a public works program that offer higher earnings (even by paying the formal minimum wage), as well as potential non-economic benefits on psychological well-being and behavior. In this context, the program has very small average impacts on employment or hours worked, leading to large foregone earnings.

A basic framework to consider self-selection mechanisms shows that a reasonable theoretical expectation is that the impact on earnings is almost zero for the ‘marginal’ participant, and equal to the amount of the transfer for those with no outside employment opportunities. In this context, the distribution of individual impacts over the population is likely to vary substantially. Consistent with this, we demonstrate, using new methods from machine learning, large heterogeneity in program impacts during the program, with more modest heterogeneity after the program. In fact, heterogeneity in program impacts on earnings during the program are so large that they suggest that improvements in targeting is a first-order program design question, and perhaps a more critical issue than other program design aspects such as those related to program content like the value-added of complementary skills training.

Results from machine learning techniques suggest potential trade-offs between maximizing contemporaneous and post-program benefits. Yet traditional heterogeneity analysis shows that a range of practical targeting mechanisms perform as well as the machine learning benchmark, leading to stronger contemporaneous and post-program benefits without sharp trade-offs. This implies that cost-effectiveness could be boosted by departing from self-targeting based on the formal minimum wage. Indeed, we show that a range of potential self-targeting or targeting rules, which could be implemented at minimum additional costs, could substantially raise cost-effectiveness as measured solely based on earnings' gain. Still, even with this improved targeting, post-program impacts would still need to be sustained for at least 3 years for the program to be cost-effective based on participants' earnings gains alone.

Does it mean that public works program are not worth it? While our results so far could be interpreted that way, this is under the important caveat that the cost-effectiveness ratios we currently estimate are based on participants' earning alone, with zero impacts on earnings assumed beyond 13 months after the program. The cost-effectiveness calculations are also conservative as they do not include other social benefits of the public works program, such as the value of the new or upgraded infrastructure or the reduction of negative externalities (for example, incapacitation effects leading to reduction in crime or illegal activities) the program may have generated. These additional benefits are viewed as one of the advantage of workfare programs compared to traditional welfare, although they are rarely formally evaluated. Accounting for such externalities, a public works program with improved targeting may well become cost-effective. It may be particularly socially or politically desirable if social planners put a high weight on externalities and non-economic social benefits related to social cohesion relative to direct economic benefits. In future work, we will attempt to provide a back-of-the envelope calculation as to how much these additional benefits might need to be to justify public works from a cost-benefit perspective.

7. References

Alik-Lagrange, A., and Ravallion, M. (2015). Inconsistent Policy Evaluation: a Case Study for a Large Workfare Program. *NBER Working Paper, 21041.*

Almeida, R., and Galasso E. (2010). Jump-starting Self-employment? Evidence for Welfare Participants in Argentina. *World Development* 38(5): 742-755.

Amaral, S., Bandyopadhyay, S. and Sensarma R. (2015). Public Work Programs and Gender-based Violence: The Case of of NREGA in India. Department of Economics, University of Birmingham Discussion Papers.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27), pp.7353–60.

Beegle, K., Galasso, E., and Goldberg, J. (2015). Direct and Indirect Effects of Malawi's Public Works Program on Food Security. *World Bank Policy Research Working Paper Series* No. 7505.

Bertrand, M., Crépon, B., Marguerie, A. and Premand P. (2016). Impacts à Court et Moyen Terme sur les Jeunes des Travaux à Haute Intensité de Main d'œuvre (THIMO) : Résultats de l'évaluation d'impact de la composante THIMO du Projet Emploi Jeunes et Développement des compétences (PEJEDEC) en Côte d'Ivoire. Washington DC : Banque Mondiale et Abidjan : BCP-Emploi.

Besley, T., and Coate, S. (1992). Workfare versus Welfare Incentive Arguments for Work Requirements in Poverty-Alleviation Programs. *American Economic Review*, *82*(1), 249–261.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2014). Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. *NBER Working Paper,* No *20142*.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review*, *96*(4), 988–1012.

Blattman, C., and Ralston, L. (2015). Generating Employment in Poor and Fragile States: Evidence from Labor Market and Entrepreneurship Programs. *Mimeo*

Breiman L. (2001). Random forests, *Machine Learning*, 45, 5-32.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S. (2016). Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*, 106(5), pp.124–127.

Christiaensen, L. and Premand, P.. (2017). Côte d'Ivoire Jobs Diagnostic: Employment, Productivity, and Inclusion for Poverty Reduction. World Bank, Washington, DC.

Datt, G., and Ravallion, M. (1994). Transfer benefits from public-works employment: Evidence for rural India. *The Economic Journal*, *104*(427), 1346–1369.

Davis, B.J.M.V. and Heller, S.B. (2017a). Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs. *American Economic Review: Papers & Proceedings*, 107(5), pp.546–550.

Davis, J.M.V. and Heller, S.B. (2017b). Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs.

Deininger, K., and Liu, Y. (2013). Welfare and Poverty Impacts of India’s National Rural Employment Guarantee Scheme Evidence from Andhra Pradesh. *Policy Research Working Paper Series*, No. 6543.

Deininger, K., Nagarajan, H.K. & Singh, S.K., 2016. Short-Term Effects of India’s Employment Guarantee Program on Labor Markets and Agricultural Productivity. *Policy Research Working Paper Series*, No.7665.

Djebbari, H., and Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, *145*(1–2), 64–80.

Fetzer, T. (2014). Social Insurance and Conflict: Evidence from India. EOPP Working Paper Number 53.

Filmer, D. and Fox, L. (2014). Youth Employment in Sub-Saharan Africa. *Africa Development Series.* Washington, DC: World Bank.

Galasso, E. and Ravallion M. (2004). Social Protection in a Crisis: Argentina's Plan Jefes y Jefas. *World Bank Economic Review* 18(3): 367-399.

Galasso, E., Ravallion, M and Salvia A. (2004). Assisting the transition from workfare to work: A randomized experiment. *Industrial and Labor Relations Review* 58(1): 128-142.

Gilligan, D. O., Hoddinott, J., and Taffesse, A. S. (2009). The Impact of Ethiopia’s Productive Safety Net Programme and its Linkages. *Journal of Development Studies*, *45*(10), 1684–1706.

Hastie T., Tibshirani R., and Friedman J. (2011), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The Review of Economic Studies*, *64*(4), 487–535.

Kleinberg, J., Ludwig J., Mullainathan, S. and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings,* 105(5), pp.491–495.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *NBER Working Paper*.

Imbert. C. and Papp, J. (2015). "Labor Market Effects of Social Programs: Evidence from India's Employment Guarantee." *American Economic Journal: Applied Economics* 7(2): 233-263.

INS and AGEPE (2014). Enquête nationale sur la situation de l'emploi et du travail des enfants (ENSETE 2013).

Islam, M. and Sivasankaran A. (2015). How does Child Labor respond to changes in Adult Work Opportunities? Evidence from NREGA. Mimeo

Jalan, J. & Ravallion, M., 2003. Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. *Journal of Business & Economic Statistics*, 21(1), pp.19–30.

Li, T. and S. Sekhri (2013). The Unintended Consequences of Employment Based on Safety Net Programs. Mimeo, University of Virginia.

McBride, L. and Nichols, A. (2016). Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. *The World Bank Economic Review*.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives—Volume*, 31(2—Spring), pp.87–106.

Ravallion, M., Datt. G and Chaudhuri S. (1993). "Does Maharashtra's Employment Guarantee Scheme Guarantee Employment? Effects of the 1988 Wage Increase*." Economic Development and Cultural Change* 41(2): 251-275.

Ravallion, M., Galasso, E, Lazo, T. and Philipp E. (2005). "What Can Ex-Participants Reveal about a Program's Impact?" *Journal of Human Resources* XL(1): 208-230.

Ravi, S., and Engler, M. (2015). Workfare as an Effective Way to Fight Poverty: The Case of India's NREGS. *World Development*, *67,* 57–71.

Rosas, N., and Sabarwal, S. (2016). Public Works as a Productive Safety Net in a Post-Conflict Setting : Evidence from a Randomized Evaluation in Sierra Leone (No. 7580). *Policy Research Working Paper Series*.

Shah, M. and B. M. Steinberg (2015). "Workfare and Human Capital Investment: Evidence from India." National Bureau of Economic Research Working Paper Series No. 21543.

Subbarao, K.; Del Ninno, C ; Andrews C. and C. Rodriguez-Alas, (2013). Public Works as a Safety Net: Design, Evidence, and Implementation. Washington, DC: World Bank.

Wager, S., and Athey, S. (2016). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Working Paper*, 1–43.

Zimmermann, L., 2015. Why Guarantee Employment? Evidence from a Large Indian Public-Works Program. *Working Paper*, (April).

Table 1: Sample compared to National Population

	Public Works (THIMO) Midline survey data Control group	National Labor Survey 2013 Youth 18-30 Urban areas
Employment status (primary occupation)		
Inactive	7,8%	34,9%
Active	92,2%	65,1%
Unemployed	6,0%	16,0%
Wage employment (including informal)	50,6%	25,1%
Self employment (non agricultural)	26,8%	39,2%
Self employment in agriculture	2,2%	4,7%
Others	14,0%	15,0%
Education (diploma)		
No diploma	47,5%	47,1%
CEPE (completed primary school)	22,8%	21,7%
BEPC (completed middle school)	16,8%	18,5%
BAC or more (completed secondary school)	12,1%	12,7%

Table 2: Balance checks and Summary statistics

	(1)	(2)	(3)	(4)	(5)
	Treatment	Control	Test	Test across	Nb
	group	group	(1)-(2)	4 Arms	Obs.
	(pooled)		(p-value)	(p-value)	
Individual characteristics					
Live in urban area	94,2%	92,7%	0,14	0,22	4 099
Age	24,61	24,56	0,58	0,16	4 099
Nationality : Ivorian	95,5%	96,9%	0,04	0,08	4 099
Nb of children	0,80	0,82	0,70	0,90	4 099
Education					
Did not complete primary school (No diploma)	46,9%	48,1%	0,56	0,88	4 098
Has completed primary school (CEPE)	24,8%	22,8%	0,19	0,61	4 098
Has completed middle school (BEPC)	18,0%	16,6%	0,31	0,30	4 098
Has completed secondary school (BAC or +)	10,1%	12,3%	0,06	0,28	4 098
Is a student	4,9%	7,6%	0,00	0,04	4 099
Previous Vocational Training	38,8%	41,2%	0,16	0,11	4 096
Including : traditional / informal apprenticeship	73,2%	71,2%	0,52	0,67	1 613
Household					
Household size (total number of members)	6,03	6,05	0,92	0,52	4 097
Number of rooms	3,16	3,17	0,87	0,66	4 099
Nb of children (<18 ans)	2,00	1,96	0,51	0,77	4 099
Is head of household	24,8%	23,5%	0,46	0,85	4 099
Share of members working (last 7 days)	55,5%	55,9%	0,62	0,86	4 097
Household Assets (total number) (last 3 months)					
Total	13,86	13,85	1,00	0,18	4 099
Transport	0,73	0,79	0,26	0,11	4 099
Agriculture	4,71	4,64	0,91	0,25	4 099
Household Equipment	1,64	1,65	0,98	0,65	4 099
Communication	6,77	6,78	1,00	0,90	4 099
Savings					
Has Saved (last 3 months)	48,6%	49,8%	0,57	0,79	4 099
Share of formal savings (among those who saved)	25,5%	25,6%	0,57	0,94	1 988
Has a Savings Account	11,2%	10,4%	0,47	0,83	4 099
Savings Stock (FCFA)	28 777,05	26 843,99	0,37	0,83	4 042
Face Constraints to repay loans	19,9%	23,1%	0,03	0,22	4 099
Face Constraints to access credit	49,9%	49,8%	0,96	0,66	4 099
Constraints and expenditures					
Nb of days with no meals (last 7 days)	0,83	0,79	0,46	0,80	4 099
Highly constrained for basic needs expenditures	70,3%	73,8%	0,04	0,15	4 099
Transportation expenditure (last 7 days)	1 924,28	1 772,95	0,15	0,59	4 095
Communication expenditure (last 7 days)	1 731,15	1 623,06	0,36	0,88	4 092
Employment					
Has an activity	79,1%	80,5%	0,28	0,37	4 099
Searched for a job (last 6 months)	76,9%	78,9%	0,17	0,20	4 099
Risk and Time preferences					
Risk aversion level (scale 0 to 10, 0=very averse)	4,69	4,74	0,66	0,43	4 099
Is Risk averse (based on lotteries)	74,0%	71,6%	0,18	0,19	4 099
Patience level (scale 0 to 10, 10=very patient)	3,33	3,42	0,39	0,81	4 095
Preference for present (actualization rate for 1 month)	0,57	0,57	0,95	0,46	4 099
Cognitive Skills (% success in answers or tasks at each test)					
Raven test (deduction)	23,4%	23,4%	0,92	0,09	4 093
NV7 test (spatial vision)	27,0%	26,4%	0,24	0,11	4 099
Dexterity (sorting nuts test)	38,0%	37,4%	0,02	0,03	4 094
Dexterity (nuts and bolts test)	33,4%	33,7%	0,21	0,26	4 083

