Using Causal Machine Learning to Explore Heterogeneous Treatment Effects:

A case-study application to girls' empowerment programme in Africa

May 2021

Hertie School

Master of Public Policy

Thesis by

Ofer Dotan and Mariana Saldarriaga

Supervised by Prof. Leonardo Iacovone

Acknowledgments

We would like to express our heartfelt thanks to all of the people who assisted us in the often-arduous process of writing this thesis. Without these trusted supporters, our investigative excursion into the world of causal forests would most certainly have looked much different. First, we would like to thank Prof. Leonardo Iacovone, PhD, for his steadfast confidence in our ability to successfully conduct this research. Thank you for encouraging us to think critically, for teaching us what good research is all about and for answering both our small and large questions with seemingly unending patience. Our thanks also go to the data science lab for their support and assistance in the case of all our theoretical and code-related queries. Specifically, we would like to thank William Lowe, PhD, for his continuous and invaluable input to our project, and Olga Gasparyan and Bruno Gasparotto Ponne for their direct help with segments of our code. And, finally, our love and thanks go to our families and friends for their support.

Table of Contents

Acknowledgments 2
Acronyms 6
Executive Summary 7
1. Introduction 8
1.1. Choice of Case Study 10
1.2. Structure of the paper 11
2. Literature review 13
2.1. Causal Inference and Heterogeneous Effects Estimation 13
2.2. Heterogeneous Effect Estimation with Machine Learning 14
2.2.1. The Causal Forest 16
2.3. State of the Art: Applications in Empirical Research 16
3. Theoretical framework 20
3.1. Women's Empowerment 20
3.2. Enablers and Constraints Driving Women's Empowerment 21
3.3. Research Hypotheses 25
4. Methodology 27
4.1. Data 27
4.2. Pre-processing of the data 27
4.2.1. Operationalisation of the Outcomes 27
4.2.2. Selected Variables for the Analysis 29
4.3. Applying Causal Forest to the Data 31
4.4. Model Specification 31
4.5. Analytical Strategy 32
4.6. Model Evaluation 34
4.6.1. Assumptions 35
5. Descriptive statistics/Baseline characteristics 37
6. Results 40
6.1. The Average Treatment Effect (ATE) 40
6.2. Replications of the results using OLS 40
6.3. Assessing Heterogeneous Treatment Effects 41
6.3.1. Variation in CATE estimates 42
6.3.2. "Best linear fit" method 43
6.3.3. Assessing Heterogeneity Across Subgroups 44

6.3.4. Assessing Heterogeneity Within Subgroups 47
6.3.4.1. Expenditure on Goods and Gender Empowerment 51
6.4. Evaluation Causal Forest Fit 53
6.4.1. Assessing the Overlap Assumption 53
6.4.2. Assessing bias of the causal forest predictions 54
7. Return on Investment (ROI) 55
8. Discussion 58
8.1. Methodological Considerations 58
8.2. Substantive Results 59
9. Lessons Learned and Limitations 62
References 64
Appendix A: The Empowerment and Livelihood for Adolescents (ELA) Programme 70
Appendix B: Survey to Experts 71
Appendix C: Survey Analysis 75
Appendix D: Missing Observations Map 81
Appendix E: Tables of Selected "Important" Variables 82
Appendix F: Additional Analysis 85
Appendix G: Estimated CATES by Ntiles 87
Appendix H: Standard Errors of the Trained Causal Forest 88
Appendix I: Example Tree from the Trained Causal Forest 89
Appendix J: Return on Investment Calculations 90

List of Figures

Figure 1. Histogram of Out-of-bag CATE predictions from trained Causal Forest by Outcome 42
Figure 2. ATE within Ntiles (as defined by predicted CATE). Outcome: Expenditure on Goods [UGX] 45
Figure 3. ATE within Ntiles (as defined by predicted CATE). Outcome: Gender Empowerment 46
Figure 4. ATE within Ntiles (as defined by predicted CATE). Outcome: Control Over Body 47
Figure 5. Causal Forest Estimated CATEs. Outcome: Expenditure on Goods [UGX] 49
Figure 6. Causal Forest Estimated CATEs. Outcome: Gender Empowerment 49
Figure 7. Histogram of propensity scores by outcomes 53
Figure 8. Histogram of bias by outcomes 54

List of Tables

Table 1. Descriptions of Variables 29
Table 2. Descriptive statistics of continuous variables by treatment status at baseline 37
Table 3. Descriptive statistics of discrete variables by treatment status at baseline 38
Table 4. Results of the t-test on outcomes of interest 39
Table 5. Causal Forest Estimated Average Treatment Effects by Outcome 40
Table 6. Linear regression Estimations of Average Treatment Effects by Outcome 41
Table 7. Best Linear Fit for Expenditure causal forest model 43
Table 8. Best Linear Fit for Gender Empowerment causal forest model 43
Table 9. Best Linear Fit for Control Over Body causal forest model 43
Table 10. Pairwise ATE Ntiles differences. Outcome: Expenditure on Goods [UGX] 45
Table 11. Pairwise ATE Ntiles differences. Outcome: Gender Empowerment 46
Table 12. Linear regression Estimated CATEs. Outcome: Expenditure on Goods [UGX] 50
Table 13. Linear regression Estimated CATEs. Outcome: Gender Empowerment 51
Table 14. Return on Investment 57

Acronyms

ML	Machine Learning
ATE	Average Treatment Effect
CATE	Conditional Average Treatment Effect
RCT	Randomised Control Trial
ELA	Empowerment and Livelihood for Adolescents
CART	Classification and Regression Trees
GRF	Generalised Random Forests
IGA	Income Generating Activities
COB	Control Over Body
BLF	Best Linear Fit
BART	Bayesian Additive Regression Trees
LASSO	Least Absolute Shrinkage and Selection Operator
CI	Confidence Intervals
GEI	Gender Empowerment Index

Executive Summary

This thesis offers new evidence on the relationship between Ugandan adolescent girls' characteristics and how they modify the extent to which they benefited from an Empowerment and Livelihood for Adolescents (ELA) programme aimed at empowering them. Beyond estimating the Average Treatment Effect (ATE) of the programme, we ventured to estimate the Conditional Average Treatment Effect (CATE). Hence, with the goal of examining treatment heterogeneity, this study re-analysed data from a completed Randomised Control Trial of the ELA and applied to it a causal forest model, a machine learning (ML) method that enables causal inference.

By comparing this innovative new model with an older linear interactions model, we sought to reveal the advantages of using machine learning to credibly detect treatment heterogeneity. Specifically, the causal forest detected heterogeneity in the underlying models, finding moreover that the programme had diverging effects depending on the specific youth club the girls attended to follow the ELA programme. Our research found that older girls benefited more from participation than their younger peers and that girls from urban areas experienced more benefits from participation than girls from rural areas. However, further research is needed to determine if these signals and directions of effects found in our results, are statistically significant.

Moreover, the causal forest proved to be more flexible than the linear model as it is based on non-parametric estimation technique. It also facilitated a data-driven approach for the determination of relevant subgroups for the study of heterogeneity. On the other hand, the casual forest proved to be suboptimal in an application to dataset of approximately 5,000 observations, meaning that larger samples need to be introduced to this model, for it to function better. Hence, further research should focus on applications to large scale programme evaluations that has at their disposal large databases. Lastly, a feedback loop between the data we collect and the models we use to analyse it is also discussed, suggesting that if we are to credibly understand subgroup experience of programmes, we need to ensure that evaluations are mindful in their design for the need of large subgroup sample sizes.

1. Introduction

"Targeting" is the name given to the practice of asserting which of society's members should receive a particular social good. The targeting of social resources and protections to specific groups of people is often seen as being unavoidable, especially given that societies operate with limited resources, and the practice has often been employed as a means to elevating poverty (Schuck & Zeckhauser, 2005). The justifications for targeting are many, but at their core lies the intention of maximising social welfare by increasing efficiency (Coady et al., 2004). In addition, various notions of fairness mean that targeting is seen as a morally justifiable practice due to its redistributive impact (Kelly, 1992; Schuck & Zeckhauser, 2005).

It only follows, then, that in order to be able to better target populations, we first need to learn what works and for whom. One way to do this is to assess the impact of interventions (treatments) by conducting a randomised control trial (RCT) (Duflo, 2020; Meager, 2019). The rationale behind such an investigation is the following: although we *cannot* measure what would have happened to a specific individual if she had not undergone a programme aimed at improving her situation, we *can* leverage statistical computation to simulate this reality and estimate what is, on average, the overall impact of the programme, that is, the average treatment effect (ATE) (Pearl & Mackenzie, 2018). This practice enables us to assess whether a specific policy intervention is helpful in achieving set goals and also whether it is cost-effective (Duflo, 2020; Meager, 2019). The ability to establish causal relations between the implementation of treatments and their outcomes can improve the policymaking process at several points, including the screening of target populations and fostering more effective decision-making with regard to programme allocation (Duflo, 2020).

In real life, however, individuals respond differently to treatment, meaning that there is more to discover than the mere ATE of an intervention. Some of this heterogeneity in treatment effects comes from unobservable factors such as an individuals' motivation to benefit from a given programme offered to them. However, observable characteristics of individuals can also explain the heterogeneous effects of treatment. As unobservable characteristics are difficult to measure, in the remainder of the thesis, when we speak of heterogeneity, we are referring to a non-random variability that affects the direction and size of treatment effects on individuals within a studied population (Osler, 1939; Varadhan & Seeger, 2013).

The prospects of learning more about the ways in which different individuals respond differently to treatment and being able to predict these effects are important to decision-makers for the following two reasons: first, this information could help them understand how a programme can be improved and how to further develop its cost-effectiveness by focusing on elements that drive positive impacts (Chang et al., 2020). Second, it could facilitate the more effective allocation of the programme in the future, for example, by offering it to those subgroups that are likely to benefit most from it (Athey et al., 2016; Bhattacharya & Dupas, 2012; Dehejia, 2005; Frölich & Melly, 2013; Hirano & Jack R. Porter, 2009; List et al., 2019; Manski, 2004; Wager & Athey, 2018). This type of inquiry aims to estimate the conditional average treatment effect (CATE), thereby analysing which subgroups benefit more than others, depending on their observable characteristics.

Treatment heterogeneity is often hard to detect due to noise in the data, meaning that factors such as measurement errors and data collection practices compromise the ability of the data to represent reality in a credible way. Traditional statistical methods generally endeavour to detect heterogeneity by testing the interaction of baseline covariates, namely information on the individuals comprising the sample population, with a treatment indicator. The trouble with this approach is known as the multiple hypothesis testing problem, which results in skewed findings (Davis & Heller, 2018). A linear investigation such as this may also fall short of detecting non-linear variation in treatment effects.

The good news is that causal machine learning (ML) methods have become a tool to estimate heterogeneity along a potentially larger number of covariates (Knaus et al., 2018). These models are based on non-parametric analysis, thereby allowing for a more flexible, non-linear estimation of heterogeneity in high-dimensional spaces (Athey & Imbens, 2015b; Wager & Athey, 2018). For this reason, they are better able to mine the data in order to detect relations between indicators and outcomes. Therefore, in this thesis, we developed a ML causal inference model using a method by Wager and Athey (2018) called "causal forest". This algorithm is specifically designed to make it possible for researchers to determine treatment heterogeneity, thereby enabling policy learning by adding nuance to the understanding interventions' effects on subgroups. At the moment, there are still relatively few applications of this method being carried out. Our thesis seeks to rectify this situation by providing further evidence about this method's capacity to reveal more accurate information on what these programmes do and for whom (Baiardi & Naghi, 2021; Davis & Heller,

2018).

1.1. Choice of Case Study

This thesis takes a case study approach in that we applied the causal forest technique to a dataset taken from a completed randomised evaluation of an Empowerment and Livelihood for Adolescents (ELA) programme set up to empower adolescent girls in Uganda (Bandiera et al., 2014).[1] This intervention sought to address the high poverty, child marriage, early pregnancy and low education rates among Ugandan girls (Bantebya et al., 2014). This social intervention was carried out in Uganda in 2014 and provided adolescent girls aged between 14 and 28 ELA vocational skills training and informational sessions on marriage, reproduction and sex, all directed towards empowering them. Please note that for the sake of this analysis we have chosen to define empowerment as the enhancement of an individual's choice and ability to shape their own life (Do & Kurimoto, 2012).

The importance of female empowerment is reflected in its inclusion in Goal 5 of the United Nation's Sustainable Development Agenda, which aims to "achieve gender equality and empower all women and girls" (Duflo et al., 2011; Razavi, 2016; United Nations, 2021). Unfortunately, women tend to remain in disadvantaged positions as compared to men, especially with regard to economic participation and political and public decision-making processes (Vollmer et al., 2017). One of the most relevant aspects is the extent to which women are able to make their own sexual and reproductive health decisions, which may also determine their economic participation (Do & Kurimoto, 2012). Women's empowerment is therefore often associated with a transformation of power relations between women and men, especially in the home (Duflo, 2012; Kabeer, 1991; Malhotra et al., 2002; Panda, 2000; Young, 1993).

Accordingly, vast literature exists on the provision of skills programmes that aim to combat poverty and inequality by developing individuals' capacities in three main domains: 1) Cognitive skills including but not limited to the promotion of individuals' ability to reason, comprehend complex ideas and adapt to their environment; 2) Socio-emotional skills facilitating individuals improving interpersonal situations, work in teams and develop self-control, autonomy and self-reliance, and; 3) Technical skills to jump-start economic emancipation (Adams, 2008; Bruce &

[1] For a more elaborate description of the ELA programme, see Appendix A.

Hallman, 2008; Coles et al., 2015; Panda, 2000; Vollmer et al., 2017). Similarly, the ELA addresses three core dimensions of female empowerment: social empowerment, economic empowerment and control over their bodies (Bandiera et al., 2014).

The vocational training programme aimed at improving the following in the treated adolescent girls: vocational tools to increase skills that allow economic empowerment and financial independence, and life skills to increase social empowerment of the girls. Here, social empowerment refers to helping women make informed choices about sex, reproduction and having a partner. Livelihood assets, cultural barriers and background characteristics of women are key determinants of livelihood outcomes (Sheheli, 2012). Therefore, as in the original paper, we specifically explored the effect of the programme on economic and social empowerment dimensions.

Because the ELA encapsulates the main aspects of the general approach to how gender inequality should be tackled through policy and proves to produce a significant impact, we see it as having significant potential to be transferred to other contexts. With this analysis, we wished to further strengthen girls' empowerment by deepening the understanding of who benefited from the ELA. Understanding heterogeneity can help address and improve the programme's design and inform the targeting of adolescent girls for such interventions in the future. In other words, predicting the heterogeneous effects of this intervention can assist policymakers in improving their programmes and provide them with information that can help scaling them in an optimal manner. Our ultimate goal here is to contribute to women's empowerment through the facilitation of cost- effective programme allocation and better implementation. And, finally, as a way of substantiating our methodology and better understanding our results, we also approached experts in impact evaluation and development to provide us with their interpretations regarding the potential sources of heterogeneity in the data we explore.

1.2. Structure of the paper

Our study followed two lines of investigation: first, it carried out a methodological inquiry into the capacity of ML – and especially the causal forest algorithm – to credibly detect treatment heterogeneity and, second, it undertook an analysis designed to uncover which Ugandan adolescent girls benefited the most from an ELA programme. Our goal is to strengthen

policymakers' understanding of the diverse effects the intervention may have, depending on participants' characteristics at the beginning of the programme.

Following the introduction, we provide the reader with a literature review covering development in the fields of causal inference using ML, specifically for estimating heterogeneous treatment effects. Drawing on this and the survey results, we construct a theoretical framework and deduct the hypotheses as to the anticipated sources of heterogeneity. We then explain our methodological approach to the analysis and provide an interpretation of the results we obtained. Descriptive statistics are then presented. Finally, we describe our results and discuss the implications of our findings.

2. Literature review

To contextualise our study, the following review briefly considers current themes regarding causal inference with ML and specifically the estimation of heterogeneous treatment effects of randomised evaluations and observational studies.

2.1. Causal Inference and Heterogeneous Effects Estimation

At the heart of any causal estimation lies the problem that we cannot observe the counterfactual, that is, what would have happened but for the treatment. The 'potential outcomes framework', first introduced by Jerzy Neyman (1923) and later established by Donald Rubin (1974), formalised this problem by asserting that we can statistically establish causality by positing to each observation two potential outcomes: (i) one observed which is treated and (ii) one observed if the observation was not treated. Then, and if randomisation is credible, the ATE is defined as the mean difference between treated and non-treated observations (Imbens & Rubin, 2015).

In addition to estimating the ATE of policies, scholars and policymakers are also interested in detecting the impact heterogeneity of interventions, that is, whether the treatment effect depends on the direction and magnitude of treatment effects for individuals within a population (Varadhan & Seeger, 2013). One of the most common analytic approaches for examining different treatment effects is subgroup analysis. This means that instead of merely estimating the ATE, the intention is to estimate the CATE. The term conditional here implies that the effects depend on and vary across baseline characteristics (Kuenzel, 2019). Information about how different subgroups benefit from a programme can help assess if the programme is generalisable to a population with specific characteristics or to better understand the driving mechanisms behind a particular programme's effects. Specifically, estimating heterogeneous treatment effects lies at the heart of many modern statistical problems**.** In fact, heterogeneous treatment effect estimation is closely related to the problem of optimal policy estimation and evaluation (Athey & Wager, 2019). Ultimately, such estimation may help assign future subjects to what is predicted to be the most effective treatment for them, which translates into more efficient cost-benefit targeting transfers to beneficiaries (Athey & Imbens, 2015b, 2019; Duflo, 2020).

2.2. Heterogeneous Effect Estimation with Machine Learning

Standard ML prediction models are designed and optimised for predicting the outcome in a test sample, which means that the model is selected by optimising the precision of the fit on the test set (Baiardi & Naghi, 2020). In empirical economic research, estimating a causal effect is more relevant. Thus, the fit and the predicting accuracy are sacrificed in order to know more about the fundamental relationships relevant for causal inference and policymakers' decisions (Athey & Imbens, 2019). In this sense, researchers have recently proposed ML tools for causal inference (Athey et al., 2016; Chernozhukov, Chetverikov, et al., 2018; Chernozhukov, Demirer, et al., 2018; Wager & Athey, 2018), and to address the problem of predicting subgroup effects of a treatment where there are many attributes to each unit and where the functional form between these attributes and the treatment effects is not determined (Athey et al., 2016). In short, these models aim to determine which subgroups have diverging ATE. Causal ML, and specifically supervised learning, which is the focus of our research, is a computational tool that can predict outcomes depending on the inputs, namely, what would y be if x (Athey & Imbens, 2015b, 2015a, 2019; Chernozhukov, Demirer, et al., 2018; Knaus et al., 2018; Mullainathan & Spiess, 2017; Wu et al., 2008). Such analysis assumes there is reasonably large data from an experiment where treatment was assigned randomly based on observable characteristics of the target population, or when exogenous variation can be simulated using an instrumental variable. Indeed, data on subgroups has to be big enough for the model to detect heterogeneity.

Several methods in supervised machine learning[2] have been developed to predict outcomes as a function of covariates in similar contexts (Athey et al., 2016). In general, as Athey and Imbens (2015b) explain, the analytical approach has been to construct a model of the relationship between treatment effects and attributes while penalising high model complexity, selected by cross-validation[3] methods to ensure high performance of the model without "overfitting". One promising strand of literature demonstrates that modified ML algorithms based on regression trees can be used to predict an outcome conditional on the indicators (Athey & Imbens, 2015a; Breiman et al., 2017). The idea of building regression trees is to split the sample based on feature values, and

[2] Methods like regression trees, random forests, LASSO and support vector machines.

[3] This method "compares a set of models with varying values of the complexity penalty and selects the value of complexity parameter for which out-of-sample predictions best match the data using a criterion such as mean squared error (MSE)" (Athey & Imbens, 2018, p 2).

given a set of splits,[4] average the within each leaf estimates, namely the conditional expectation inside every tree leaf (Athey et al., 2016). Such algorithms that build the regression trees produce a partition of the sample according to covariates, which means all units in the partition receive the same prediction (Athey & Imbens, 2015a). This means that regression trees do not directly estimate conditional average treatment effects.

There has also been interest in exploring the application of random forests. This method uses local approximations of conditional mean functions (Athey et al., 2019; Breiman et al., 2017). One relevant critique of this method is that it does not manage complicated models well, and estimates seem to be poor when the data is too small (Athey, 2020). Another challenge is how to form subgroups correctly from a theoretical standpoint, considering that this is a theoretical driven analysis and there are many ways in which subgroups can be categorised (Athey et al., 2018). Despite the complexity, it is possible to use empirical data to discover whether there is any relevant heterogeneity in treatment effect by covariates after the end of the experiment (Chernozhukov, Demirer, et al., 2018).

In general, one advantage of causal ML models is that they can control for confounders because we are not aiming at identifying the exact effects of one parameter on another, but rather to see what is the most probable value of an outcome given a set of predictors. Accordingly, ML is handy relative to conventional methods used to model how outcomes vary depending on many attributes (Athey, 2015; Baiardi & Naghi, 2021; Davis & Heller, 2018; Wager & Athey, 2018). However, the biggest challenge of applying ML tools to estimate general heterogeneous causal effects is obtaining valid inference. Because of the lack of strong assumptions, in high dimensional settings, generic ML tools may not produce consistent estimators of the CATEs, i.e., the difference in the expected potential outcomes between treated and control states conditional on covariates (Chernozhukov, Demirer, et al., 2018).

Another advantage of causal ML methods is that they employ regularisation to reduce variances and exchange regularisation bias with overfitting practice which produces better model performance (Chernozhukov, Demirer, et al., 2018). Regularisation is specifically relevant for random forest combined with a 'doubly robust' outcome modification. In random forests

[4] The complexity of the model can be penalised here. It is chosen a constant time the number of splits/leaves, with the constant chosen through out-of-sample cross-validation (Athey & Imbens, 2015a).

algorithms, the predictions are calculated by a weighted average of close, "nearby" observations combined with a data-driven approach for discerning which of the close-by observations should be weighted more heavily (Athey et al., 2019, 2017). These algorithms have displayed strong prediction and classification capacities. However, in order to facilitate causal interpretation, they need to be adapted. For that, Wager and Athey (2017) provide a set of conditions under which predictions made by random forests are both asymptotically unbiased and Gaussian, and therefore allowing statistical inference. Specifically, establishing credible confidence intervals is essential to fostering decision-making based on predictions. Here, asymptotic normality[5] results are of particular importance.

2.2.1. The Causal Forest

A special case of random forest was recently developed by Susan Athey, Stefan Wager and Julie Tibshirani and presented in the General Random Forests (GRF) package: the causal forest[6] (Athey et al., 2016, 2018; Wager & Athey, 2018). This algorithm modifies a specific ML estimator from classification and regression trees (CARTs), targeted in the prediction of the outcomes (Breiman et al., 2017)- to focus on the estimation of the individual average treatment effect (IATE) which is the newest sophisticated feature of the causal forest (Knaus et al., 2018). It is also designed to detect the maximum heterogeneity in treatment effects, as it splits the sample along the gradient of the mean difference with the pseudo-outcomes (Buchwald, 2019). As it is still in its early phase, one evident disadvantage is that this algorithm is prone to overfitting and can consequently produce selection bias (Knaus et al., 2018). Local centering could potentially partly correct selection bias because weighted estimators of the causal forest have the potential to be robust to confounding, but more applications are needed to be sure (Athey et al., 2018).