Table 3: Estimated impacts *during* and *post* program

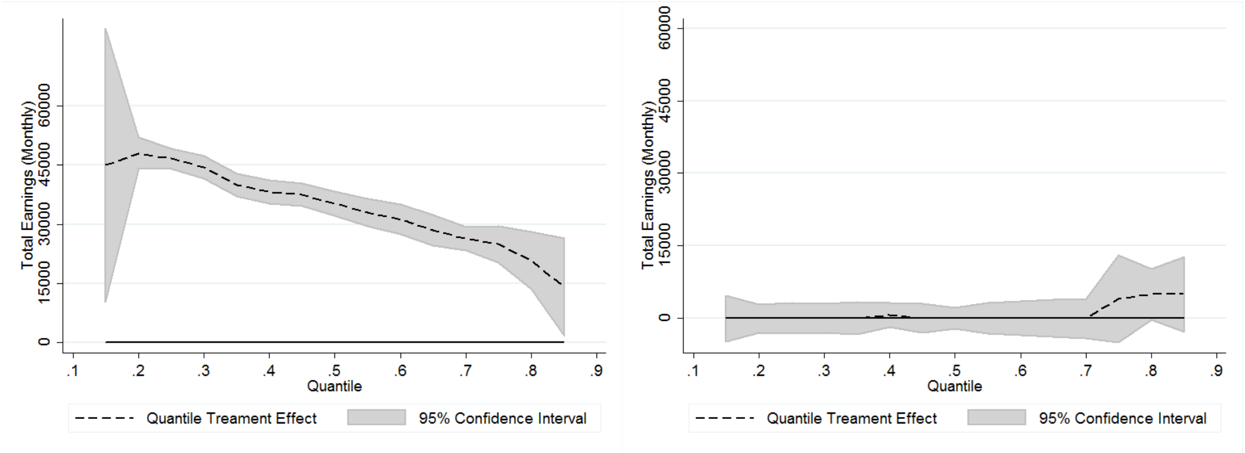
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Has an	Wage Employed	Self Employed	Total Hours	Hours worked	Hours worked	Total Earnings	Earnings in	Earnings in	Total	Savings	Well Being	Behavior
	Activity	(in at least	(in at least	worked	in Wage Empl.	in Self Empl.		in Wage Empl.	in Self Empl.	Expenditures	(stock)	index	index
		1 activity)	1 activity)	(weekly)	(weekly)	(weekly)	(monthly)	(monthly)	(monthly)	(monthly)		(z-score)	(z-score)
Panel A : Impacts *during* the program (∼ 4,5 months after program starts)													
Public Works Treatment (ITT)	0.12***	0.44***	-0.09***	3.49***	14.04***	-6.71***	20885.31***	35385.33***	-12624.85***	15085.18***	39633.27***	0.20***	0.13***
	(0.01)	(0.02)	(0.02)	(1.19)	(1.21)	(0.89)	(6194.08)	(3699.69)	(4633.09)	(1552.68)	(3086.16)	(0.04)	(0.04)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control	0.86	0.53	0.33	40.93	22.90	12.19	60051.55	30916.20	25713.13	48043.04	21751.72	0.00	-0.00
Observations	2958	2958	2958	2958	2958	2958	2912	2912	2912	2945	2958	2934	2946
Panel B : *Post* program impacts (pooled treatment) (12 to 15 months after program ends)													
Public Works Treatment (ITT)	0.01	0.01	0.01	1.21	-0.27	1.36	5621.62**	-972.88	6223.36***	1916.44	10833.24**	0.09**	0.00
	(0.01)	(0.02)	(0.02)	(1.29)	(1.18)	(1.14)	(2422.04)	(1347.95)	(2125.11)	(1503.10)	(4511.48)	(0.04)	(0.04)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control	0.87	0.55	0.33	42.27	24.13	13.23	48463.49	25352.70	19718.45	52227.70	54437.32	0.00	-0.00
Observations	3934	3934	3934	3934	3934	3934	3934	3934	3934	3814	3934	3932	3933
Panel C : *Post* program impacts (by treatment arms) (12 to 15 months after program ends)													
Public Works Treatment (ITT) (PW)	0.01	0.01	0.00	-0.45	-0.33	-0.28	4100.49	-244.47	3736.88*	1810.96	11097.87**	0.12***	0.05
	(0.02)	(0.02)	(0.02)	(1.60)	(1.59)	(1.26)	(2731.38)	(1635.02)	(2213.02)	(1668.30)	(5176.12)	(0.05)	(0.05)
Self-Empl. training (SET)	-0.00	-0.02	0.02	3.07	0.85	2.18	3426.76	-1588.54	6525.02**	-545.44	4776.90	-0.01	-0.06
	(0.02)	(0.03)	(0.03)	(1.90)	(1.91)	(1.54)	(3281.93)	(1676.43)	(3240.86)	(1510.02)	(6098.60)	(0.05)	(0.05)
Wage-Empl. training (WET)	-0.00	0.01	0.01	2.11	-0.66	2.90*	1324.74	-686.35	1249.14	865.97	-5561.65	-0.09**	-0.09*
	(0.02)	(0.02)	(0.03)	(1.78)	(1.65)	(1.63)	(3208.97)	(1543.17)	(3151.02)	(1737.47)	(5831.31)	(0.04)	(0.05)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control	0.87	0.55	0.33	42.27	24.13	13.23	48463.49	25352.70	19718.45	52227.70	54437.32	0.00	-0.00
p-value PW+SET=0	0.558	0.703	0.470	0.161	0.746	0.244	0.026	0.248	0.002	0.468	0.009	0.025	0.858
p-value PW+WET=0	0.524	0.507	0.644	0.301	0.506	0.109	0.102	0.583	0.118	0.172	0.347	0.433	0.459
p-value SET=WET	0.982	0.332	0.815	0.631	0.390	0.722	0.598	0.584	0.201	0.435	0.109	0.126	0.579
Observations	3934	3934	3934	3934	3934	3934	3934	3934	3934	3814	3934	3932	3933

Robust standard errors clustered at (large) brigade level

Earnings, Expenditures and Savings are in FCFA and winsorized at 99%. Hours winsorized at 99%.

* $p < .1$, ** $p < .05$, *** $p < .01$

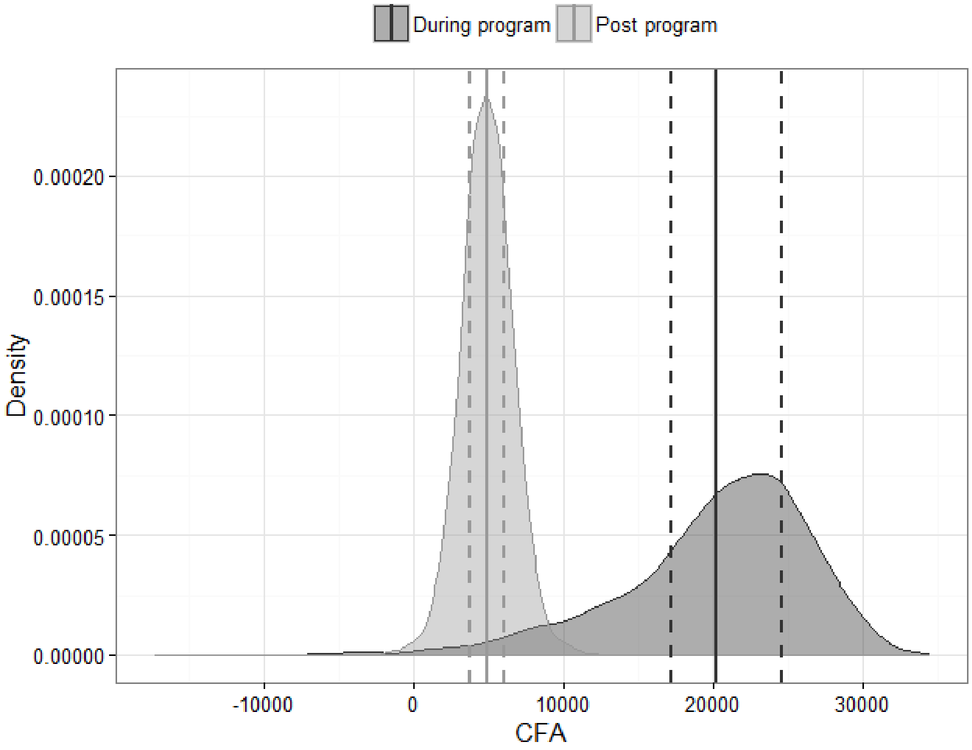
Figure 1: Quantile Analysis of Treatment effects on Earnings



(a) *During* Program

(b) *Post* Program

Figure 2: Distribution of the predicted (conditional) treatment effect on monthly earnings, obtained by a causal forest model



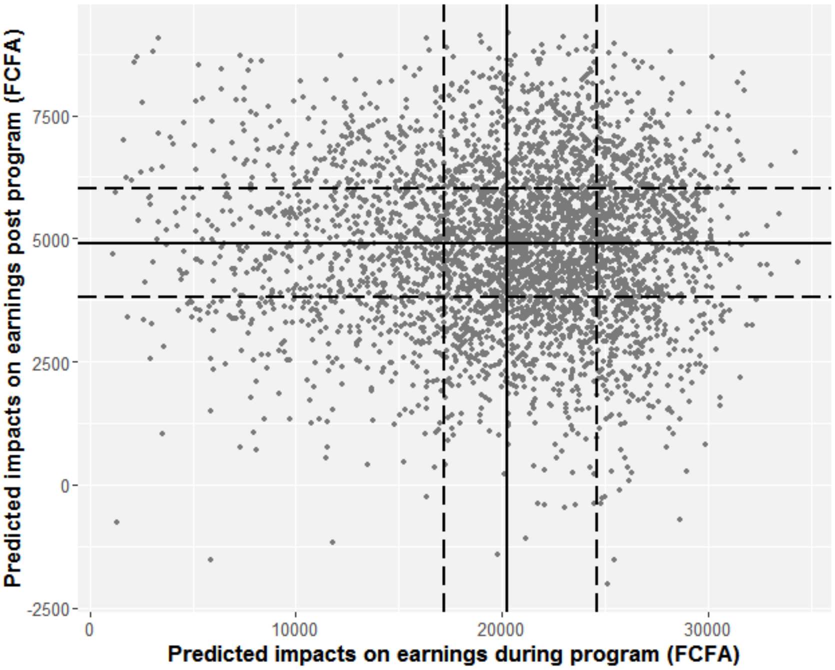
Note : Predicted (conditional) treatment effects obtained by 30 causal forest simulations, predictions made on test sample and averaged across simulations. The solid line represents the mean for each distribution, the dashed lines represents first and third quartile, delimiting bottom 25% and top 25% of the distribution.

Table 4: Heterogeneity in predicted impacts *during* and *post* program

	During program (1) Total Earnings (monthly)	**Post program** (2) Total Earnings (monthly)
Panel A : Predicted impact obtained by causal forests		
Mean (predicted CATE $\widehat{\tau}_i^{30CF}$)	20 230	4 914
Standard deviation (predicted CATE $\widehat{\tau}_i^{30CF}$)	7 614	2 713
ITT estimates (presented for comparison)	*20 885*	*5 621*
Panel B : By quartile of predicted impacts in earnings *during* the program		
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 1 (0 to 25%)	9 934	4 762
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 2 (25 to 50%)	19 033	4 878
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 3 (50 to 75%)	23 527	4 969
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 4 (75 to 100%)	28 622	5 119
Panel C : By quartile of predicted impacts in earnings *post* program		
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 1 (0 to 25%)	19 160	1 475
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 2 (25 to 50%)	20 317	4 130
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 3 (50 to 75%)	20 657	5 807
Mean ($\widehat{\tau}_i^{30CF}$) in quartile 4 (75 to 100%)	20 706	8 329

Note : Causal Forest predictions ($\widehat{\tau}_i^{30CF}$) are estimated on the test sample for each forest, then averaged across the 30 causal forest models (with different training/test sample splits).

Figure 3: Predicted impacts on earnings*during* Vs *post* program



Note : Predicted (conditional) treatment effects obtained by 30 causal forest simulations, predictions made on test sample and averaged across simulations. For presentational purpose, the scatterplot is truncated for points above (resp. below) 99th (resp. 1st) percentile.

Table 5: Baseline characteristics, for bottom and top quartile of predicted impacts during the program

	(1)	(2)	(3)	(4)
	Mean in treatment (reference)	Mean in 1st quartile of predicted impacts *during* program	Mean in 4st quartile of predicted impacts *during* program	P-value (2) - (3)
Individual characteristics				
Gender	32%	22%	43%	0,00
Age	24,60	25,07	23,81	0,00
Nb of children	0,94	0,91	0,96	0,00
Live in urban area	80%	80%	75%	0,22
Education				
Did not complete primary school (No diploma)	47%	50%	42%	0,00
Has completed primary school (CEPE)	25%	25%	25%	0,74
Has completed middle school (BEPC)	18%	14%	23%	0,00
Has completed secondary school (BAC or +)	7%	7%	7%	0,65
Previous Vocational Training	39%	49%	29%	0,00
Household				
Household size (total number of members)	6,03	8,12	4,69	0,00
Is head of household	25%	19%	29%	0,00
Assets (Household level, last 3 months)				
Total Nb of Assets	13,87	22,02	8,10	0,00
Nb of Transportation assets (aggregated)	0,74	1,33	0,33	0,00
Nb of Agricultural assets (aggregated)	4,71	8,29	2,07	0,00
Nb of Household Equipment (aggregated)	1,65	2,60	1,00	0,00
Nb of Communication assets (aggregated)	6,78	9,80	4,71	0,00
Employment				
Has an activity	79%	90%	66%	0,00
Is engaged in wage-employment	34%	41%	28%	0,00
Is engaged in self-employment	36%	49%	23%	0,00
Total nb of activities	0,98	1,20	0,76	0,00
Total Earnings (monthly)	18 043,95	32 190,24	7 553,91	0,00
Savings, Constraints and Expenditures				
Has Saved (last 3 months)	48%	53%	46%	0,00
Savings Stock (FCFA)	28 637,17	48 115,37	17 535,29	0,00
Has a Savings Account	11%	15%	9%	0,00
Face Constraints to repay loans	20%	20%	19%	0,48
Face Constraints to access credit	50%	44%	59%	0,00
Transportation expenditure (last 7 days)	1 909,77	2 989,97	1 047,59	0,00
Communication expenditure (last 7 days)	1 723,40	2 925,42	876,18	0,00
Treatment	100%	75%	74%	0,62
Nb of observations	3 020	1 101	1 129	

Table 6: Impacts *during* the program on main outcomes, for top and bottom quartile of predicted impacts on earnings during the program ($\sim$ 4,5 months after program starts)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Has an	Wage Employed	Self Employed	Total Earnings	Earnings in	Earnings in	Savings	Well Being
	Activity	(in at least	(in at least		in Wage Empl.	in Self Empl.	(stock)	index
		1 activity)	1 activity)	(monthly)	(monthly)	(monthly)		(z-score)
Panel A : Top 25% group (top quartile of predicted impacts on earnings during the program)								
Treatment * Top 25% (Q4)	0.16***	0.50***	-0.11***	24797.19**	43337.86***	-17147.72***	33972.37***	0.24***
	(0.02)	(0.03)	(0.03)	(9965.05)	(6857.94)	(6598.99)	(6427.17)	(0.07)
Treatment * Rest	0.11***	0.41***	-0.08***	19247.11***	32257.10***	-10917.21*	41655.63***	0.18***
	(0.01)	(0.02)	(0.02)	(7251.80)	(4238.61)	(5686.54)	(2966.60)	(0.05)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control for Q4	0.82	0.46	0.31	47138.63	23195.45	21915.20	20261.49	-0.10
Mean in Control for Rest	0.88	0.56	0.34	65378.47	34101.21	27279.87	22371.09	0.04
P value (T*Q4 = T*Rest)	0.06	0.01	0.50	0.63	0.16	0.45	0.23	0.46
Observations	2957	2957	2957	2912	2912	2912	2957	2933
Panel B : Bottom 25% group (bottom quartile of predicted impacts on earnings during the program)								
Treatment * Bottom 25% (Q1)	0.10***	0.43***	-0.08**	2425.74	19605.94**	-14798.41	50286.69***	0.20***
	(0.02)	(0.03)	(0.03)	(13762.03)	(7990.53)	(11139.81)	(5149.37)	(0.07)
Treatment * Rest	0.13***	0.44***	-0.09***	28052.85***	41550.13***	-11817.96**	35344.64***	0.19***
	(0.01)	(0.02)	(0.02)	(6356.25)	(3959.95)	(4727.67)	(3479.50)	(0.05)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control for Q1	0.90	0.54	0.39	85917.29	44265.23	37664.54	25664.42	0.05
Mean in Control for Rest	0.85	0.53	0.31	49989.63	25723.35	21063.96	20255.87	-0.02
P value (T*Q1 = T*Rest)	0.13	0.75	0.70	0.09	0.01	0.81	0.01	0.92
Observations	2957	2957	2957	2912	2912	2912	2957	2933

Robust standard errors clustered at (large) brigade level. Earnings and Savings are in FCFA and winsorized at 99%.

* $p < .1$, ** $p < .05$, *** $p < .01$

Table 7: Impacts *post* program on main outcomes, for top and bottom quartile of predicted impacts on earnings during the program (12 to 15 months after program ends)

	(1) Has an Activity	(2) Wage Empl. (in at least 1 activity)	(3) Self Empl. (in at least 1 activity)	(4) Total Earnings (monthly)	(5) Earnings in in Wage Empl. (monthly)	(6) Earnings in in Self Empl. (monthly)	(7) Well Being index (z-score)	(8) Value of productive assets	(9) Total start-up capital for this activity	(10) Personal (start-up) capital for this activity	(11) Launched a new business in the last 1.5 yrs
Panel A : Top 25% group (top quartile of predicted impacts on earnings during the program)											
Treatment * Top 25% (Q4)	-0.02	0.01	-0.04	5973.71	-515.15	6903.74*	0.00	5272.59	9778.31**	10428.74***	0.05
	(0.03)	(0.04)	(0.04)	(4838.79)	(2153.34)	(4168.86)	(0.07)	(3528.48)	(4685.93)	(3534.35)	(0.03)
Treatment * Rest	0.02	0.00	0.04	4863.97	-1427.42	5797.50**	0.12**	9562.68***	9977.58**	9798.56***	0.06***
	(0.02)	(0.02)	(0.02)	(3348.50)	(1931.49)	(2768.06)	(0.05)	(2704.16)	(4283.76)	(2684.16)	(0.02)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control for Q4	0.85	0.50	0.35	38734.60	19694.82	16541.44	-0.01	12326.48	20698.05	12059.25	0.23
Mean in Control for Rest	0.87	0.58	0.32	52651.72	27724.63	20981.72	0.01	14634.52	28168.35	18785.19	0.22
P value (T*Q4 = T*Rest)	0.19	0.74	0.08	0.86	0.76	0.83	0.18	0.32	0.97	0.89	0.71
Observations	3733	3733	3733	3733	3733	3733	3731	3733	3733	3733	3733
Panel B : Bottom 25% group (bottom quartile of predicted impacts on earnings during the program)											
Treatment * Bottom 25% (Q1)	0.03	-0.00	0.07*	3424.38	-2895.16	8510.54*	0.16**	13171.53***	17450.86***	13705.28***	0.13***
	(0.03)	(0.04)	(0.04)	(5753.57)	(3262.03)	(4955.64)	(0.08)	(4682.82)	(6429.07)	(4548.48)	(0.03)
Treatment * Rest	0.01	0.01	-0.00	6012.68**	-435.90	5252.29**	0.06	6568.02***	7096.12*	8614.20***	0.03
	(0.02)	(0.02)	(0.02)	(2771.35)	(1640.05)	(2399.66)	(0.05)	(2278.01)	(3893.88)	(2391.87)	(0.02)
LocXGender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean in Control for Q1	0.89	0.60	0.32	61815.22	30773.61	24273.87	-0.00	16660.28	30388.52	21607.82	0.16
Mean in Control for Rest	0.86	0.54	0.33	43586.81	23364.19	17946.33	0.00	12934.51	24351.26	15028.36	0.25
P value (T*Q1 = T*Rest)	0.51	0.86	0.10	0.68	0.50	0.54	0.23	0.18	0.16	0.31	0.01
Observations	3733	3733	3733	3733	3733	3733	3731	3733	3733	3733	3733

Robust standard errors clustered at (large) brigade level. Earnings, Asset value and Capital are in FCFA and winsorized at 99%. Treatment pooled across arms.