2.3. State of the Art: Applications in Empirical Research

Despite these advances, research on the estimation of any relevant heterogeneity with ML is still underdeveloped. Applications are mostly available in the field of medicine (Kent et al., 2018; Powers et al., 2018)[7], but rather absent when it comes to estimating heterogeneous effects of social

[5] Meaning how estimators perform as sample size gets close to infinity.
[6] Causal forest is not a generic approach, but a one estimator specific approach. A generic approach splits the causal estimation problem into several prediction problems and combines many supervised machine learning estimators (Knaus et al., 2018).
[7] See also Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., ... & Steyerberg, E. W. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of internal medicine*, *172*(1), 3545. And,

policies and development programmes (Athey & Imbens, 2019). Most researchers and practitioners have no proper guidance about which methods are suitable for their applications. Recently, and particularly in the field of econometrics and policy evaluation, different ML methods are being explored to answer different sorts of heterogeneity questions: to identify subgroups and treatment effect for subgroups (Athey et al., 2016; Carbo-Valverde et al., 2020); to test for heterogeneity across all covariates (List et al., 2019); to explore heterogeneity for the robustness of the model specification (Athey & Imbens, 2015a); to personalise estimates (Wager & Athey, 2018), and; to estimate semi-parametric models for optimal policies (Bhattacharya & Dupas, 2012; Dehejia, 2005; Hirano & Jack R. Porter, 2009; Manski, 2004).

Indeed, there already exist some applications of various supervised ML tools to predict heterogeneous causal effects, and there is a growing body of reasonably successful attempts (Athey & Imbens, 2015a, 2015c; Athey & Wager, 2019; Battioui et al., 2014; Carvalho et al., 2019; Foster et al., 2011; Imai & Ratkovic, 2013; Wager & Athey, 2018). In 2018, a simulation data of an observational study containing treatment effect heterogeneity was provided to leading researchers who were asked to analyse it with the goal of testing treatment heterogeneity hypothesis (Carvalho et al., 2019). The different estimates were later compared, showing that while all researchers reached very similar estimates of the ATE, there were diverging methods to estimate treatment heterogeneity, namely the CATE, and hence, diverging estimates (Carvalho et al., 2019). Some researchers analysed formally theoretical pre-defined subgroups to search for heterogeneity (Luo et al., 2019; Panigrahi & Taylor, 2018; Zhao et al., 2017) applied a Bayesian Additive Regression Trees (BART) algorithm that treats all variation in the same way (Carnegie et al., 2019). Moving away from these workshop results, BART is shown to credibly detect treatment effect moderation in various other observational data settings (Chipman et al., 2007; Green & Kern, 2012; Hill, 2011).

For the most part, researchers apply approaches that sit between the agnostic evaluation of heterogeneity and theoretical driven analysis of heterogeneity. That is to say their models use a data-driven selection method that uses sample splits to define which subgroups may be the source of heterogeneity, and then they formally test if that is indeed the case. This method is called

---

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomised program evaluation. *The Annals of Applied Statistics*, *7*(1), 443-470.

recursive partitioning and it is present in earlier work which proposed using regression interaction trees from the random forest family (Su, Tsai, Wang, Nickerson & Li, 2009). Perhaps most prominently today is the causal forest that enables the modelling of interactions in a high dimensionality space by constructing many regression trees and then averaging their predictions (Wager & Athey, 2018). For example, Davis and Heller (2020) used a causal forest to explore possible heterogeneity in treatment effects in observational data produced to evaluate the benefits of youth employment interventions. They found significant heterogeneity of youth employment outcomes; however, their estimates were not sufficiently precise. They were therefore cautious to make definitive claims regarding their findings.

Another econometrics application that successfully modelled causal forests used Swiss administrative social security data to simulate evaluation of job search programmes and then detect treatment heterogeneity on the outcome of employment in the subsequent 3 years (Knaus et al., 2018). Luo et al. (2019) applied causal forest to commercial settings and discovered that the more high-spending customers are more sensitive to scarcity incentive than to price incentives compared to the low-spending customers. Using causal forest, Elek and Bíró (2021) also successfully detected regional differences in diabetes prevalence in Europe depending on individuals' risk factors (i.e. socio-economics, demographic, health, lifestyle).

Carbo-Valverde et al. (2020) explored variation in bank customers' digitalisation process using causal forests and random forests. Both methods provided for credible results, while random forests produced higher accuracy. Adapting the causal tree structure, Stoffi & Gnecco (2020) established a reliable data-driven identification strategy of heterogeneous treatment effect in an instrumental variable setting. Other ML tools have also recently improved and are used for the analysis of heterogeneity, such as the Least Absolute Shrinkage and Selection Operator (LASSO) method. This method attempts global[8] approximation of conditional mean features (Knaus et al., 2018; Tibshirani, 2011), and it is combined with a covariate modification and efficiency augmentation (Chen et al., 2017; Qian & Murphy, 2011; Tian et al., 2014). Recently, Nie and Wager (2017) also tried to combine LASSO with R-learning, which in "the first step learns an approximation for the oracle objective, and the second step optimises it" (p. 3). However, forest-

[8] "By 'local' we mean that each point of the conditional mean function is approximated by the (weighted) average of neighbouring observations. By 'global' we mean the attempt to approximate the conditional mean by a flexible functional fitted to all data simultaneously" (Knaus et al., 2018, p. 8).

based estimators might actually be better performing than LASSO estimators (Knaus et al., 2018). In fact, LASSO estimators tend to be more unstable than forests since they are less likely to achieve a normal distribution. Finally, another set of relatively successful applications employ neural nets for the purpose of uncovering heterogeneity across features (Johansson et al., 2016; Kuenzel, 2019; Schwab et al., 2018).

3. Theoretical framework

Drawing on the literature review, we decided to address the analysis of heterogeneous effects by developing a causal forest application. The decision to explore heterogeneity with causal forest and not other ML models is both conceptual and pragmatic. Causal forest seems to be one of the most credible tools currently available with which to establish a valid estimation and inference of the CATEs and thereby determine heterogeneous effects. One simple reason is that this algorithm was created for this specific aim while other methods are just applied to the problem but were not catered for it specifically (Knaus et al., 2018; Lechner, 2018; Wager & Athey, 2018). Apart from that, due to the limited scope of this study, it was beyond our capacity to explore different ML models.

In the following section, in order to assert how to apply the model to the context of our case study, we consider the direct and indirect drivers that seem to affect the magnitude and direction of gender empowerment interventions effects for women in vulnerable communities. To establish the analytical approach, we drew on literature, available empirical studies and results from the survey we distributed to experts in the field. Altogether, the framework lays the foundation upon which to develop our hypotheses.

3.1. Women's Empowerment

The original concept of women's empowerment comes from the acknowledgment of inequalities between men and women in the political and economic realm, and the subordinated role of women in the family and in the community (Ngezahoguhora, 2013). Although women's empowerment is a multi-dimensional concept (Ibanez et al., 2017; Ngezahoguhora, 2013), it is broadly seen as a solution to remedy structural gender gaps (Cornwall, 2016; Montanari & Bergh, 2019). As mentioned before and following J-Pal recommendations on how to measure empowerment, we primarily use Naila Kabeer's definition of empowerment as the ability to make strategic choices (Glennerster et al., 2018; Kabeer, 1991; Malhotra et al., 2002). Moreover, this is understood as people's freedom and opportunities to improve their well-being (Ibanez et al., 2017; Kabeer, 1991; Ngezahoguhora, 2013; Sen, 1999).

The strategic life choices are based on resources, agencies, and achievements (Glennerster et al., 2018). These aspects are interconnected and depend on each other, so that empowerment can also

be related to the process of increasing personal and interpersonal power to improve their circumstances in social and private domains (Ngezahoguhora, 2013). Interpersonal social relationships and different roles in society, the family and the labour market can affect the financial and skills resources of women. Empowerment also refers to the ability of women to increase their own self-reliance, to value their independent right to make choices and to control resources (Malhotra et al., 2002). In general, gender conceptual theory refers to women's empowerment as a process which leads women to have opportunities and increase their autonomy in economic and social decision-making. This also means women's empowerment can be transformed through action as well as in situations that strengthen them through the provision skills and self-reliance (Ngezahoguhora, 2013).

3.2. Enablers and Constraints Driving Women's Empowerment

Before conducting the analysis, it is important to understand which of the adolescent girls' characteristics might affect the magnitude of the programme's impact. Beyond theory and empirical results, we asked 26 experts in the fields of impact evaluation and development programmes to estimate which characteristics of the adolescent girls are likely to impact the programme success.[9] In turn, this would enable us to deduct our hypotheses regarding the various effects we might find in the economic and social empowerment of the treated adolescent girls. Further, it would enable us to discuss our results against practitioners' opinions.

Some direct factors that enable or constrain women's economic empowerment are the following: education, access to quality and decent paid work, access to property and financial services, and skills development and training (Hunt & Samman, 2016). The baseline education level and economic status of women seems to affect the outcome on women's economic empowerment improved by the intervention. One of the reasons this is possible is because poor women consider time to be a most valuable resource and they often face competing claims as to what they should succeed in accomplishing in a given day. Thus, being poor could restrain girls from attending school or from engaging in income-generating activities, for example, due to pressure to engage in domestic duties (Coles et al., 2015). This reality is aligned with the fact that girls usually drop out earlier than their male counterparts. Somewhat regardless of economic resources, school

[9] See Appendix B to read the distributed survey and Appendix C to read the analysis of collected answers.

attendance is dependent on whether schools have adequate toilet facilities as well as whether girls are exposed to sexual harassment or to the threat of harassment on school premises (Coles et al., 2015; UNESCO, 2014).

Experts' opinions diverged regarding the effect of girl's economic status: on one hand, 28% anticipated a positive treatment effect on wealthy girls and, on the other hand, the rest of the sample predicted that more wealth would have no effect on the programme's outcomes.[10] In addition, approximately one third of the experts thought that women who were satisfied with their income would have positive effects on the treatment relative to women unsatisfied with their economic gains.[11] Regarding potential educational moderators of treatment effects in gender programmes, the group of experts we surveyed agreed that girls who are enrolled in school would have a positive impact on the programme effectiveness (40%). However, only 30% of experts anticipated that rarely attending school would have either positive or no effect on programme outcomes.[12]

Some empirical studies have also found that income gained in a month by women depends on years of education, age and family size; education and age have a positive impact, although family size has a negative impact (Okumo, 1992; Sheheli, 2012). Obayelu and Chime's (2020) results indicate that while women's age and their husband's education level are negatively related to women's empowerment, her education and status as head of the household were positively directly related to it.

Marriage or partnership may undermine adolescent girls' accumulation of education and labour market participation. In fact, delays in marriage are related to greater educational achievement (Klugman et al., 2014). Further, this reality can result in a more limited effect of the intervention to kick-start human capital building, and yet, it can improve the lives of the women within their households, for example, by way of better marriage quality (Baird et al., 2014; Bandiera et al., 2014). When marriage is critical to girls' success in society, the likelihood of promiscuity, sexually transmitted infections, abduction, single parenthood and abandonment by partners increases significantly (Coles et al., 2015). One third of the experts surveyed agreed that being married can hinder the treatment effects (30%).[13] However, the opinion of experts shows diverging views in

[10] See Appendix C, Figure C. 3.
[11] See Appendix C, Figure C. 4.
[12] See Appendix C, Figure C. 5.
[13] See Appendix C, Figure C. 6.

this matter: 20% thought being married could result in a higher positive effect of the treatment, and 24% expressed uncertainty with regard to what the effect would be. Some empirical studies have found marital status does not affect the variation in household income (Okumo, 1992; Sheheli, 2012).

The effects of interventions in women's economic empowerment with children or without children seem to be homogeneous. The survey findings showed that 30% of the experts answered that having children would not have any effect on programme outcomes.[14] Nevertheless, research into gender gaps in the labour market shows that the arrival of children has a significant impact on that gap in the long-run in comparison to men (Kleven et al., 2019).