* $p < .1$, ** $p < .05$, *** $p < .01$

Table 8: Summary of average impacts on earnings under alternative targeting approaches

	Random selection	Machine Learning prediction cond. on X		Self selection	Selection based on baseline characteristics		
	(0)	(1) (A)	(1) (B)	(2)	(3)	(4)	(5)
	Treated	Mean ($\widehat{\tau}^{CF}$) in	Mean ($\widehat{\tau}^{CF}$) in	Low	Women	Low baseline	Low baseline
	(ITT)	4^{th} quartile	1^{st} quartile	reservation		earnings	earnings
				wage for		(self-declared)	(predicted)
	(random 25%)	(top 25%)	(bottom 25%)	PW		(bottom 25%)	(bottom 25%)
Panel A : During the program							
Impact on total earnings (monthly)	20 885***	28 622	9 934	32 076***	37 150***	32 695**	36 822***
(s.e)	(6 194)			(6 913)	(6 257)	(7 738)	(9 519)
N	2 912	2 912	2 912	2 912	2 912	2 912	2 912
Panel B : Post program							
Impact on total earnings (monthly)	5 622**	8 329	1 475	8 121***	8 259**	13 305*	8 845**
(s.e)	(2 422)			(2 511)	(3 556)	(4 381)	(4 506)
N	3934	3 862	3 733	3 934	3 934	3 934	3 934

Column (1) (A and B) : $\widehat{\tau}^{CF}$ is the predicted treatment effect on earnings (conditional on covariate set X) when applying causal forest model on the test sample. Predictions are averaged across 30 simulations on different training/test splits of the data ; quartile groups are determined across simulations following a methodology described in B.3.1.

Column (2) : Outcome variable is a dummy for people who answered they would participate in the program if the wage was set to 1500 rather than 2500 FCFA / day. 45% willing to participate for lower daily wage. Variable measured at endline : balanced across treatment / control. Robustness checks done using the prediction of this variable on the control group, applied to the treatment.

Column (5) : Prediction using a restricted set of observable and non easily manipulated characteristics (gender, age, household characteristics and assets). Random forest algorithm.

Standard errors clustered at (large) brigade level for treated and individual level for control. For *Post* impacts, treatment is pooled across arms.

* $p < .1$, ** $p < .05$, *** $p < .01$

Table 9: Cost-Effectiveness Analysis

	(1)	(2)	(3)	(4)	(5)	(6)
	Survey measure	*Survey measure*	Cumulated	*Administrative data*	Cost	Nb of Years
	Impact on earnings	Impact on earnings	Impact on	Total Public	Effectiveness	after the
	during the program	post program	earnings	Works cost per	Ratio	program to
	(β^{During})	(β^{Post})	(over 20 mths)	beneficiary	(4) / (3)	be break-even
Current program (randomized)	20 885	4 100	206 695	660 478	3,2	23
Alternative Selections :						
(A) Low Reservation Wage for PW	31 926	7 409	332 790	660 478	2,0	6
(B) Women	37 314	7 998	379 191	660 478	1,7	4
(C) Low baseline Earnings (self-declared)	32 492	11 518	397 354	660 478	1,7	3
(D) Low baseline Earnings (predicted)	36 577	11 043	418 941	660 478	1,6	3

(3) Discounted sum of impacts on total earnings from program starts (month 1) up to 13 months after exit from the program (month 20).

Computed as $\sum_{k=1}^{7}(\rho^{k-4} * \beta^{During}) + \sum_{k=8}^{20}(\rho^{k-13} * \beta^{Post})$, with β^{During} (respectively β^{Post}) the *during* (respectively *post*-program) ITT estimates of monthly earnings impact. ρ is the monthly discount factor. $\rho = 1/(1+\delta)^{(1/12)}, \delta = 10\%$

(4) : Costs include both administrative implementation costs and the direct transfer made to beneficiaries.

(6) : Assuming (discounted) β^{Post} monthly impact after month 20, this is the number of years until total costs (4) equal cumulated impact on earnings (computed as (3)).

(A) : Group of individuals who answered they would participate in the program if the wage was set to 1500 rather than FCFA 2500 / day.

(C) : Bottom 25% of the distribution of baseline self-declared earnings.

(D) Bottom 25% of the distribution of baseline predicted earnings. Prediction using a restricted set of observable and non easily manipulated characteristics (gender, age, household characteristics and assets). Random forest algorithm.

Appendix A Description of the weights

We describe in this appendix the weights used in the estimations. They depend on the survey used (midline or endline data), the treatment status, the locality. All specifications in the paper use these weights. We provide a summary of the weights used with midline and endline data in A.5.

A.1 Randomization weights

A.1.1 First level of randomization : public lotteries

First, we consider weights related to the random selection into the program. Such weights should take into account the specificity of each public lottery held.

Our objective is to estimate an equation of the following type with weight w_i :

$$y_i = a + bT_i + u_i \tag{1}$$

One can easily check that estimator b is obtained as :

$$\widehat{b} = \frac{\sum_{i,T=1} w_i y_i}{\sum_{i,T=1} w_i} - \frac{\sum_{i,T=0} w_i y_i}{\sum_{i,T=0} w_i} \tag{2}$$

There are K different public lotteries [1] with N_k individuals participating to each lottery. Let's note N_{k1} the individuals from lottery k selected in the program (we will call them 'treated') and N_{k0} those who are not selected, with $N_k = N_{k1} + N_{k0}$. Among the N_{k0} , N_{k0s} are randomly selected to be surveyed and constitute the 'control group' of the experiment. The size of the population of lotteries' participants is N_P, with $N_P = \sum_k N_k = N_1 + N_0$. The size of the survey sample for the experiment is $N_E = \sum_k N_{k1} + N_{k0s} = N_1 + N_{0s}$.

We can rewrite equation (2) with weight w_{ki} ($i = 0_s; 1$ according to treatment status) assigned to individuals of the survey sample :

$$\widehat{b} = \frac{\sum_k w_{k1} N_{k1} \overline{y}^{k1}}{\sum_k w_{k1} N_{k1}} - \frac{\sum_k w_{k0s} N_{k0s} \overline{y}^{k0s}}{\sum_k w_{k0s} N_{k0s}} \tag{3}$$

To ensure this estimator can be interpreted as the weighted sum of the impact for each lottery, the following condition should hold :

$$w_{k0s} N_{k0s} = w_{k1} N_{k1} \tag{4}$$

The estimator rewrites as :

$$\widehat{b} = \sum_k \frac{w_{k1} N_{k1}}{\sum_k w_{k1} N_{k1}} \left(\overline{y}^{k1} - \overline{y}^{k0s} \right) \tag{5}$$

[1] In our experiment, public lotteries are held in each locality for each gender separately. Therefore $K = 16{*}2 = 32$

Following condition (4), we take $w_{k1} = 1$ and $w_{k0s} = N_{k1}/N_{k0s} \times N_0/N_1$. We obtain the following estimator:

$$\widehat{b} = \sum_k \frac{N_{k1}}{\sum_k N_{k1}} \left(\overline{y}^{k1} - \overline{y}^{k0s}\right) \tag{6}$$

Note that it means that a higher weight is put on lotteries for which the treated to not-treated ratio is higher[2]. It will be homogeneous across lotteries k (localities x gender) by construction of the survey sample. Note that these weights depend on the quota of treated (N_{k1}) that was assigned to each locality when the program was initially designed. It should be proportionate to the number of disadvantaged youth looking for employment, but such precise figures where not available at the time of design.

A.1.2 Second level of randomization : treatment arms assignment

Then, we want to add weights to take into account the second level of randomization in the experiment : assignment of treated individuals into 3 treatment arms denoted T_a, T_b and T_c. This is relevant when comparing treatment effects across arms, using endline survey data.

Similar to A.1.1, the equation estimated as the following form with a weight w_i:

$$y_i = \alpha + \beta_1 * T_{a,i} + \beta_2 * T_{b,i} + \beta_3 * T_{c,i} + u_i \tag{7}$$

As previously in, β_j estimators ($j = a, b, c$ for 3 options) are :

$$\widehat{\beta_j} = \frac{\sum_{i,T=j} w_i y_i}{\sum_{i,T=j} w_i} - \frac{\sum_{i,T=0} w_i y_i}{\sum_{i,T=0} w_i} \tag{8}$$

We use the same notations as A.1.1 for N_P (whole population), $N_E = N_1 + N_{0s}$ (survey sample population), N_{k1} and N_{k0s}[3].

Brigades of treated individuals (N_1) are assigned across 3 options T_a, T_b and T_c. The number of brigades assigned to the treatment arms varies by locality. We use the following notation : $N_k = N_{a,k} + N_{b,k} + N_{c,k} + N_{0,k}$ with $N_{1,k} = N_{a,k} + N_{b,k} + N_{c,k}$, and $N_P = \sum_k N_k = N_0 + N_a + N_b + N_c$ with $N_1 = N_a + N_b + N_c$.

We put a weight $w_{j,k}$ to treated individuals from lottery k who were assigned to treatment T_j, and weight w_{k0s} for non treated individuals selected in the survey sample. Similar to 3 with subscript $j = a, b, c$, (8) rewrites as :

$$\widehat{\beta_j} = \frac{\sum_k w_{j,k} N_{j,k} \overline{y}^{j,k}}{\sum_k w_{j,k} N_{j,k}} - \frac{\sum_k w_{k0s} N_{k0s} \overline{y}^{k0s}}{\sum_k w_{k0s} N_{k0s}} \tag{9}$$

[2] One could have chosen another option for the weight : $w_{k1} = N_k/N_{k1} \times N_1/N_P$ and $w_{k0s} = N_k/N_{k0s} \times N_0/N_P$. In that case, there will be a higher weight for lotteries where the demand for the program was higher (compared to the quota assigned).

[3] For endline survey, quantity N_{k0s} was increased compared to baseline. It affects weights computation through N_{k0s} and N_{0s} but other quantities remain unchanged.

To ensure this estimator can be interpreted as the weighted sum of the impact for each lottery, we need a condition similar to (4) and if this holds, the estimator can be written as :

$$\widehat{\beta}_j = \sum_k \frac{w_{j,k} N_{j,k}}{\sum_k w_{j,k} N_{j,k}} \left(\overline{y}^{j,k} - \overline{y}^{k0s} \right) \tag{10}$$

Similar to A.1.1, we choose the following weights :

- $w_{j,k} = N_{k1}/N_{j,k} \times N_j/N_1$ with $j = a, b, c$ [4]
- $w_{k0s} = N_{k1}/N_{k0s} \times N_0/N_1$

A.2 Sub-sampling weights (midline survey only)

The sample for midline survey ('during' program) includes the control group (N_{0s}) and a sub-sample of the treatment group.

Let's consider that we draw a random sub sample of group l in proportion $P_l = N_l^S/N_l$ (S the drawing variable). Original weights have to be multiplied by S/P_l to take sub-sampling into account.

Therefore, in group $l = k, 1$ for which one draws N_{k1}^S individuals out of N_{k1}, the original weight w_{k1} is updated to $\omega_{k1}^S = w_{k1} \times N_{k1}/N_{k1}^{S_{k1}}$. The weights for the control units, w_{k0s}, are unchanged as there is no sub-sampling of this group for midline survey.

A.3 Control group and potential ex-post enrolment in the program (endline survey only)

Individuals from control group need specific weights when using endline data, because some of them have been able to participate to the following waves of the program[5]. More precisely, control units were allowed to apply to wave 3 (apply meaning to take part to the lotteries) and wave 4. Such behavior could be tracked using administrative data. At the end of the fourth wave of THIMO program, each individual from the control group were in one the following 7 situations (for each locality) :

1. The individual applied to wave 3 (C_3), was selected as 'beneficiary' of wave 3 after public lotteries (T_3) and was therefore not allowed to apply to wave 4 ($\bar{C}_4$). This group is noted $C_3T_3\bar{C}_4$.
2. The individual applied to wave 3 (C_3), was not selected after public lotteries ($\bar{T}_3$), applied to wave 4 (C_4) and was selected as 'beneficiary' of wave 4 after lotteries (T_4). This group is noted $C_3\bar{T}_3C_4T_4$.
3. The individual applied to wave 3 (C_3), was not selected after public lotteries ($\bar{T}_3$), applied to wave 4 (C_4) and was not selected after lotteries ($\bar{T}_4$). This group is noted $C_3\bar{T}_3C_4\bar{T}_4$.

[4] Note : $\sum_j w_{j,k} = w_{k1} = 1$, which is the weight used for midline data when there is only one treatment group.

[5] Recall that the study exploits wave 2 (out of 4 waves) of the THIMO program

4. The individual applied to wave 3 (C_3), was not selected after public lotteries ($\bar{T}_3$) and he did not apply to wave 4 ($\bar{C}_4$). This group is noted $C_3\bar{T}_3\bar{C}_4$.
5. The individual did not apply to wave 3 ($\bar{C}_3$), applied to wave 4 (C_4) and was selected as 'beneficiary' of wave 4 after public lotteries (T_4). This group is noted $\bar{C}_3C_4T_4$.
6. The individual did not apply to wave 3 ($\bar{C}_3$), applied to wave 4 (C_4) and was not selected after public lotteries ($\bar{T}_4$). This group is noted $\bar{C}_3C_4\bar{T}_4$.
7. The individual did not apply to wave 3 ($\bar{C}_3$), and did not apply to wave 4 ($\bar{C}_4$). This group is noted $\bar{C}_3\bar{C}_4$.

This idea is that we don't want to include in the estimations control units that have benefited from further waves of the program (waves 3 and 4), which are precisely groups $C_3T_3\bar{C}_4$, $C_3\bar{T}_3C_4T_4$ and $\bar{C}_3C_4T_4$ following the notations introduced before. To compensate for that, we want to put a higher weight on individuals who had the exactly same behavior (towards wave 3 and 4) but were (randomly) not selected into the program. We introduce a new multiplicative weight for control units ($\widetilde{w}_{k0s,j}$) to control for that.

Intuitively, if there had been only one wave at which individuals could apply (C) and be selected (T), weights would have been :

- $\widetilde{w}_{k0s,\bar{C}} = 1$
- $\widetilde{w}_{k0s,C,\bar{T}} = N_{k0s,C} \big/ N_{k0s,C,\bar{T}}$
- $\widetilde{w}_{k0s,C,T} = 0$

Taking into account the two waves, the weights follow the seven groups previously defined :

- $\widetilde{w}_{k0s,C_3T_3\bar{C}_4} = 0$
- $\widetilde{w}_{k0s,C_3\bar{T}_3C_4T_4} = \frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}} \times 0 = 0$
- $\widetilde{w}_{k0s,C_3\bar{T}_3C_4\bar{T}_4} = \frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}} \times \frac{N_{k0s,C_3\bar{T}_3C4}}{N_{k0s,C_3\bar{T}_3C_4\bar{T}_4}}$
- $\widetilde{w}_{k0s,C_3\bar{T}_3\bar{C}_4} = \frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}} \times 1 = \frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}}$
- $\widetilde{w}_{k0s,\bar{C}_3C_4T_4} = 1 \times 0 = 0$
- $\widetilde{w}_{k0s,\bar{C}_3C_4\bar{T}_4} = 1 \times \frac{N_{k0s,\bar{C}_3C4}}{N_{k0s,\bar{C}_3C_4\bar{T}_4}}$
- $\widetilde{w}_{k0s,\bar{C}_3\bar{C}_4} = 1$

One can easily check that the sum of weights gives the total number of individuals in control group [6]

A.4 Tracking weights

We want to add a weight taking into account the differential response rate of individuals during each survey (midline and endline. More precisely, one can consider that a given survey consists in two phases a and b :

[6] $(\frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}} \times N_{k0s,C_3\bar{T}_3C_4} + \frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T}_3}} \times N_{k0s,C_3\bar{T}_3\bar{C}_4}) + (N_{k0s,\bar{C}_3C_4} + N_{k0s,\bar{C}_3\bar{C}_4})$ That is : $N_{k0s,C_3} + N_{k0s,\bar{C}_3} = N_{k0s}$

- The main data collection phase (a), during which the response rate is $R_{a,j}$ for group $j = 1, 0$.
- An additional tracking phase (b), targeting attritors from the first phase. We note $R_{b,j}$ the response rate of the tracking phase for group $j = 1, 0$.

To determine the tracking sample, we first define a sub-sample of 'eligible' attritors[7] $E_{b,j}$ from which a random sub-sample will be drawn in proportion $\pi_j = NE^S_{b,j}/NE_{b,j}$ (j is an index for treatment status x locality).

Individuals who responded (only) during the tracking phase should have a different weight than those who responded during the main survey phase. To take this selection into account, tracking respondants should be weighted by $\omega^T_j = (R^S_{a,j} + \lambda_j s_j R^S_{b_j}(1 - R^S_{a,j})E^S_{b,j}$, with λ_j to be determined, so final weight is $\omega^{S,f}_j = \omega^S_j \times \omega^T_j$.

The sum of the weights on population j is therefore : $\omega_j \times (N^S_{a,j} + \lambda_j NER^S_{s,b_j})$, with NER^S_{s,b_j} the number of individuals who responded during tracking phase (in group S drawn). We make the hypothesis that residual non response $R^S_{b,j}$ is random. The population for which we want to be representative is the respondent population of phase a and the respondent population of phase b. This lead us to take $\lambda_j = NE^S_{b,j}/NER^S_{s,b,j}$

In group j, weights will be set such as[8] :

- $\omega^S_j \times 1$ for phase a respondents
- $\omega^S_j \times NE^S_{b,j}/NER^S_{s,b,j}$ for phase b respondents

Tracking weights ω^T_j are multiplied to the previous weights.

A.5 Synthesis of the weights used for midline and endline data

[7] Among the attritors of phase (a) some individuals are considered 'non eligible' for tracking in order to exclude them from the tracking draw. Non eligible attritors are those considered to be impossible or quasi impossible to reach (which is why we don't want to put extra effort on them) : dead individuals, individuals who migrated to another country, (for endline) individuals who were already impossible to find at baseline 1.5 years ago.

[8] In theory, ω_j should be adjusted so that it does not use correction N_j/N^S_j but rather the correction corresponding to the total of eligibles $N_{a,j} + NE_{b,j}$. However, this number is only known for selected units $S_j = 1$. Therefore we will ignore this aspect, which is fair considering that units where randomly drawn. Finally, it means that we estimate the unknown amount $N_{a,j} + NE_{b,j}$ by $N^S_{a,j} + NE^S_{b,j} \times N_j/N^S_j$

Randomization weights w_k		**Sub Sampling weights** ω^S_k		**Tracking weights** ω^T_j	
Application criterion	Weight computation	Application criterion	Weight computation	Application criterion	Weight computation
Treated	$w_{k1} = 1$	Treated	$w^S_{k,1} = N_{k1}/N^{S_{k1}}_{k1}$, k=locality	Respondents main phase ($R_a = 1$)	$\omega^T = 1$
Control	$w_{k0s} = N_{k1}/N_{k0s} \times N_0/N_1$, k=locality x gender	Control	$w^S_{k,0} = 1$	Non Respondents main phase ($R_a = 0$)	$\omega^T_j = NE^S_{b,j}/NER^S_{s,b,j}$ if respondent in tracking phase ($E_b = 1$ et $R_b = 1$), j=locality x treatment status
					$\omega^T = 0$ if non respondent (but sampled) in tracking phase ($E_b = 1$ et $R_b = 0$)
					$\omega^T = 0$ if not sampled for tracking phase ($E_b = 0$)
Final weight: $w^F_{k,i} = w_{k,i} \times \omega^S_{k,i} \times \omega^T_{k,i}$, $i = 0, 1$ (treatment status), $k \in [\![1, 32]\!]$ (locality x gender)					

Table 10: Synthesis of the weights used with midline data

Randomization weights $w_{j,k}$		**Post-enrollment weights $\widetilde{\omega}_{k,j}$**		**Tracking weights ω_j^T**	
Application criterion	Weight computation	Application criterion	Weight computation	Application criterion	Weight computation
Treatment arm $T_a,\ T_b$ or T_c	$w_{j,k} = N_{k1}/N_{j,k} \times N_j/N_1,\ j = a, b, c$	Selected to participate to wave 3 or 4 (groups $C_3T_3\bar{C_4},\ C_3\bar{T_3}C_4T_4$ et $\bar{C_3}C_4T_4$)	0	Respondents main phase $(R_a = 1)$	$\omega^T = 1$
Control	$w_{k0s} = N_{k1}/N_{k0s} \times N_0/N_1$, k=locality x gender	Group $C_3\bar{T_3}C_4\bar{T_4}$	$\frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T_3}}} \times \frac{N_{k0s,C_3\bar{T_3}C4}}{N_{k0s,C_3\bar{T_3}C_4\bar{T_4}}}$	Non Respondents main phase $(R_a = 0)$	$\omega_j^T = NE_{b,j}^S/NER_{s,b,j}^S$ if respondent in tracking phase ($E_b = 1$ et $R_b = 1$), j=locality x treatment status
		Group $C_3\bar{T_3}\bar{C_4}$	$\frac{N_{k0s,C3}}{N_{k0s,C_3\bar{T_3}}}$		
		Group $\bar{C_3}C_4\bar{T_4}$	$\frac{N_{k0s,\bar{C_3}C4}}{N_{k0s,\bar{C_3}C_4\bar{T_4}}}$		$\omega^T = 0$ if non respondent (but sampled) in tracking phase ($E_b = 1$ et $R_b = 0$)
		Group $\bar{C_3}\bar{C_4}$	1		$\omega^T = 0$ if not sampled for tracking phase ($E_b = 0$)
Final weight: $w_{k,i}^F = w_{j,k} \times \widetilde{\omega}_{i,l} \times \omega_{k,i}^T$, $j = 0, a, b, c$ (treatment status), $i = 1, 0s$, l post-enrollment group, $k \in [\![1, 32]\!]$ (locality x gender)					

Table 11: Synthesis of the weights used with endline data

Appendix B Applying Causal Forests to study heterogeneous treatment effects

In this section we provide details on our implementation of the causal forest algorithm, including the selection of the sample, the list of covariates and the quartile group dummies used in tables 6 and 7.
We used **causalTree** package[9] which implements causal trees as introduced in Athey and Imbens (2016) and contains causalForest fonction corresponding to Wager and Athey (2016) causal forests. The two papers previously cited provide comprehensive details on causal trees and forests structure and properties. Davis and Heller (2017a) provides a clear synthesis of how causal forests works.

B.1 Sample for the Causal Forest algorithm

As a supervised learning algorithm, causal forests require to work on samples for which both covariates and outcomes are observed. In our case, this means observing baseline covariates (set of K covariates, X^K) and midline and/or endline outcome of interest. Therefore we first need to address two main concerns in our data : attritors across survey rounds and missing values among baseline covariates. On top of that, some specificities of our surveys have to been taken into account : the reduced sample used for midline survey and the share of control individuals who got into later waves of the public works program between midline and endline surveys. It leads to three potential samples we can use for the algorithm : a 'midline' $(X_i^K, Y_i^{During}, W_i)$ (respectively 'endline' (X_i^K, Y_i^{Post}, W_i)) sample that can be used to build and apply the model to predict 'during' (respectively 'post') conditional treatment effects. A third (marginally smaller) sample can be used when one wants to study how effects vary between 'during' and 'post' program (intersection of non attritors and non-missing outcomes for both surveys).