The theoretical discussion on girls' age focuses on the fact that while in childhood girls are dependent on their parents, in adolescence, this slowly comes to an end, meaning that the girls either manage to gain income themselves or they become dependent on men (Bandiera et al., 2014; Obayelu & Chime, 2020; Okumo, 1992; Sheheli, 2012). This fact makes age an important factor when thinking about interventions to empower women, as a lack of trajectory in the labour market might reduce the likelihood girls will invest in their human capital. Thus, age plays a crucial role in the realm of economic empowerment and family planning, which makes interventions to boost human capital more likely to succeed for younger adolescent girls (Bandiera et al., 2014). Also, younger girls sometimes exhibit very little control over their body, which makes them vulnerable to sexual exploitation by their teachers. This reality then leads these girls to attend school and activities less which may also hinder their progress (Bandiera et al., 2014). Ultimately, the experts do not seem to agree on the effect of age in the treatment effects. One third of them thought that relatively older adolescent girls are likely to benefit more from the programme. However, here too we see that 28% of experts anticipated that age would not change the treatment effect.[15]

Rural women also face more challenges in participating in income generating activities. Insufficient economic resources and skills, the lack of cooperation of their partners and family members, various religious values and social insecurity can be causes of the lower ability of rural women to pursue income-generating activities (Sheheli, 2012). Obayelu and Chime (2020) found that the main cause was the large percentage of rural women in households with traditional gender

[14] See Appendix C, Figure C. 7.
[15] See Appendix C, Figure C. 8.

norms. According to Sheheli (2012): "rural women were always trying to improve their livelihood status and played a crucial role in supplementing the family income, but patriarchal domination coupled with restricted finances marginalises their livelihood status in the society" (p. 180). This is contrary to what experts in the survey predicted, namely that girls from rural areas would have more positive treatment effects (33%), and that the factor of whether girls come from rural or urban localities would not impact programme effects at all (29%).[16]

There are also underlying factors such as labour market characteristics, gender norms and discriminatory social norms that can affect economic empowerment (Hunt & Samman, 2016). Social barriers to female empowerment, captured by variables like the programme location, the stated identities and the perceptions of the role of women in society, are also taken into consideration when determining the heterogeneous effects of the interventions for different subgroups (Ibanez et al., 2017). Glennerster et al. (2018) posited that the process of empowerment is also shaped by and interacts with the cultural, social, political and economic norms and institutions that define an individual's opportunities in a specific context. For example, traditional attitudes toward gender roles in the private sphere, understood as the domestic roles of women and men in households or families, might limit the success of empowerment interventions by not allowing women's ability to translate resources into agency and self-agency successfully. It is also still the case that male roles are given greater social values than female roles, since men are seen as the family's financial provider and women as homemakers and caretakers of children and elders. One third of the experts surveyed perceived that holding traditional values can hinder the treatment effects of the programme.[17]

As previously discussed, most of the literature and empirical studies relate to the drivers and effects of promoting women's economic empowerment. Nevertheless, few studies mention drivers for women's social empowerment. According to Klugman et al. (2014), drivers and determinants of women's voice and agency are related to their endowments, for example, their health, education, assets and economic opportunities. Gender norms surrounding women's agency can also limit the impact of many interventions that aim to support and empower women (Chang et al., 2020). Traditional gender roles relative to childcare and housework are limiting (Klugman et al., 2014).

[16] See Appendix C, Figure C. 9.
[17] On traditional attitudes towards gender roles, one of every five experts surveyed said holding traditional values also could result in a more positive effect of the treatment. See Appendix C, Figure C. 10.

When employed, women have double work hours, because they typically remain the head of domestic tasks, childcare and elderly care in the home. In addition, social gender norms can limit women's mobility and ability to network, as well as restrict their representation in important institutions. In this sense, Chang et al. (2020), concluded that in order to improve women's agency, gender norms should also be addressed, rather than placing the focus solely on achieving increased access to financial resources.

3.3. Research Hypotheses

Drawing on the literature review and the theoretical framework, we devised the following hypotheses about the heterogeneity of treatment effects in economic and social empowerment dimensions.

In general, we assume that the causal forest will be better compared to the analysis in the original paper in detecting heterogeneous treatment effects.[18] However, we anticipate that some of the estimates will not be very precise, specifically when exploring heterogeneity across subgroups.

As stated by various empirical results and the opinion of the experts, we expect that in all three outcomes, adolescent girls from rural villages will have lower effects compared to girls from more urban villages.[19] Moreover, we also expect to see diverging treatment effects depending on the implementation of the programme made by the BRAC club branches' personnel in the participating communities (Cornwall, 2016).

With regard to age, we expect to find different treatment effects across age groups, because experts predicted age to have no effect, while theory indicates higher effects for older girls in relation to women's economic empowerment and lower effects for younger girls in relation to COB empowerment as part of social empowerment. However, we are unsure what the direction of these effects will be in each of the three outcomes we analyse.

[18] In Bandiera et al. (2014), they presented ITT estimates on the two summary indices, IGA and COB along the dimensions: rural, rich and age.

[19] This is contrary to what approximately one third of what experts thought, namely that rural girls would benefit more from the program. One possible explanation to this discrepancy is that experts, working in the field, have representativeness bias hindering them from properly asserting the probability of an outcome. Otherwise, it can also be the case that when answering the questions, experts had in their minds, rural girls who, as a whole, depending on other characteristics, would have benefited more. Lastly, it might be the case that the benefits experts were referring to are outcomes that have not yet materialised by the time of the follow-up survey (Gugerty & Karlan, 2018).

Based on the theoretical framework, we expect to find no significant different treatment effects in women with children and without children. However, we anticipate that the treatment effects for girls with partners will hinder in comparison to single participants.

We predict that girls who invest more time in education will have better treatment effects, as education is one positive driver of economic and social empowerment. We also suspect that higher economic status and income satisfaction will have higher positive treatment effects relative to girls with low financial capacity and income satisfaction.

Finally, we anticipate that adolescent girls who were more empowered before the programme, in particular in relation to social gender norms and their agency and voice, will have a higher treatment effect, because the programme is likely to strengthen their capacity to make informed and strategic life choices.

4. Methodology

4.1. Data

We made use of cross-sectional data from an RCT conducted in Uganda to train the analytical model aimed at predicting heterogeneous effects. The dataset is available online at the open-source platform of the American Economic Association (Bandiera et al., 2020). It has a relatively large sample size with 2,002 observations in the control group and 3,964 in the treatment group. Furthermore, the dataset, which is based on a survey design, contains a variety of covariates that can be exploited to determine heterogeneous effects. Specifically, information on the status of the adolescent women (i.e., their age, marital status and socio-economic background), their market participation and financial literacy, measures of economic and social empowerment and their subjective views regarding family roles before and after the intervention (Bandiera et al., 2014). At follow-up, there were 4,888 observations.

4.2. Pre-processing of the data

We started the pre-processing by running the STATA creation file[20] made by the researchers in the original experiment. This code generates the indexes across the different time points (baseline, end-line and follow-up) and standardises other variables for comparability of results.

4.2.1. Operationalisation of the Outcomes

To operationalise economic and social empowerment, we proceeded in the following manner. Aligned with Bandiera, et al. (2014), we chose COB and the Gender Empowerment Index (GEI) as outcomes to measure social empowerment. The COB index measures reproductive health knowledge and whether girls have the ability to decide to engage in sexual acts.[21] The GEI variable includes questions about roles in labour, education and the household.[22] This index variable indicates opinions toward gender roles in education, labour market and household chore tasks. Based on our findings on how to measure women's empowerment, we believe this is a good variable that indicates changes in women's opinions that will enhance the ability to make informed and strategic choices when related to creating a family, studies and their professional life.

[20] See Bandiera et al. (2020) [code file] "creation.do".
[21] See Table 1 for outcome description
[22] See Table 1 for outcome description.

Finally, to assess the economic empowerment in the treated girls, we chose the variable Expenditure on Goods that measures monthly consumption expenditure on a basket of goods.[23] Although gender and development policymakers and practitioners see control of assets and freedom to earn income as crucial elements in the 'empowerment' (Ngezahoguhora, 2013), most of the adolescent girls who participated in the programme are not employed and hence do not receive an income.[24] We therefore use expenditures as a proxy for the economic situation the girls had, since it indicates the consumption of the adolescent girls households.[25] Based on this outcome, we would also be able to calculate the return on investment the programme may have, and thus to provide a monetary value of targeting.

All three variables we choose as outcomes were measured both in the end-line survey and in follow-up. For this analysis, we took only the follow-up measurement, as we were interested in how the outcomes of the intervention have consolidated over time. For each causal forest analysis of an outcome, we exclude missing observations in the Y outcome matrix.[26] This step brings us to 4,854 observations for COB and Empowerment, while Expenditure on Goods is left with 4,814 observations.

After these steps, the data still contains 7% of missing observations.[27] In order to proceed with the analysis including the missing observations, we assumed that the missing observations are missing at random (MAR), meaning that there might be a structured difference in observed and missing values, however this difference can be explained by the observed sample (Bhaskaran & Smeeth, 2014).[28]

The GRF, which is the causal forest R package, makes it possible to retain missing observations in the X matrix, however not in the outcome Y matrix. We therefore choose only complete cases for the three outcomes we analysed, reaching a final sample of 5,188 observations. Keeping missings in the X matrix may increase the credibility of the model, because it eliminates the need

[23] See Table 1 for variable description.
[24] At follow up only 546 adolescent girls who received the treatment had some income compared to 2729. In control, only 167 received an income and 1405 did not report any gains.
[25] Monthly expenditure is a proxy for earning as in low-income settings it is hard to measure self-employment activities (Bandiera et al., 2014).
[26] For the linear replication models, we used the complete data set, including 4,888 observations.
[27] See Appendix D, Figure D. 1.
[28] This assumption differs from the assumption that the data is missing completely at random (MCAR), meaning that data distribution of observed and unobserved values would be similar to one another.

to impute values and generate artificial observations by allowing the model itself to determine whether a missing observation is informative and thus predictive of the outcome (Athey et al., 2019).

4.2.2. Selected Variables for the Analysis

Table 1. Descriptions of Variables

Variable name	Variable measurement in the RCT' survey
branchno	Branch number (10 unique values)
children	1 if has a child
age_quant	Age in years, divided into quantails: (10,15) (15, 17) (17-28)
rural	1 if lives in a rural village
rich	1 if income is larger than sample mean income
satisf_income	0 = not satisfied with earning, 6 = satisfied. Recoded into dummy (< 3 = 0, 3-6 = 1)
Expenditure_goods/ foll_expend_goods	"The variable measures monthly consumption expenditure in the following goods categories: "jewellery/ornaments, cosmetics/makeup, clothes, hairdressers, shoes/footwear, going to restaurants/bars/tea shop/cafe, talk time for your mobile phone and presents/gifts." The researchers in the original study excluded the first 1% outliers with high expenditures, reducing concerns that the estimates are not representative or driven the top of the distribution" (Bandiera et al., 2014, p. 35).
partner	1 if cohabitating/married/has a partner
empl	1 if employed
enroll_school	1 if enrolled in school
study_hours_quant	Hours spent studying if enrolled. Proxy to investment in education.
worry_job	Never worries that she will not get a good job
work_married	Would you like to work outside the house after you get married? 0=no, 1=yes
ablework_married	Will you be able to work outside the house after you get married? 0= no, 1=yes
want_respect	I want to be a respected person in my village 1=not at all, 10=a lot
who_decidehusband	Who will decide on your husband? Self=1, parents=2, self with parents=3, relatives=4
index_aspiration_quant	Recodified continuous variable, using the quantiles for creating groups. "The aspiration index averages the mean of the following standardised variables: "At what age you would you like your son to get married?", "At what age you would you like your daughter to get married?",