We directly exclude from each algorithm sample attritors from follow-up surveys (outcomes being unobserved). Attrition was limited in our surveys, but one has to keep in mind that the prediction model derived from the data would not apply to 'attritors'.
Missing values among baseline covariates are replaced by the mean in the rest of the sample, and a binary indicator for missing value on this covariate is created and included in the final set of covariates[10]. Individuals with missing values for the outcome of interest (among non attritors) are dropped from the sample.

Differences between midline and endline survey sample sizes create constraints. Approximately one third of the full sample was not included in the midline survey. These individuals form what we call the 'Midline-out(-of-sample)' and cannot be used to build the model to predict treatment effects during the program . However, these units will be included in the sample when we build the post program prediction. How do we deal with that ? When comparing 'during' to 'post' program predicted treatment effects, we recover a prediction for the 'Midline-out' individuals by

[9] Susan Athey, Guido Imbens and Yanyang Kong (2016). causalTree: Recursive Partitioning Causal Trees. R package version 0.0. We made small changes to the algorithm implemented by adding stratification for subsampling, see B.2.2.

[10] For consistency we checked on the forests the frequency of splits made on these missing-value dummies, which is indeed marginal over the trees compared to the main baseline covariates.

applying the model built with midline sample (exactly as one would do with a test sample, as the 'Midline-out' is independent from the sample used to build and estimate the model).[11]
When working on the endline algorithm, we drop control individuals who might have participated in a later wave of the public works program as it could lead to under-predict treatment effects in the model. Recall that 200 individuals were sampled to be added to the control group at baseline, to compensate for that : because these individuals were not part of the baseline survey, the machine learning model cannot be applied to them (prediction relying on observed X^K).

The final sample size (N_{all}) for the algorithm depends on the number of missing variables for the outcome considered, and whether you build separately the model with midline and endline survey data or jointly. The total sample that can be used for the midline algorithm ranges between 2,884 and 2,958 units ; for endline algorithm between 3,745 and 3,910 units ; to build jointly midline and endline predictions the sample is reduced at 3,700 units.

For the features (X^K), we use an extensive set of covariates ($K = 101+$ dummies for missing values in covariates) measured at baseline (Table 13) that covers both individual and household characteristics and include measures on time and risk preferences, personality traits and cognitive skills. As previously mentioned, we add binary indicators to this set for missing values in each covariate. This is a large set of features compared to recent applications ($K = 19$ in Davis and Heller, 2017a.)

B.2 Building the model

B.2.1 Causal forest parameters

First, some parameters have to be set ahead of the procedure ('tuning' the forest) : the test sample fraction, the subsampling share, the fraction for re-estimation, the minimum units of control group within a leaf and the number of covariates considered for each tree. We provide in Table 12 a description of each parameter to set, its impact on the causal forest algorithm, the value chosen and its justification. There is no clear rule so far for the tuning of these parameters. However, most of our choices are driven by sample size constraints : indeed, even if we have a relatively 'large' sample for an experiment on public policies in developing countries, it is small compare to standard 'big data' samples (e.g. in tech industry).[12]

The splitting criteria used follows Athey Imbens (2016) recommendation[13] ('Causal Tree' criterion, 'CT'). This criteria can be seen as an objective function (Q) used to determine the partition of the

[11] The 'Midline-out(-of-sample)' units form an 'independent' sample for which baseline covariates are observed, therefore one can apply the prediction model to this sample and use it later, including for inference, in the spirit of test samples. When comparing 'during' and 'post' program, we randomly split the 'Midline-out(-of-sample)' to add it to both the training and test samples that will be used for the endline algorithm. The test sample of the endline prediction remains a 'true' test sample as none of the units where used for the algorithms, whether midline or endline.

[12] Note that this list of parameters to set excludes the penalty term controlling for the 'complexity' of the tree-model (multiplied to the number of splits). In regression trees, the penalty parameter is 'empirically tuned' by cross-validation. With causal forest, there is no cross-validation during tree building as trees are grown deep and the CT splitting criteria proposed by Athey and Imbens (2016) directly incorporates a penalty term.

[13] Athey Imbens (2016) detail advantages and drawbacks of four potential splitting rules, the preferred one being the 'CT' criteria. For three of these criteria, an 'adaptative' as well as an 'honest' version of the criteria exist. We use the honest version of the criteria, 'CT-H'. More details provided in B.2.3.

covariate space. It is maximized at each step (as a tree grows) such that the algorithm splits the data into sub-groups when heterogeneity in treatment effects is detected (more details in B.2.3).

The share of the sample used to construct the model ($S_{tr} = (1-\alpha)N_{all}$) compared to the share used to make inference ($S_{te} = \alpha N_{all}$) is chosen by the user. Again, there is no clear 'rule' to decide on that, although in large datasets test samples are traditionally small (10%). In our case, the modest total sample size lead us to consider 'statistical power' for the selection of this parameter. We take $\alpha = 50\%$, so half of the sample for training (and the same size for predictions and inference).

B.2.2 Step-by-step causal forest procedure

Figure 4 provides an illustration of the main steps described below to assist the reader in the understanding of the methodology. The procedure builds on causal trees introduced by Athey and Imbens (2016) and corresponds to the causal forest (with honesty) procedure described by Wager and Athey (2016).

We start with the full sample of observations (X_i^K, Y_i, W_i) from the experiment where X^K denotes the covariates space (K baseline measures), Y is the observed outcome of interest (with $Y_i = Y_i(W_i)$), and $W_i \in \{0,1\}$ the treatment status. We note N_{all} the total sample size used for the algorithm. Some parameters of the causal forest have to be set upfront (see B.2.1).

Main steps for causal forest model :

1. The full sample is (randomly) split in two non-overlapping samples : one training sample (S_{tr}) on which the causal forest (CF) model will be built [15], and one test sample [16] (S_{te}) to which the CF model will be applied and to make inference. α is the share of observations assigned to the test sample (e.g. 50%). We have $N_{all} = |S_{tr}| + |S_{te}| = N_{tr} + N_{te}$.
2. Build your model on the training sample (S_{tr}) of size $N_{tr} = (1-\alpha) * N_{all}$.
   - Draw B 'bootstrapped' subsamples (S_{tr}^b) out of S_{tr} sample. Each subsample contains $N_{tr}^b = \beta * N_{tr}$ observations. The model is a forest of B honest trees, each one being constructed on one of the bootstrapped sub-samples of the training sample S_{tr}.[17]
   - Each tree b is obtained following the next steps, repeated for $b = 1$ to B :

     2.1 Randomly select *ncov* baseline covariates among the set of K covariates. Only these covariates will be considered as 'candidates' for the splits in tree b.

[14] This parameter is in **rpart** package for random forests, but it is not currently implemented in the causal forest packages (neither **causalForest** by Wager, nor **causalTree** by Athey). The next parameter, *ncov* plays a similar role and is the one currently implemented to 'decorrelate' trees in causalForest function of **causalTree**

[15] Training sample or 'in-sample observations' in the literature.

[16] Test sample, ' hold out' sample or 'out-of-sample' units in the literature. Test samples are traditionally used to assess the performance of the predictive model built (or to compare its performance to other models), compared to the training sample on which the model performance might seem 'too good' due to over-fitting. Following Athey and Imbens (2016), we use the test sample to make inference as it is independent from the construction of the model.

[17] If $\alpha = \beta = 50\%$, then $25\% N_{all}$ is used to build the model of a tree.

Table 12: Parameters of the causal forest

Parameter	Description	Value
B Number of trees in the forest	Compared to forests, single trees suffer from high variance : given a 'small' dataset, predictions obtained on a given training set compared to another can differ a lot. One way to reduce the variance of such method is to do bootstrap aggregating ('bagging'), using the fact that averaging a set of (independent) observations reduces variance. In the absence of multiple (independent) training datasets, one can draw repeated samples from the single training dataset by generating B different bootstrapped training datasets on which B trees will be grown. The trees are grown deep and are not pruned, the predictions are averaged across all trees. Using a large value of B will not lead to overfitting. Usually B is set as a trade-off between computational (time) cost and reduction of the test error rate. A large number of trees also provide more stability in the predictions obtained by reducing the Monte Carlo error introduced by subsampling : predictions will vary less across different forests.	10,000
$n_{minsize}$ Minimum number of treatment and control units per leaf	This parameter is specific to causal forests (compared to classic random forests) and is introduced by Athey and Imbens (2016). Setting a too low value can increase the noise of the prediction (increases the variance) as the predicted CATE could be estimated with very few control and treatment units in a given leaf. Setting a too high value forces the trees to be less deep with bigger leaves, which decreases the 'precision' of the prediction (so increases bias) : on the one hand it will mechanically predict less heterogeneity, on the other hand it will make predictions more consistent across trees / forests.	10
β Fraction of the subsample used to build each tree	Smaller subsamples reduce dependence across trees but increase the variance of each estimate (for a tree). In our case, the choice of this parameter is driven by statistical power constraints, as predictions in causal forest will computed on a sample of size $(1-\alpha)*\beta*\delta*N$. To set it at a 'maximum' level while keeping the benefits of subsampling, we set it at the intermediate level (50%).	0.5
δ Fraction of the subsample used for training	It determines the size of the training subsample used to build the partition of the covariate space, the rest of the sample being used to estimate treatment effects within leaves. Although original **causalForest** package fixed it at 50%, the updated causalForest function in **causalTree** allows the user to fix it to a different value. We have no reason to allocate more units to one of the two tasks : both require reasonable sample size to perform well, and in our case total sample size is already a constraint, so we set it at 50%.	0.5
v Nb of covariates X^k considered at each split (within a tree)	The predictor subset size (v) is what makes random forests different from bagged trees (bagging = bootstrapped aggregating, see comment on B). Let K be the total number of predictors (X). In random forests, each time a split is considered in a tree, a random sample of $v < K$ predictors is considered for the split. In bagged tree, $v = K$. This procedure 'decorrelates' the trees : trees are less likely to be similar, so predictions will be less correlated, and the average of these predictions across trees will generate a larger reduction in variance. For random forests, people typically use $v = \lfloor\sqrt{K}\rfloor$ for classification tree and $v = \lfloor K/3 \rfloor$ for regression tree.	NA[14]
$ncov$ Nb of covariates X^k considered before building each tree (within a forest)	Underlying idea is similar to the one above : 'decorrelating' the trees, such that the aggregation of predictions across trees will reduce the variance of the prediction. This parameter sets the number of covariates that are randomly subsampled *before building a tree*, compared to previous parameter v for which covariate subsampling occurs each time a split is considered in a tree.	$\lfloor K/3 \rfloor$

To be called 'honest' (Athey and Imbens, 2016), the tree has to be built in a two step procedure on two separate samples of S^b_{tr} :

2.2 (Randomly) split the sample S^b_{tr} in two : one sample ($S^b_{tr,tr}$) to build the structure of the tree (the partition of the data) ; the other sample ($S^b_{tr,re}$) to 're-estimate' the prediction for each leaf of the tree. δ is the share of S^b_{tr} used for re-estimation[18].

2.3 Using only $S^b_{tr,tr}$, build the structure of the tree : split the covariate space to determine a partition (into leaves) that maximize the variance of the treatment effects across leaves while limiting overfitting by penalizing any split that would increase within-leaf variance. Each final leaf of the tree (also called 'node') will correspond to an element of the partition and determine a subgroup of individuals based on observed covariates. We provide below more details on the the splitting criteria used in B.2.3. This step defines the assignment rules that map an observation to a specific leaf based on its covariate values only. Another way to see this step is that you build a mapping between vectors of X (attributes) and (not computed yet) predictions.

2.4 Using only the 're-estimation' sample $S^b_{tr,re}$, assign a prediction value to each leaf (or terminal node) given the partition previously determined. More precisely : assign each observation of $S^b_{tr,re}$to its corresponding leaf (based on the partition) and compute the "predicted treatment effect" conditional on being in leaf l. Treatment effect in leaf l is estimated as

$$\widehat{\tau_{l,b}} = \overline{Y_{w=1,l}}^{S^b_{tr,re}} - \overline{Y_{w=0,l}}^{S^b_{tr,re}}$$

in a given tree b. Another way to see this step is that the mapping is completed by computing the predictions.

At this step, a final honest tree b is obtained : a partition of the covariate space is defined, and there is a mapping associating any vector (X^K_i) to its 'conditional average treatment effect' (CATE) $\widehat{\tau_{l,b}}$.

2.5 Apply the model to the test sample to get a predicted treatment effect for each unit of S_{te}. More precisely : each unit is assigned $\widehat{\tau_{l,b}}$ depending on the leaf to which it was mapped (given observed vector of covariates (X^K_i)).

- *Remark :* Note that for each tree, a share of $(1-\beta)$ of S^b_{tr} was not subsampled for tree building (neither to build the partition, neither to estimate predictions within leaves). We call them '*quasi test observations*' (noted $S^b_{tr,te}$) [19], as they are independent from the tree model as much as the (separate) test sample.

3. For each unit of the test sample, the causal forest model assigns a prediction that is the

[18] If $\delta = 50\%$, then the structure of the tree (respectively the estimation of treatment effect within each leaf) is done using $50\% * 25\% N_{all} = 12,5\% N_{all}$ for a given tree.

[19] Davis and Heller (2017a) apply each causal tree model to these units and refer to them as 'adjusted training' sample. They show that the aggregated prediction obtained across trees on the adjusted training sample are very close to the prediction obtained on the test sample. In Davis and Heller (2017b), they do not use test samples but apply the causal forest model on the 'adjusted training sample' to recover for the whole sample predicted treatment effects later used for inference. We do not follow the same methodology : we keep a test sample but perform several causal forests (on different splits of the data) to recover an aggregated predicted treatment effect for each unit of the sample.

average of the predictions obtained in each tree-model :

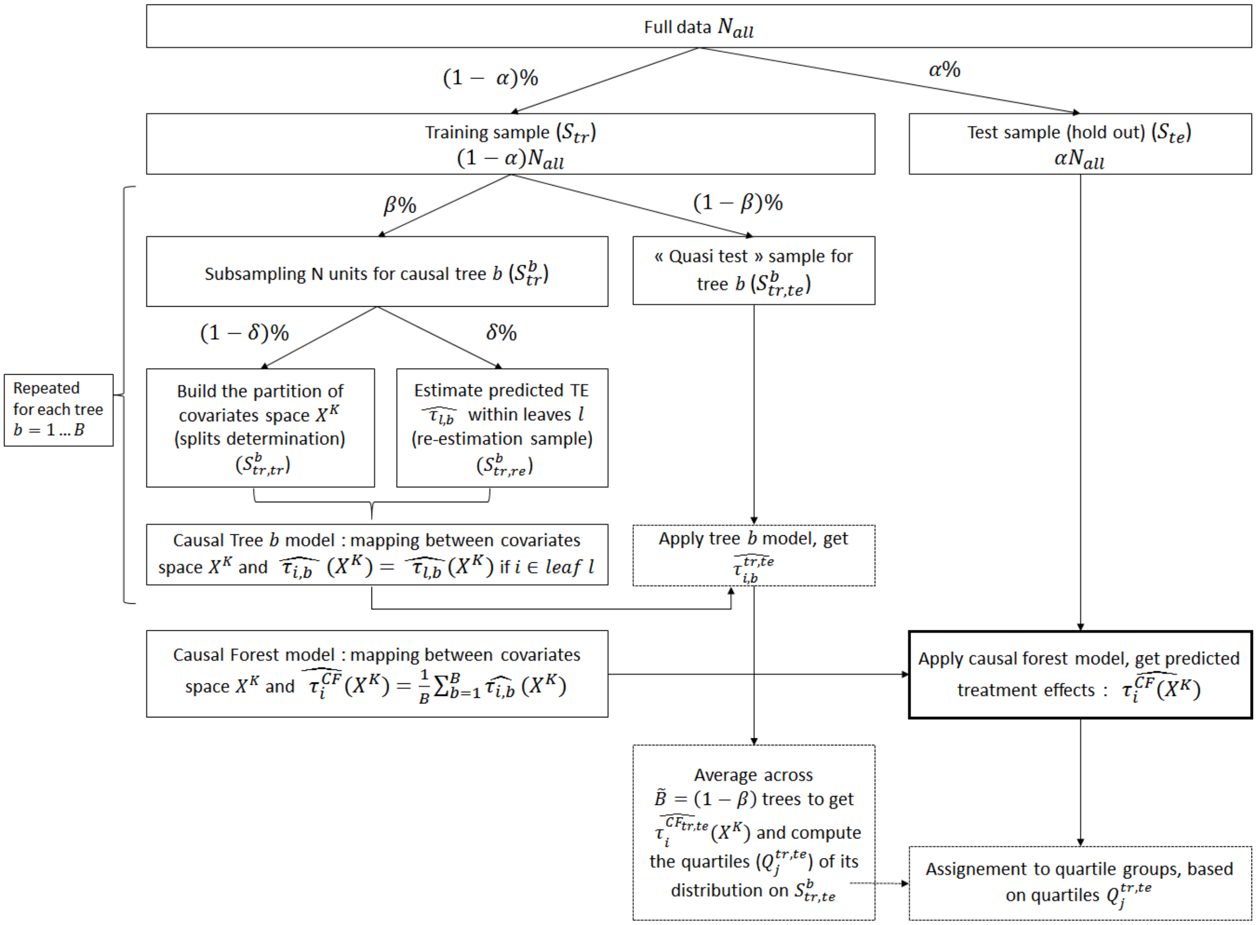
$$\forall i \in S_{te}, \widehat{\tau}_i^{CF} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\tau_{i,b}}(X_i^K)$$

This means that you get an individual prediction $\widehat{\tau}_i^{CF}$ for αN_{all} individuals, i.e. 50% of the initial sample in our application.

A comment on random partition and subsampling of the data : Note that several steps require either partitioning or subsampling our dataset : the determination of test sample (α%) versus training sample at the forest level ; the (β%) subsampling of the training dataset for each tree-building ; the (δ%) split of this subsample into tree-building and leaf-reestimation samples for each tree too. We adapted the algorithm so that each of these 'subsampling' procedures is stratified by our randomization blocks ($locality * gender$) as well as treatment status, therefore 96 strata in total[20]. Such stratification ensures that the sample on which the partition is constructed, as well as the sample used to estimate leaf treatment effect and the test sample are representative of these strata.

[20] In particular, for endline data, we want to make sure that treatment arms shares are identical between training and test samples.

Figure 4: Applying Causal Forests, Main Steps


Full data N_{all}
$(1-\alpha)\%$
$\alpha\%$
Training sample (S_{tr})
$(1-\alpha)N_{all}$
Test sample (hold out) (S_{te})
αN_{all}
$\beta\%$
$(1-\beta)\%$
Subsampling N units for causal tree b (S^b_{tr})
« Quasi test » sample for tree b ($S^b_{tr,te}$)
$(1-\delta)\%$
$\delta\%$
Repeated for each tree $b = 1 \dots B$
Build the partition of covariates space X^K (splits determination) ($S^b_{tr,tr}$)
Estimate predicted TE $\widehat{\tau_{l,b}}$ within leaves l (re-estimation sample) ($S^b_{tr,re}$)
Causal Tree b model : mapping between covariates space X^K and $\widehat{\tau_{i,b}}\ (X^K) = \widehat{\tau_{l,b}}(X^K)$ if $i \in leaf\ l$
Apply tree b model, get $\widehat{\tau^{tr,te}_{i,b}}$
Causal Forest model : mapping between covariates space X^K and $\widehat{\tau^{CF}_i}(X^K) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\tau_{i,b}}\ (X^K)$
Apply causal forest model, get predicted treatment effects : $\widehat{\tau^{CF}_i(X^K)}$
Average across $\tilde{B} = (1-\beta)$ trees to get $\widehat{\tau^{CF_{tr,te}}_i}(X^K)$ and compute the quartiles ($Q^{tr,te}_j$) of its distribution on $S^b_{tr,te}$
Assignement to quartile groups, based on quartiles $Q^{tr,te}_j$

B.2.3 More details on the tree splitting procedure

Each tree-based model b is constructed on a (bootstrapped) subsample S^b_{tr}. Recall that this sample is randomly split in two : a sample to build the structure of the tree ($S^b_{tr,tr}$) and a sample to estimate the treatment effect within each leaf ($S^b_{tr,re}$). Let $N^b_{tr,tr}$ be the number of observations in $S^b_{tr,tr}$, and $N^b_{tr,re}$ for $S^b_{tr,re}$ with $N^b_{tr,re} = \delta N^b_{tr}$.

The splitting rule we use is the honest causal tree splitting rule (CT-H) defined by Athey Imbens (2016). It is an (adjusted) mean square error criteria, which rewards a split finding heterogeneity in treatment effects (left term in the following formula) and penalizes a split that increases variance in leaf estimates (right term) :

$$Q = \frac{1}{N^b_{tr,tr}} \sum_{i \in S^b_{tr,tr}} \widehat{\tau_{l,b}}^2(X_i^K) - \frac{2}{N^b_{tr,tr}} \sum \left(\frac{\widehat{Var(Y_{l,w=1}})}{p} + \frac{\widehat{Var(Y_{l,w=0}})}{1-p} \right))$$

with $p = P(W_i = 1)$. This criteria is applied to each leaf to decide whether to perform the split or not, starting with the full $S^b_{tr,tr}$ at the beginning of the tree.

Compared to the traditional CART 'adaptative' splitting criteria, two main modifications are made by Athey Imbens (2016) : (i) the criteria incorporates the fact that 'honesty' (i.e. using an independent sample for leaf means estimation) generates unbiased estimates (which already reduces over-fitting); (ii) the criteria explicitely incorporate the fact that more splits (finer partitions) generate greater variance in leaf estimates, to limit over-fitting. With adaptative splitting criteria, to penalize splits that increase the variance within leaves a penalty term of the form $\lambda|T|$ is added to the goodness-of-fit criteria, where $|T|$ (number of splits) captures the complexity of the tree and λ the penalty paramater is determined by cross-validation.

The initial node contains all observations $N^b_{tr,tr}$. The following steps (1 to 4) are repeated for each sub-node. The tree stops growing when all nodes are terminal nodes.

1. One X^k at a time among X^K set, and considering all values x taken by X^k , form candidate splits of the current node into two potential sub-nodes (based on $(X^k \leq x)$ or not)
2. Consider only splits creating sub-nodes in which there is at least $n_{minsize}$ treatment and $n_{minsize}$ control observations (if there is no splits such that this constraint holds, this is a terminal node).
3. Choose the split that maximizes objective function Q capturing how much the treatment effect estimates vary across the two subgroups with an included penalty for generating within-leaf variance.
4. Perform this split if it increases Q relative to no splits. If there is a split : there are two new nodes and we repeat the procedure from step 1. If there is no split : this is a terminal node.