	“In your opinion, what is the suitable age for a woman to have the first baby?”, “How many children would you like to have (including the ones she might already have)?”, “Suitable age for marriage for a female?”, “Suitable age for marriage for a male?” (Bandiera et al. [code book], 2014).
index_iga_quant	Recodified continuous variable, using the quantiles for creating groups. “The income generating activities index averages the mean of the following standardised variables: “The respondent does any kind of income generating activities”, “Self-employed”, “Employed”, “Expenditure on goods in the last month”, “The entrepreneurial ability index is the cumulative and rescaled score aggregating the self-assessed ranks to the following activities (where 10 was the highest rank and 1 the lowest): "Run your own business", "Identify business opportunities to start up new business", "Obtain credit to start up new business or expand existing business", "Save in order to invest in future business opportunities", "Make sure that your employees get the work done properly", "Manage financial accounts", "Bargain to obtain cheap prices when you are buying anything for business (inputs)", "Bargain to obtain high prices when you are selling anything for business (outputs)", "Protect your business assets from harm by others", "Collecting the money someone owes you". The index for satisfaction with earnings/income is the reversed and rescaled respondent's self-assessment on a 7 point score (where originally "1" is completely happy and "7" is not at all happy)” (Bandiera et al., 2014, p. 40).
index_empowerment_quant/ foll_index_empower_quant	Recodified continuous variable, using the quantiles for creating groups. “The gender empowerment index is a variable that cumulates the number of times a respondent answers "Both/Same" to the following questions: "Who should earn money for the family?", "Who should have a higher level of education in the family?", "Who should be responsible for washing, cleaning and cooking?", "If there is no water pump or tap, who should fetch water?", "Who should be responsible for feeding and bathing children?", "Who should help the children in their studies at home?" and "Who should be responsible for looking after the ill persons?" The other possible answers given to the respondent were "Male" and "Female". The index is then rescaled such that 100 indicates that the respondent answered that the female should (at least partly) be responsible for all the activities.” (Bandiera et al., 2014, p. 32).
index_controlbody_quant/ foll_index_controlbody_quant	Recodified continuous variable, using the quantiles for creating groups. “The pregnancy knowledge index equals 1 if the respondent correctly identifies the statement "A women cannot become pregnant at first intercourse or with occasional sexual relations" as true or false. The HIV knowledge index is based on the number of statements correctly identified as true or false. The relevant statements are "A person who has HIV is different from a person who is ill with AIDS", "During vaginal sex, it is easier for a woman to receive the HIV virus than for a man", "Pulling out the penis before a man climaxes keeps a women from getting HIV during sex", "A women cannot get HIV if she has sex during her period", "Taking a test for HIV one week after having sex will tell a person if she or he has HIV." and "A Pregnant woman with HIV can give the virus to her unborn baby", “Uses any other (than condum) form of contraceptives”, “Always uses a condum”, “Do you have any children?”, “Has a partner”, “Have you been involved in any sexual intercourse that were not willing to do” (Bandiera et al., 2014, p. 40).
club_sometimes	Adolescent attends club meeting at least time but less than 3 times a week
club_frequently	Adolescent attends club meeting at least 3 times a week

4.3. Applying Causal Forest to the Data

The GRF package is built on the random forests' theory (Breiman, 2001) and boils down to taking the difference of two weighted means, in our case of binary treatments (Knaus et al., 2018). Weighting helps to define an adaptive local neighbourhood around the covariate value of interest and shows that the estimator can be consistent and asymptotically normal for a fixed covariate space. Also, Wager and Athey (2018) have shown that, in the presence of confounding, the causal forest model performance can be improved by introducing local-centering. Another important feature is honesty, which uses different subsamples for making predictions and for growing the trees, reduces prediction bias and ensures consistent estimates.[29] Specifically, honesty means that in training the causal forest, the real measured outcome of each observation is used to establish the within-leaf treatment effect or to assert where to split the sample, but not both (Wager et al., 2018).

The first challenge of causal inference is that individual treatment effects can never be observed, as we are missing the counterfactual, i.e., what would have happened to the individual had she not undertaken a specific treatment. Consequently, the model aims at estimating credible, stable and unbiased averages of the individual treatment effects, namely ATE combined with estimation of the CATE of subgroups in the studied population:

$$\tau(x) = E[Yi(1) - Yi|Xi = x]$$

4.4. Model Specification

Despite the fact that the original study treats treatment assignment as an instrument that generates exogenous variation, we do not use the causal forest function as the instrumental forest function. The reason for this is that when the instrument and treatment assignment coincide, both functions essentially operate in the same manner (Athey et al., 2019; Friedberg et al., 2020).

To estimate the treatment effects using causal forests, we started by training the outcome propensities, *Y.forest*, and the treatment propensities, *W.forest*. Following Athey and Wager (2019), we then took the predictions of the *Y.forest* and the *W.forest*, namely the *Y.hat* and the *W.hat* and specify them in the causal forest training parameters as priors for the predictions.[30] This

[29] Honest forests first randomly divide a subsample in two and use only the first for the splitting process. The second half of the subsample feed into the tree's leaf nodes, thereby, 'repopulating' the leaf nodes with new observations after the splitting.
[30] These propensities can also be estimated using a different method like boosting.

can also help eliminate confounding effects. We grew 20,000 trees to ensure estimates are not sensitive to the number of trees. We introduced a village cluster as in the original analysis and set the *equalize.cluster.weights* = TRUE to account for differing cluster sizes. Clustering meant to improve variance estimates and the accuracy of predictions. We then assumed that our sample is not small and set the *honesty* as true;[31] our goal was to reduce bias in the tree predictions. We also set *tune.parameter* = TRUE to automatic parameter selection and set *orthog.boosting* = TRUE to avoid wasteful computation.[32]

The first, raw causal forest included all the possible covariates we could include. Because of the flexibility of the causal forest and its ability to capture non-linear relations, we took the approach of first training a model which includes all the potentially relevant baseline covariates which were available in the dataset. Doing so, we introduced 22 different variables[33], including both objective information regarding socio-economic factors and subjective information regarding girls' attitudes and aspirations at baseline.

We then trained a second causal forest that only selected the important variables. These are simply determined by a weighted sum capturing number of splits a feature had in every depth layer in the causal forest. In that sense, we only include in our final causal forest those variables who weighed more, so we can make more splits in all those features that can potentially be treatment modifiers (Athey et al., 2019). Lastly, we repeated the exact same process across all three outcomes.

4.5. Analytical Strategy

Our first interest was to estimate the overall effectiveness of the programme on the full sample and on the three outcomes of our interest with village cluster and standard errors. This is achievable by using the built-in *average_treatment_effect* function, inside the package GRF, which implicitly assumes a similar treatment effect across heterogeneous individuals with different characteristics.[34]

---

[31] If sample is small, honesty = TRUE can hurt the performance, however, without this parameter, bias is likely to increase.
[32] Without setting this parameter to TRUE, the causal forest would perform double splitting, on features that impact treatment effects and treatment propensities. Because the estimates were already decoupled when we introduced the causal forest model Y.hat and W.hat from trained regression forests.
[33] See Table 1 for the variables list.
[34] The average treatment effect function is based on a variant of augmented inverse propensity weighting (Robins et al., 1994; Wager & Athey, 2018).

For the study of treatment heterogeneity, we undertook the following strategy: for each outcome, we first performed two separate tests to check whether heterogeneity is present in the underlying signal. Firstly, we tested the BLF is a test for the presence of heterogeneity that set the CATE[35] to be a linear function of the causal forest out-of-bag treatment effect estimates[36] (Chernozhukov, Demirer, et al., 2018; Wager & Athey, 2018). Secondly, we artificially constructed three subgroups, ntiles, according to the magnitude of ATE estimates as measured from the out-of-bag predictions and see, using F-test, whether these effects are significantly different from one another, which would suggest the presence of heterogeneity.[37]

We then analysed the results through a predetermined threshold. Specifically, if both tests were insignificant, we did not move forward to conduct formal analysis to test heterogeneity across covariates, unless we had a theoretical justification to expect variation. Then, we moved to test heterogeneity along these specific covariates. Contrastingly, in outcomes where one or two of the initial tests indicated the presence of heterogeneity, we went forward to try and understand its source. For that, we performed a formal analysis of heterogeneity across covariates. One structured analysis is to pre-specify subgroups and estimate their CATEs (Fang, 2020). To establish that, we calculated, again using the GRF ATE function, ATEs in each level of specific covariates, including a quantification of the uncertainty of these estimates using 95 percent confidence intervals. Specifically, we focused this analysis on variables that can later be used for targeting and that were selected as "important" by the model or estimated by our theoretical framework as potential drivers of heterogeneity. This analysis enabled us to test each covariates' levels effects, when all else remains equal. Due to data constraints, we decided not to try and establish subgroups that account for interactions between covariates because that would have resulted in subgroups with too little observations, which would have prevented accurate analysis. [38]

[35] A CATE is an average treatment effect specific to a subgroup of population. The subgroup is defined by individuals' characteristics (Fang, 2020).
[36] It checks if the model properly is well calibrated and if it captures treatment heterogeneity (Athey & Wager, 2019).
[37] The results are non-monotonic meaning that ntiles' estimates may not be ordered correctly, especially when samples are small (Athey, 2019).
[38] One can potentially construct subgroups with interaction between two variables' levels, however, due to our relatively small sample size, this process will result in subgroups which are too small for the model to produce credible estimates, and therefore we avoided this practice.

4.6. Model Evaluation

Before running the model, we predetermined how we will assess their performance. Our strategy here involves two dimensions. First, we were interested in evaluating whether causal forest is more flexible than linear models in terms of model specification. Secondly, we wanted to determine if it is more capable of detecting treatment heterogeneity.

To achieve this, we compared our results to the linear analysis in the original paper as well as replicated the main results from the paper using our pre-processed data set. This allowed us to first validate the sanity and quality of the ATE results, and then to be able to assert whether causal forest is advantageous compared to a linear model.

Following the methodology of Bandiera et al., (2014), we calculated the intent-to-treat (ITT) impact of the programme, where the instrument is treatment assignment as girls self-selected to participate. We estimated the impact of the program on the outcomes y_{ij1} for every girl *i* living in community *j* at follow-up (t=1):

$$y_{ij1} = a + BX_{ij0} + \gamma treat_j + \delta y_{ij0} + \varepsilon_{ij1}$$

We ran an OLS and control for age, branch number BX_{ij0} and several baseline values y_{ij0} (t=0). To simulate our causal forest models, we constructed the linear models based on the theoretical framework and accounting for possible interactions effects. This included variables that the causal forest found to weigh more in the splitting procedure and variables that the theoretical framework emphasised, specifically whether they had a partner, children, are enrolled in school or from a rural area. We also added a village cluster in the error term ε_{ij1} because we wanted to control for unobservables within communities that determine outcomes. The following equations describe our models' specifications:

$$\begin{aligned} expenditure_goods_{ij1} &= a + B_1 age_{ij0} + B_2 branchno_{ij0} + \gamma treat_j + \delta_1 enroll_school_{ij0} \\ &+ \delta_2 expenditure_goods_{ij0} + \delta_3 rich_{ij0} + \delta_4 empl_{ij0} + \delta_5 partner_{ij0} + \delta_6 children_{ij0} \\ &+ \delta_7 rural_{ij0} + \delta_8 index_empowerment_quant_{ij0} + \delta_9 enroll_schoool{:}partner_{ij0} \\ &+ \delta_{10} enroll_school{:}rich_{ij0} + \delta_{11} expenditure_goods{:}rural_{ij0} + \delta_{12} enroll_school{:}rich_{ij0} \\ &+ \delta_{13} empl{:}partner_{ij0} + \delta_{14} index_empowerment_quant{:}rich_{ij0} + \varepsilon_{ij1} \end{aligned}$$

$$
\begin{aligned}
index_empowerment_{ij1} \\
&= a + B_1 age_{ij0} + B_2 branchno_{ij0} + \gamma treat_j + \delta_1 enroll_school_{ij0} \\
&+ \delta_2 expenditure_goods_{ij0} + \delta_3 rich_{ij0} + \delta_4 empl_{ij0} + \delta_5 partner_{ij0} + \delta_6 children_{ij0} \\
&+ \delta_7 rural_{ij0} + \delta_8 index_empowerment_quant_{ij0} \\
&+ \delta_9 index_empowerment_quant{:}\, enroll_school_{ij0} \\
&+ \delta_{10} index_empowerment_quant{:}\, empl_{ij0} + \delta_{11} index_empowerment_quant{:}\, rural_{ij0} \\
&+ \varepsilon_{ij1}
\end{aligned}
$$

$$
\begin{aligned}
index_controlbody_{ij1} \\
&= a + B_1 age_{ij0} + B_2 branchno_{ij0} + \gamma treat_j + \delta_1 enroll_school_{ij0} \\
&+ \delta_2 expenditure_goods_{ij0} + \delta_3 rich_{ij0} + \delta_4 rural_{ij0} + \delta_5 index_controlbody_{ij0} \\
&+ \delta_6 index_empowerment_quant_{ij0} + \delta_7 index_controlbody{:}\, rural_{ij0} \\
&+ \delta_8 index_controlbody{:}\, age_{ij0} + \delta_9 index_controlbody{:}\, enroll_school_{ij0} + \varepsilon_{ij1}
\end{aligned}
$$

We also compared the causal forest results of the CATEs to the results of linear models with treatment-by-treatment interaction effects. The model specification for each covariate checked was:

$$y_{ij1} = a + BX_{ij0} + \gamma treat_j + \delta treatment{:}\, M_{ij0} + \varepsilon_{ij1}$$

Because this is an RCT, we assume the covariates M_{ij0} are randomly assigned, and hance treatment-by-treatment interactions may be interpreted causally. To avoid certain unbalance of the covariates and expect more precision, we control for age and branch number BX_{ij0}. Following our interests, defined by the hypothesis, the covariates chosen to check with linear models were age, children, expenditure, enrolled in school, GEI, and rural.