The structure of the tree is completed when all nodes are terminal nodes. These terminal nodes form a partition of the covariate space, they are the 'leaves' of the tree, and determine the subgroups. Note that the partition was obtained only using $S^b_{tr,tr}$.

B.3 Using predicted treatment effect (CATE) to study heterogeneity

B.3.1 Assignment to quartiles of the distribution

To study heterogeneity we want to focus on the upper and lower tails of the distribution of the predicted treatment effect. In particular, we are interested in the average CATE ($\widehat{\tau_i}^{CF}$) within each quartile, more specifically the 'top' quartile (predictions above quantile 75%) and the 'bottom' quartile (predictions below quantile 25%).

One concern we might have is that using the test sample (S_{te}) distribution of $\widehat{\tau}^{CF}$ to assign test sample units to a given quartile-group is biased as we are using 'endogenous' cut-offs. Ideally, we would like to use quartiles of $\widehat{\tau}^{CF}$ distribution estimated on a separate sample, and assign test sample units based on these 'exogeneous' thresholds.
However, we obviously don't want to use for that $\widehat{\tau}^{CF}$ distribution on either $S^b_{tr,tr}$ or $S^b_{tr,re}$ (units used in any b tree to determine the partition of the covariate space for the former, and to estimate the leaf treatment effect for the latter) as this would produce 'endogenous' cut-offs too, even more likely to be biased as they come from a distribution of predictions very likely to overfit the data[21].

[21] We indeed see that the distribution of $\widehat{\tau}^{CF}$ on units used for training (splitting or re-estimation) suffered from

One good sample candidate is the previously mentioned 'quasi test observations' (see B.2.2) : for a given tree b, sample $S^b_{tr,te}$ was not used at all to build the model due to the sub-sampling of S_{tr}, (recall that only β% of the training sample S_{tr} is used for a given tree). Therefore, for each tree b one can use $S^b_{tr,te}$ to apply the model of tree b just 'as' one would do with the test sample.

Figure 4 shows how 'quasi test' units are used and where quartile assignment intervenes in the overall procedure. These steps are in dashed lines as this approach has been newly introduced and departs from the causal forest implementation currently described in the literature.

Detailed steps to obtain the quartile 'cut-offs' using $\widehat{\tau}^{CF}$ distribution on a sample separate from both test and 'training':

1. For each tree-model (B in total), apply the model to the 'quasi test observations' ($S^b_{tr,te}$) (we note $\widehat{\tau_{i,b}}^{tr,te}$ the prediction obtained). For each tree, it gives a prediction on $(1-\beta)$ % of the sample, that is $50\% * 50\% = 25\% N_{all}$ if $\alpha = \beta = 50\%$.
2. For each observation in S_{tr}, recover a 'quasi-test' prediction at the forest level by averaging 'quasi-test' predictions on $S^b_{tr,te}$:

$$\widehat{\tau_i}^{CF_{tr,te}} = \frac{1}{\tilde{B}_i} \sum_b \widehat{\tau_{i,b}}^{tr,te}$$

   $\tilde{B}_i$ is the number of trees in which observation i was not sub-sampled, with $\tilde{B}_i$ approximately equal to $(1-\beta)B$ as a unit from S_{tr} should be subsampled in S^b_{tr} in β% of the trees in the forest on average.
3. It gives a prediction $\widehat{\tau_i}^{CF_{tr,te}}$ for $(1-\alpha)N_{all}$ observations. We compute the quartile thresholds on $\widehat{\tau_i}^{CF_{tr,te}}$ distribution : $Q^{tr,te}_j$ with $j \in 1, 2, 3$.
4. Assign each observation from S_{te} to one of the four quartile groups by comparing $\widehat{\tau_i}^{CF}$ to the three previous cut-offs.

Using these quartiles, we can now :

- Define two (binary) indicators of interest on the test sample : I_{Q1} (bottom 25% group) and I_{Q4} (top 25% group), with I_{Q1} (respectively I_{Q4}) equal to 1 if $\widehat{\tau_i}^{CF} \leq Q^{tr,te}_1$, (respectively $\widehat{\tau_i}^{CF} \geq Q^{tr,te}_3$), $\forall i \in S_{te}$.
- Compute average predicted (conditional) treatment effects for the bottom 25% group and for the top 25% group , with $\overline{\widehat{\tau}^{CF}}^{top25\%} = \frac{1}{\#\{I_{Q4}=1\}} \sum_{i \in S_{te}} \widehat{\tau_i}^{CF} * \mathbb{1}\{I_{Q4}(i) = 1\}$ and $\overline{\widehat{\tau}^{CF}}^{bot25\%} = \frac{1}{\#\{I_{Q1}=1\}} \sum_{i \in S_{te}} \widehat{\tau_i}^{CF} * \mathbb{1}\{I_{Q1}(i) = 1\}$

B.3.2 Extension : generating M causal forests on different training samples

The initial split of the sample between training S_{tr} and test sample S_{te} introduces sampling noise in the predictions, due to finite sample size. One way to limit this noise is to run M times the

over-fitting : it leads to a higher number of predictions in both tails of the distribution (sensitivity to outliers), increasing (respectively decreasing) the value of the 75% quantile (respectively 25% quantile) compared to the test sample or quasi-test sample.

causal forest procedure described above, on different initial splits of the data. It generates M-causal forests predictions $\widehat{\tau_m^{CF}}$ (and quartiles indicators $I_{Q1,m}$ and $I_{Q4,m}$). Aggregating results should reduce the sampling noise and stabilize results across different forests. This is also a way to recover a prediction for every unit of the sample.

However, aggregating the results of M forests when the test sample $S_{te,m}$ differs across simulations is not straightforward.

1. For predicted treatment effect : If M is large enough, any individual of the full sample should be at least once in a test sample $S_{te,m}$, so we can get a predicted treatment effect (CATE) for every individual of the inital sample. We define the aggregated prediction as follows :

$$\widehat{\tau_i^{MCF}} = \frac{1}{\tilde{M}_i} \sum_{\tilde{M}_i} \widehat{\tau_i^m}$$

   where m is a causal forest in which i was assigned to test sample $S_{te,m}$ and $\tilde{M}_i$ the number of forests in which observation i was assigned to the test sample.

2. For binary indicators : we compute the likelihood of being predicted in each quartile group (following the procedure described B.3.1) across the M simulations. We create dummies I_{Qi}^L, with $i = 1; 4$: the dummy is equal to one if the probability of being predicted in bottom quartile (resp. top quartile) is greater or equal to 25% (the random reference). [22]

[22] This is a 'large' definition for this indicator. One could also use a 'restricted' definition in which we units that have the same probability of being predicted in the quartile of interest as another quartile are recoded to zero (i.e. units not 'strictly' predicted to be in bottom (resp. top) quartile). The difference across these two definitions should disappear as M increases (it will be less likely to be predicted 'as many times' in a given quartile as in another quartile).

Table 13: List of baseline covariates used as features in causal forest algorithm

Variable description	Type	Variable description	Type
Individual characteristics		Nb of fixed phones	Continuous
Gender	Binary	Nb of mobile phones	Continuous
Age	Continuous	Nb of cars	Continuous
Nb of children	Continuous	***Employment***	
Live in urban area	Binary	Total nb of activities	Continuous
Nationality : Ivorian	Binary	Total nb of wage-employment activities	Continuous
Education		Total nb of self-employment activities	Continuous
Has been to school at least once	Binary	Is engaged in (at least one) casual activity	Binary
Did not complete primary school (No diploma)	Binary	Total nb of independent agricultural activities	Continuous
Has completed primary school (CEPE)	Binary	Owns a farm	Binary
Has completed middle school (BEPC)	Binary	Log of total Earnings (monthly) (log (1+x))	Continuous
Has completed secondary school (BAC or +)	Binary	Nb of days worked (last 7 days)	Continuous
Previous Vocational Training	Binary	Aspire to be wage-employed in future	Binary
Is a student	Binary	Aspire to be self-employed in future	Binary
Literate in French	Binary	Searching for a wage job	Binary
Household characteristics		Searching for an independent activity (to start)	Binary
Household size (total number of members)	Continuous	Search for a job using personal relationships	Binary
Number of rooms	Continuous	Search for a job contacting directly employers	Binary
Nb of children (<18 ans)	Continuous	Search for a job using job ads	Binary
Is head of household	Binary	Search for a job through a public agency	Binary
Is the partner of the head of household	Binary	Search for a job through a private agency	Binary
Is a children of the head of household	Binary	Search for a job by taking entrance examinations (*)	Binary
Other relationship to the head of household	Binary	Search for a job by any other mean	Binary
Nb of members who did not complete primary school	Continuous	***Savings, Expenditures and Constraints***	
Nb of members who complete primary school only	Continuous	Has Saved (last 3 months)	Binary
Nb of members who complete middle school and more	Continuous	Savings Stock (FCFA)	Continuous
Nb of members working	Continuous	Nb of savings channels	Continuous
Share of other hh members engaged in self-empl.	Binary	Has a Savings Account	Binary
Share of other hh members engaged in wage empl.	Binary	Has loans to repay (borrowed money)	Binary
Nb of friends engaged in self-empl.	Continuous	Face Constraints to repay loans	Binary
Nb of friends engaged in wage employment	Continuous	Face Constraints to access credit	Binary
Nb of family members (non hh) engaged in self-empl.	Continuous	Transportation expenditure (last 7 days)	Continuous
Nb of family members (non hh) engaged in wage-empl.	Continuous	Communication expenditure (last 7 days)	Continuous
Assets (at household level)		Nb of days with no meals (last 7 days)	Continuous
Nb of livestock	Continuous	Nb of days with leisure activities (last 4 weeks)	Continuous
Nb of poultry	Continuous	***Preferences, Personality traits, Cognitive skills***	
Nb of other farm animals	Continuous	Risk aversion level (scale 0 to 10, 0=very averse)	Continuous
Nb of plow	Continuous	Is Risk averse (based on lotteries)	Binary
Nb of field sprayer	Continuous	Preference for present (actualization rate for 1 mth)	Continuous
Nb of carts	Continuous	Personality trait - Centrality of work	Continuous
Nb of wheelbarrow	Continuous	Personality trait - Tenacity	Continuous
Nb of bicycles	Continuous	Personality trait - Desire for Achievement	Continuous
Nb of motorcycles	Continuous	Personality trait - Polychronicity (**)	Continuous
Nb of pirogues	Continuous	Personality trait - Desire for Power	Continuous
Nb of refrigerators	Continuous	Personality trait - Organization	Continuous
Nb of freezers	Continuous	Personality - Trust in others	Continuous
Nb of air conditioning units	Continuous	Personality trait - Taste for Managing people	Continuous
Nb of fans	Continuous	ZTPI Future forward score	Continuous
Nb of stoves	Continuous	ZTPI Fatalist present score	Continuous
Nb of computers	Continuous	CESD Positive Affect score	Continuous
Nb of radio stations	Continuous	Score at NV7 test (spatial vision)	Continuous
Nb of TV	Continuous	Score at 1st Dexterity test	Continuous
Nb of TV antenna	Continuous	Score at 2nd Dexterity test	Continuous
Nb of players (video, music)	Continuous	Score at Raven test (deduction)	Continuous

Note that on top of this list of covariates, as described in section B.1 we add to this set binary indicators coded one for missing values in the associated covariate.
(*) entrance examinations for public administration jobs (**) taste for handling several activities in parallel