4.6.1. Assumptions

Unconfoundedness Assumption

The unconfoundedness assumption means that treatment is exogenous, or that a credible instrument is available, Formally, unconfoundedness can be represented as:

$$W_i \perp (Y_i(0), Y_i(1) \mid X_i)$$

In relation to the experiment on the basis of our data, and because girls could decide whether they participate or not, the treatment itself is the instrument that enables causal interpretation of the

programme's effects on girls' outcomes. In other words, unconfoundedness requires that the covariates presented in the model include all confounding factors that affect both the assignment to a treatment group and the outcome.

Overlap Assumption

First, we check the overlap assumption by looking whether the propensity scores are bounded away from 1 and 0 (Elek & Bíró, 2021; Wager & Athey, 2018). This assumption requires a positive probability of treatment for each individual no matter her characteristics (covariates). Further, it validates that at every level of covariates, both treated and untreated girls are observed. Formally, it tests whether

$$0 < \Pr(W_i = 1 \mid X_i = x) < 1 \; \forall \, x \, .$$

Model bias

Bias in the estimation of ATE can emerge if pre-treatment variables that are related to both potential outcomes and the treatment exit. Bias can be the sparsity of the product of the conditional expected outcome or sparsity in the propensity score. So, it is not so much a problem when a pre-treatment variable is only correlated with either the potential outcomes or the treatment. This condition may challenge the estimation of the conditional expected potential outcome or propensity score but will not produce bias per se (Athey et al., 2017). Accordingly, we measured bias by applying this method which captures the "effort" the propensity scores and outcome models put in to get unbiased estimates. This is then scaled by the standard errors to produce the bias estimates.

$$bias(x) = (e(x) - p) \, X \, (p(u(0, x) - u_0) + (1 - p)(u(1, x) - u_1)$$

5. Descriptive statistics/Baseline characteristics

The following tables, Table 2 and Table 3 show the summary statistics of the data as related to the status of the adolescent girls.

Table 2. Descriptive statistics of continuous variables by treatment status at baseline

Control

Statistic	Mean	St. Dev.	N
Age	16.412	2.956	1,973
Study Hours	60.650	19.515	1,321
Expenditure on Goods	11,249.470	17,924.470	1,989
Gender Empowerment	31.262	24.889	2,002
Control Over Body	0.000	1.000	1,609
Income Generating Activity	-0.000	1.000	1,609
Aspirations	-0.000	1.000	1,609

Treatment

Statistic	Mean	St. Dev.	N
Age	16.349	2.802	3,922
Study Hours	61.891	18.267	2,499
Expenditure on Goods	12,518.730	19,508.340	3,937
Gender Empowerment	28.453	24.226	3,964
Control Over Body	-0.061	0.983	3,279
Income Generating Activity	0.075	1.128	3,279
Aspirations	0.030	0.997	3,279

Note. Descriptive statistics of continuous baseline variables used in the causal forest. This table presents a description of the original variables, and not the recodified one into quantiles. Sample size of two-years panel is 4,888. The indexes: IGA, COB, and Aspirations are composed of several standardized variables. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1. The Gender Empowerment index cummulates answers regarding gender roles and the index Control Over Body represents the extent of reproductive health literacy and family planning attitudes. The IGA index is composed of the entrepreneurship index, if the girls had any income generating activity, if they were employed or self employed, and their expenditure on goods. The Aspirations index is composed by standardised variables measuring opinions about marriage and childbearing. Higher values indicate less traditional social norms. See Table 1 for full specification of the outcomes.

Table 3. Descriptive statistics of discrete variables by treatment status at baseline

Treatment status	Children		Partner		Rural		Age			Study hours				Enrolled school		Rich		Satisfied income		Employed	
	0	1	0	1	0	1	[10,15]	(15,17]	(17,28]	[0,50]	(50,60]	(60,72]	(72,166]	0	1	0	1	0	1	0	1
0	1741	196	1742	223	1001	1001	872	456	645	366	349	329	277	560	1442	1002	987	599	517	1931	71
1	3442	402	3526	364	1975	1989	1771	875	1276	599	660	600	640	1149	2815	1965	1975	1324	1038	3745	219

Treatment status	Expenditure			Empowerment			Control over body			Aspirations			
	[0,1.77e+03]	(1.77e+03,1.03e+04]	(1.03e+04,1.45e+05]	[0,14.3]	(14.3,42.9]	(42.9,100]	[-3.37,-0.608]	(-0.608,0.0553]	(0.0553,0.612]	(0.612,5.27]	[-5.63,-0.154]	(-0.154,0.479]	(0.479,2.61]
0	695	677	617	753	755	494	403	437	381	388	532	448	629
1	1283	1303	1351	1666	1490	808	819	785	841	834	1183	1301	795

Note:

Number of observations depending on treatment status. Sample size is 5,966. Variable specifications: Children: 1 if has children, Partner: 1 if has a partner, Rural: 1 if from a rural area, Enrolled school: 1 if enrolled, Rich: 1 if income above mean, Satisfied income: 1 if any level of satisfaction, 0 if not satisfied, Employed: 1 if employed. The outcome Expenditure refers to monthly consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The outcome Gender Empowerment index (rescaled from 0 to 100) is based on multiple questions relating to gender roles in labor markets, education and household chores. Higher values indicate more empowerment. The index Control Over Body is composed of answers to questinos measuring the extent of reproductive health literacy and family planning attitudes. Higher values indicate better control over body. The Aspirations index is composed by standardised variables measuring opinions about marriage and childbearing. Higher values indicate less traditional social norms. See Table 1. for full specification of the outcomes.

Furthermore, given the assumptions of a randomised experiment,[39] we began to analyse the data by comparing the average outcome in treatment and control. The results of the T-test on the three primary outcomes of our interest can be seen in Table 4 below:

Table 4. Results of the t-test on outcomes of interest

Variable	Means difference	Control mean	Treatment mean	Statistics	P-value
Expenditures	4962.82	15999.82	20962.64	-6.39	0.00
Control over body	0.54	0.00	0.54	-19.84	0.00
Gender Empowerment	2.58	28.58	31.16	-3.80	0.00

Note:
Ttest of the difference in means between treatment and control groups on the full sample. The outcomes were measured with two years follow-up. Signicance at 95%. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The Gender Empowerment index cummulates answers regarding gender roles and the index Control Over Body represents the extent of reproductive health literacy and family planning attitudes. See Table 1. for full specification of the outcomes.

The output indicates that there is a significant difference between the treatment and control groups (see column p-values). The p-value statistics means that the difference in our interest outcomes between the two groups seems significantly large enough.

[39] Specifically, based on the arguments made in Bandiera et al. (2014), we assumed independence, meaning that the treatment status is independent of potential outcomes, and the Stable Unit Treatment Values Assumption (SUTVA) hold.

6. Results

The findings of the original paper indicate that the programme had a quantitatively substantial positive impact on the economic and social empowerment of the treated adolescent girls relative to girls in control communities.[40]

6.1. The Average Treatment Effect (ATE)

Table 5 indicates that all outcomes effects are positive and have the same positive direction, showing a sizable positive effect of the programme. The reported treatment effect on expenditure indicates that by follow-up expenditure is 4798 points higher for adolescent girls in treated communities than for those in control communities. Next, the ATE estimate for empowerment is 2.34 points higher for the treated girls in respect to the girls in the control group. Finally, the effect on COB (an index composed of variables such as childbearing, marriage and sex) for adolescent girls is 0.55 points higher than for girls who did not take part in the programme.

Table 5. Causal Forest Estimated Average Treatment Effects by Outcome

	Expenditure on Goods	Gender Empowerment	Control Over Body
Estimated ATE	4798.227	2.344997	0.5471981
Standard error	1284.566	1.197601	0.0741198

Note:
Estimated ATEs by a trained Causal forest. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The Gender Empowerment index cummulates answers regarding gender roles and the index Control Over Body represents the extent of reproductive health literacy and family planning attitudes. See Table 1. for full specification of the outcomes.

6.2. Replications of the results using OLS

Comparing between the causal forest and the linear replication results, the estimated treatment effects are overall similar. For Expenditure on Goods and the GEI outcomes are higher when using linear regressions (Table 5 and Table 6). The ATE estimated by linear regression for the outcome

[40] See Table 1 for specification of outcomes.

COB is slightly smaller than the ATE estimated by the causal forest.[41] However, on three outcome, the standard errors, which are also clustered by community, are slightly bigger using the causal forest relative to the linear model.

Table 6. Linear regression Estimations of Average Treatment Effects by Outcome

	Dependent variable:		
	Expenditure on Goods	Gender Empowerment	Control Over Body
	(1)	(2)	(3)
Treatment	4,882.331***	2.708***	0.516***
	(727.739)	(0.670)	(0.028)
Observations	4,540	4,579	4,768
R^2	0.115	0.063	0.176
Adjusted R^2	0.110	0.058	0.173
Residual Std. Error	23,934.620 (df = 4513)	21.620 (df = 4553)	0.841 (df = 4747)
F Statistic	22.472*** (df = 26; 4513)	12.209*** (df = 25; 4553)	50.856*** (df = 20; 4747)

Note: *p<0.1;* ***p<0.05;*** p<0.01

Estimated ATE by linear regression on the three outcomes at two-year follow-up. Standard errors in brackets. All models control for branch number, age, if they were rich, partner, if they lived in rural area, if they had children, enrolled in school, if they were employed and what were their expenditure on goods. Moreover, the model for Expenditure on Goods also control for the gender empwoerment index. The model for Gender Empowerment controls for the index gender empowerment at baseline. The model for COB controls for index COB and for index gender empowerment at baseline. Finally, all models account for several interactions. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1. The Gender Empowerment index cummulates answers regarding gender roles and the index COB represents the extent of reproductive health literacy and family planning attitudes. See Table 1. for full specification of the outcomes.

6.3. Assessing Heterogeneous Treatment Effects

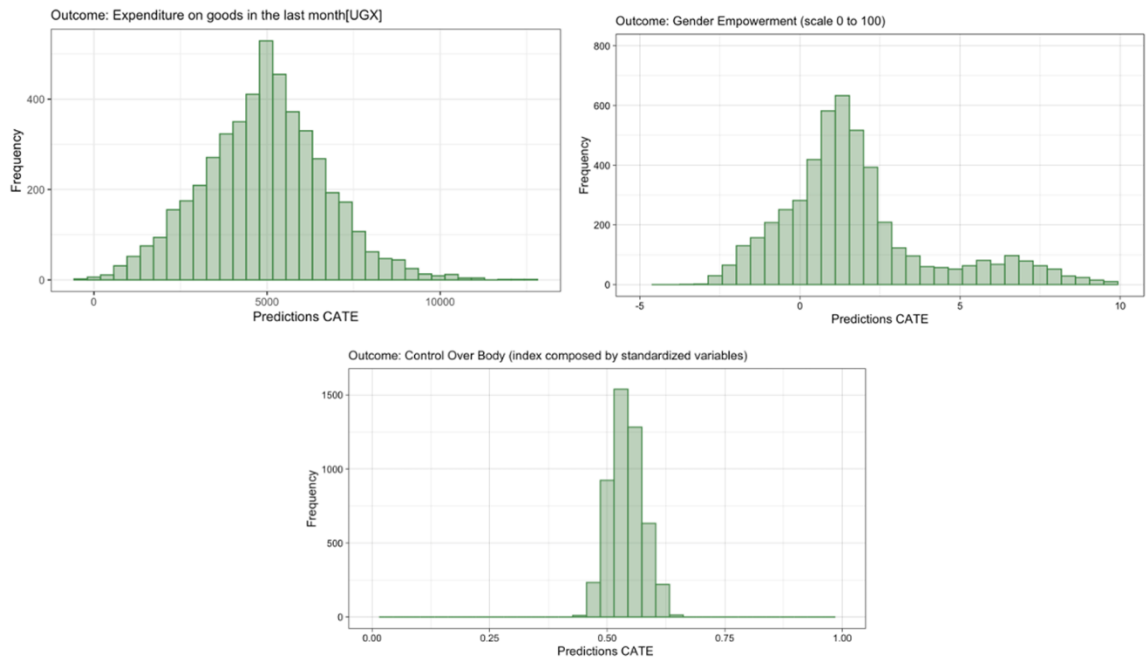
Following our theoretical framework, we assumed that while some treatment effects can be average for some individuals, certain subgroups will have treatment effects that are considerably different from the ATE. In this section, we present the results of our causal forest heterogeneity analysis. This analysis examines whether heterogeneity exists between the subgroups and across specific covariates, i.e., individual characteristics of the treated adolescent girls.

[41] When inserting to the causal forest only the variables that were introduced in the linear analysis, the ATEs are overestimated. Table F. 1. This seem to imply that when using causal forest, it is important to feed the model with all possible baseline indicator, to allow for a credible non-parametric estimation of the effects.

6.3.1. Variation in CATE estimates

We observe the histogram of the CATE estimates in every outcome to see if the causal forest is estimating heterogeneity. Figure 1 exhibits variation in the out-of-bag CATE predictions from the causal forests.[42] The predictions are produced using trees that were not used in the training of the model. From this, we already sense that the ATE varies more for the outcome expenditure, while it does not vary for the outcome COB. In gender empowerment, we observe minor variation. Importantly, we cannot conclude from these histograms whether any of this variation is meaningful.

Figure 1. Histogram of Out-of-bag CATE predictions from trained Causal Forest by Outcome



Note: Predictions made on held-out data from the full sample that was not used for training.

[42] See Appendix H, Figure H. 1 to observe the standard errors.

6.3.2. "Best linear fit" method

Following Chernozhukov, Demirer et al. (2018) and Athey and Wager (2019) we use the method "best linear fit" that seeks to fit the CATE as a linear function of the out-of-bag causal forest estimates. On the one hand, the result from running the *test_calibration* function shows that the model is well calibrated on all of the three outcomes. Table 7, Table 8, and Table 9 show that the coefficients of the mean forest predictions are significant and close to 1. This is a measure of the quality of the estimates of treatment heterogeneity (Athey & Imbens, 2019; Athey & Wager, 2019; Wager & Athey, 2018). We can conclude our causal forests for the three outcomes of interest have a good fit.

Table 7. Best Linear Fit for Expenditure causal forest model

```
Best linear fit using forest predictions (on held-out data)
as well as the mean forest prediction as regressors, along
with one-sided heteroskedasticity-robust (HC3) SEs:

                              Estimate Std. Error t value    Pr(>t)
mean.forest.prediction         0.95639    0.25460  3.7564 8.719e-05 ***
differential.forest.prediction 0.24915    0.59670  0.4176    0.3381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 8. Best Linear Fit for Gender Empowerment causal forest model

```
Best linear fit using forest predictions (on held-out data)
as well as the mean forest prediction as regressors, along
with one-sided heteroskedasticity-robust (HC3) SEs:

                              Estimate Std. Error t value    Pr(>t)
mean.forest.prediction         1.40521    0.72853  1.9288   0.02691 *
differential.forest.prediction 2.09625    0.48149  4.3537 6.831e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 9. Best Linear Fit for Control Over Body causal forest model

```
Best linear fit using forest predictions (on held-out data)
as well as the mean forest prediction as regressors, along
with one-sided heteroskedasticity-robust (HC3) SEs:

                              Estimate Std. Error t value    Pr(>t)
mean.forest.prediction         0.99637    0.13190  7.5542 2.503e-14 ***
differential.forest.prediction -4.85962    3.01724 -1.6106    0.9463
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On the other hand, with this first approach to test heterogeneity, we conclude that the model is capturing heterogeneity in the underlying signal only in one of the three outcomes of interest, namely the GEI. The differential forest prediction coefficient is positive and statistically significant, which means we can reject the null hypothesis that treatment effects are homogenous

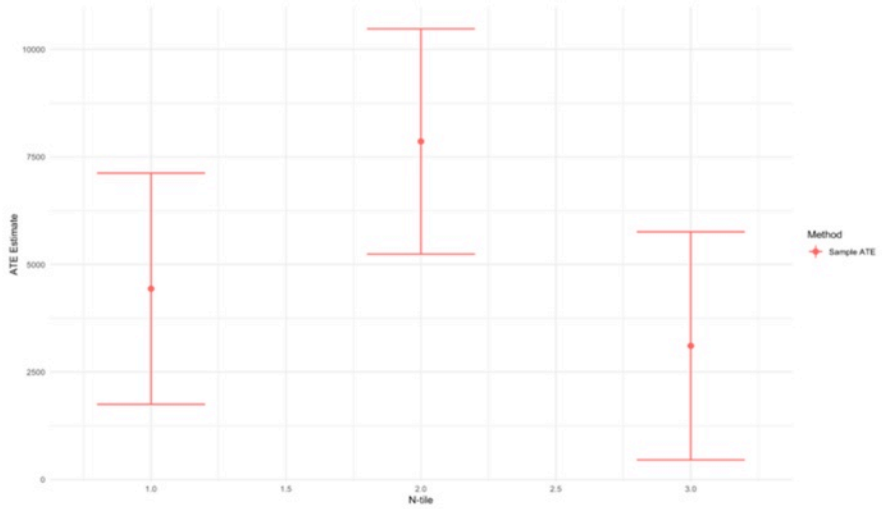
(Athey & Wager, 2019; Chernozhukov, Chetverikov, et al., 2018). Although the p-value of the differential forest prediction coefficient acts as a first test for the presence of heterogeneity, Athey and Wager (2019) are cautious to mention that particular covariates may be strongly correlated with the outcome and not show up in the splitting procedure.

6.3.3. Assessing Heterogeneity Across Subgroups

We then opted for a second test for heterogeneity that involves grouping the sample based on the different strength of the predicted ATEs. In other words, we created subgroups based on ntiles of the predicted treatment effect, to after compute the ATE within each subgroup in all the three outcomes we explored, presented in Figure 2, Figure 3 and Figure 4. Because of the exogenous variation in the RCT, we expect this method will result in unbiased estimates of the subgroup specific treatment effect. For Expenditure on Goods and the GEI outcomes, the results of the F-test show that we can reject the null hypothesis that the estimated ATE are constant across ntiles. Specifically, as Table 10 shows, we find statistically significant differences between ntile2 and ntile3 for the Expenditure on Goods outcome. The effects across ntiles for the GEI in Table 11 are statistically significant between ntile 1 and ntile3 and between ntile 2 and ntile 3. However, for the outcome COB index, the estimated ATEs are constant across ntiles.[43]

[43] Because we cannot reject the null-hypothesis, meaning the ATE is constant across subgroups, we did not perform a pairwise test to estimate differences between ntiles.

Figure 2. ATE within Ntiles (as defined by predicted CATE). Outcome: Expenditure on Goods [UGX]



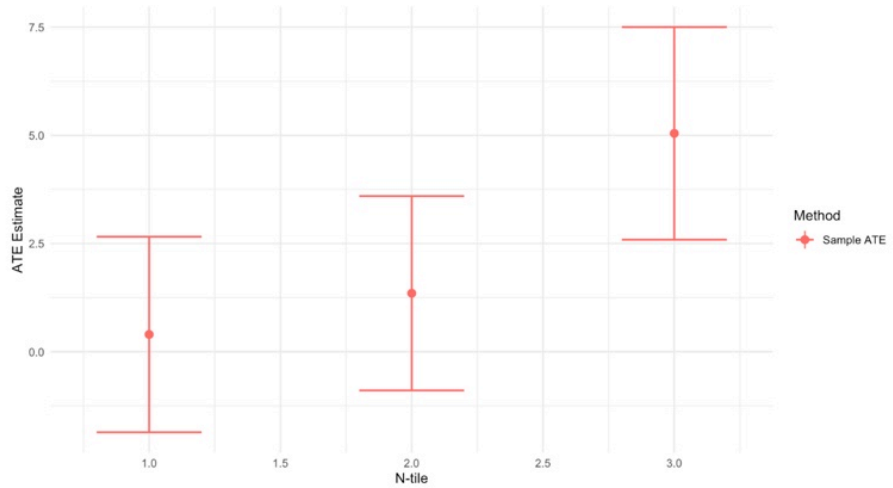
Note: ATE per subgroups defined by out-of-bag CATE. 95% CI.

Table 10. Pairwise ATE Ntiles differences. Outcome: Expenditure on Goods [UGX]

rowname	tile1	tile2	tile3
tile1			
tile2	(1914.081)		
tile3	-1327.335 (1926.536)	-4751.339 (1900.855)	

Note: Standard errors in parenthesis.
Significance: no background colour p≥0.05,
grey background p<0.05, black background p<0.01.
When adjusting for multiple testing,
if p-value below 0.016 then it is statistically different.

Figure 3. ATE within Ntiles (as defined by predicted CATE). Outcome: Gender Empowerment



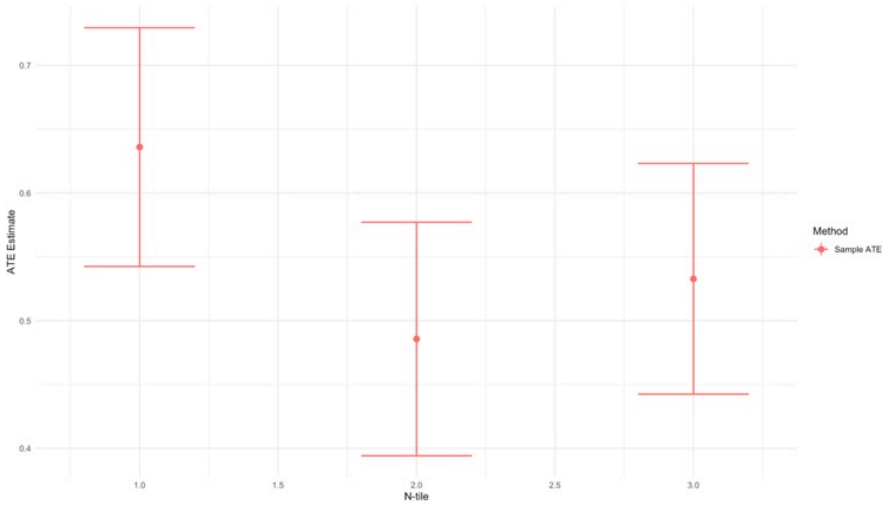
Note: ATE per subgroups defined by out-of-bag CATE. 95% CI.

Table 11. Pairwise ATE Ntiles differences. Outcome: Gender Empowerment

rowname	tile1	tile2	tile3
tile1	NA	NA	NA
tile2	0.952 (1.625)		
tile3	4.646 (1.703)	3.694 (1.698)	NA

Note: Standard errors in parenthesis.
Significance: no background colour p≥0.05,
grey background p<0.05, black background p<0.01.
When adjusting for multiple testing,
if p-value below 0.016 then it is statistically different.

Figure 4. ATE within Ntiles (as defined by predicted CATE). Outcome: Control Over Body



Note: ATE per subgroups defined by out-of-bag CATE. 95% CI.

After these two first checks to detect the presence of heterogeneity, the overall picture in our three outcomes is mixed. The causal forest predicts treatment heterogeneity in the gender empowerment outcome with the two first approaches. The results of the analysis of the ATEs within subgroups for expenditures show significant differences between levels of expenditure across subgroups of the treated girls, although the BLF test did not detect the presence of heterogeneity. Here, it is important to mention that the BLF is only a linear projection and as Athey and Wager (2019) discussed, a non-significant result is not sufficient to reject the presence of heterogeneity. Indeed, despite non-significant results of the BLF, we do detect heterogeneity when performing the ntiles analysis. We could not detect statistically significant different treatment effects across the ntiles for the COB outcome. Any treatment heterogeneity present is weak and therefore we decided not to explore it further.

6.3.4. Assessing Heterogeneity Within Subgroups

After we detected heterogeneity in treatment effects within subgroups, we ventured to explore the presence of heterogeneity in Expenditure on Goods and GEI by estimating the CATEs across

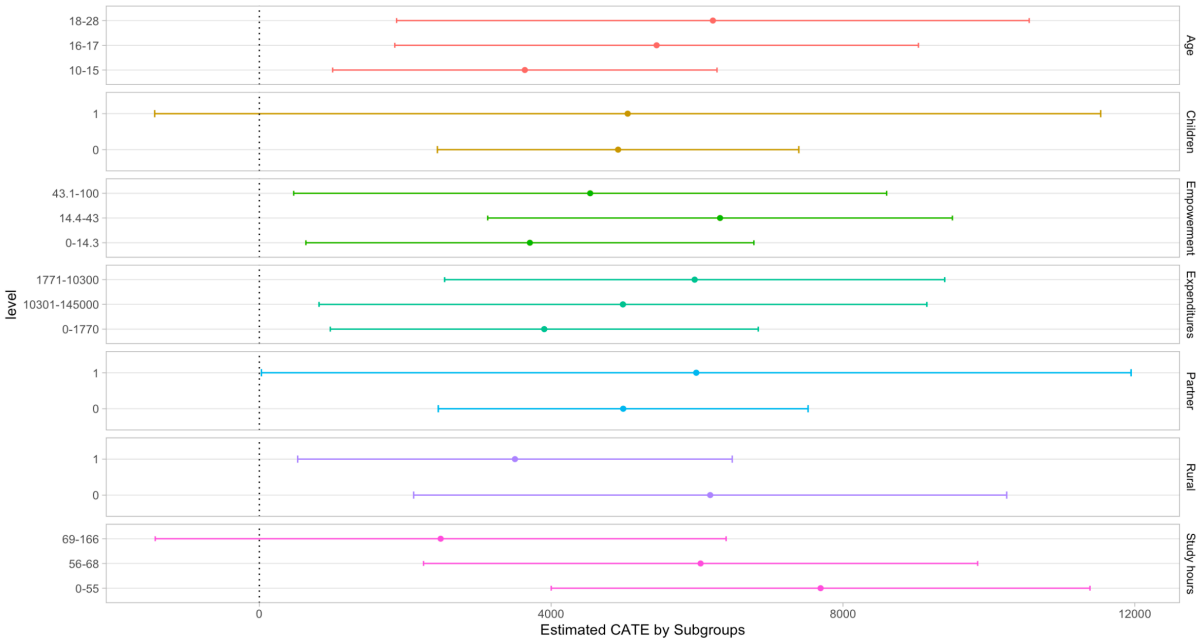
different levels of specific covariates.[44] Subgroup analysis is the most common analytical approach to examine heterogeneity (Varadhan & Seeger, 2013), as it has the potential to generate a fuller picture of how treatment effects differ between subgroups.

To formally test our respective hypotheses in the outcomes where heterogeneity was detected by the omnibus tests of the causal forest, we generally followed a data-driven approach. Hence, we selected variables the causal forest selected as important to make the splitting. These are shown in Appendix E. However, because it is possible that one covariate is not "important" in splitting, but rather varies strongly with treatment effects, we further considered theory and prior knowledge of treatment effect modifiers as conceptualised in our Theoretical Framework (Athey, 2019).

In Figure 5 and Figure 6 below, the dots show the conditional ATE along each level of the covariate, the lines of their respective confidence intervals, while the vertical black line shows the boundary for a zero impact. Overall, the results for this analysis do not show statistical differences in the estimated CATEs of the different levels of the covariates, and we observe overlapping confidence intervals. Nevertheless, the estimated CATEs from the causal forest can help us understand the direction of effects depending on covariates and thereby to learn about the working mechanisms of the ELA programme. We therefore discuss the direction of the effects for both outcomes. Before doing so, we compare the overall results from the causal forest with estimations from linear models we constructed, as seen in Table 12 and Table 13.

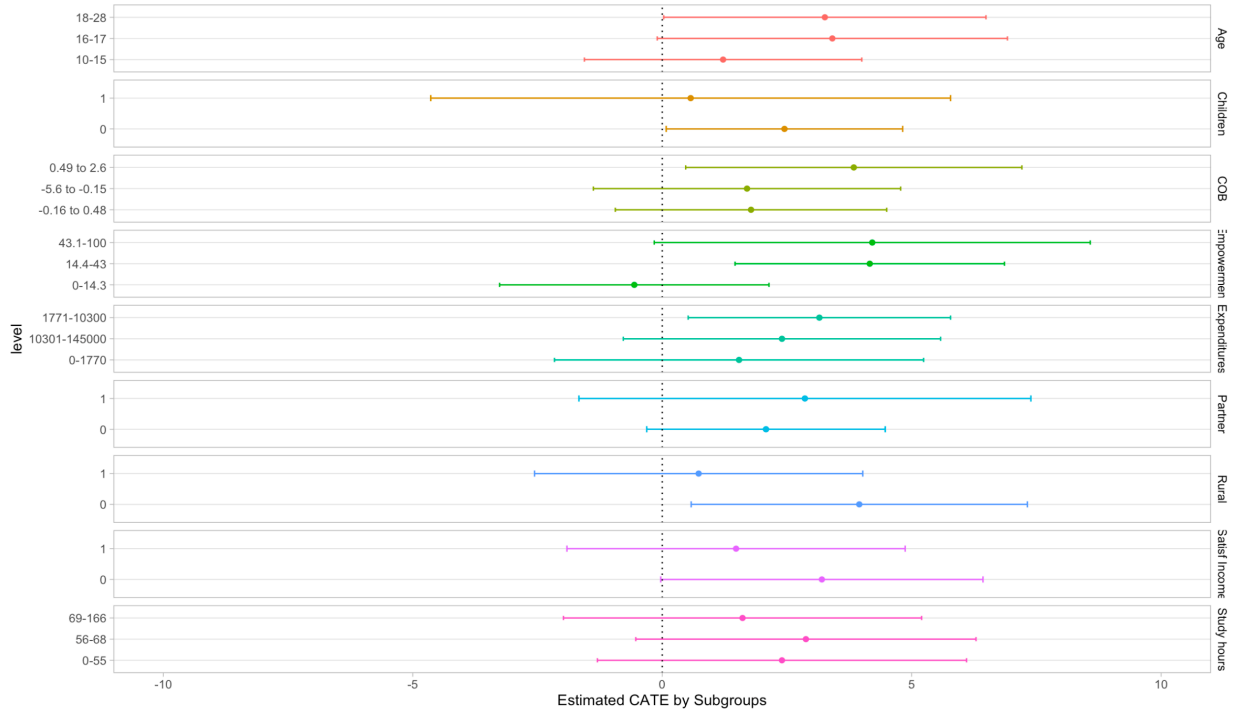
[44] Continuous variables were recoded into discrete values by taking the variables' quantiles. For this analysis, covariates were chosen in respect to the causal forest selected "important" variables, our theory-derived hypotheses, and the opinion of the surveyed experts.

Figure 5. Causal Forest Estimated CATEs. Outcome: Expenditure on Goods [UGX]



Note. 95% CI are calculated based on the standard error of each covariate level.
Each dot indicates the estimated ATE for the subgroup.
The vertical black dotted line designates the boundary of zero impact.
The outcome Expenditure refers to monthly consumption of goods in several goods categories.
The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$.
See Table 1. for full specification of the outcome.

Figure 6. Causal Forest Estimated CATEs. Outcome: Gender Empowerment



Note. 95% CI are calculated based on the standard error of each covariate level.
Each dot indicates the estimated ATE for the subgroup.
The vertical black dotted line designates the boundary of zero impact.
The outcome Gender Empowerment index is based on multiple questions relating to gender roles in labor markets, education and household chores (scale 0 to 100).
See Table 1. for full specification of the outcome.

First, in comparison to the estimated CATEs of the linear models looking at the treatment-by-treatment interaction effects (Table 12 and Table 13), we observe similar magnitudes and directions of the conditional average treatment estimates on both outcomes. The causal forest results seem more consistent with the direction of the magnitudes across the levels of the covariates, meaning that in higher quantiles of the variables the estimated average treatment effects are consistently increasing or decreasing more, depending on the specific case, than in lower quantiles with respect to the comparison group. For example, while for the two outcomes the linear predictions estimate girls between 16 and 17 years-old to have the highest effects compared to the other age categories, causal forest estimates indicate that the higher the age, the larger the estimated effect.

Table 12. Linear regression Estimated CATEs. Outcome: Expenditure on Goods [UGX]

	Dependent variable:					
	Expenditure Goods					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	3,994.817***	4,656.465***	4,089.037***	3,790.707***	4,192.456***	4,762.956***
	(1,113.492)	(794.578)	(1,275.160)	(1,397.556)	(1,213.464)	(1,062.819)
Treatment:Age[16-17]	1,455.527					
	(1,920.056)					
Treatment:Age[18-28]	626.175					
	(1,705.821)					
Treatment:Has children		-652.961				
		(2,459.189)				
Treatment:Medium Expenditure			1,532.891			
			(1,807.363)			
Treatment:High Expenditure			78.892			
			(1,838.129)			
Treatment:Enrolled School				1,002.203		
				(1,649.128)		
Treatment:Empowerment Score[14-43]					1,015.454	
					(1,720.439)	
Treatment: Empowerment Score[44-100]					-511.418	
					(1,962.771)	
Treatmnent:Lives Rural						-459.220
						(1,488.552)
Observations	4,791	4,655	4,752	4,791	4,791	4,791
R^2	0.108	0.110	0.110	0.109	0.109	0.108
Adjusted R^2	0.105	0.107	0.107	0.106	0.106	0.106
Residual Std. Error	24,093.840 (df = 4776)	24,024.180 (df = 4640)	24,066.890 (df = 4735)	24,086.420 (df = 4776)	24,090.370 (df = 4774)	24,092.540 (df = 4777)
F Statistic	41.339*** (df = 14; 4776)	40.840*** (df = 14; 4640)	36.665*** (df = 16; 4735)	41.575*** (df = 14; 4776)	36.393*** (df = 16; 4774)	44.486*** (df = 13; 4777)

Note: *p<0.1;* ***p<0.05;*** p<0.01

Note. Estimated CATEs by linear regression on dependent variable Expenditure on Goods at two-year follow-up. Standard errors in brackets. All models control for branch number and age. The outcome Expenditure refers to monthly consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1. See Table 1. for full specification of the outcome. The comparison groups are: model 1 - Age[10-15], model 2 - Has no children, model 3 - Low Expenditure, model 4 - Not enrolled in school, model 5 - Empowerment Score[0-13], model 6 - Lives in urban area.

Table 13. Linear regression Estimated CATEs. Outcome: Gender Empowerment

	Dependent variable:					
	Gender Empowerment					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	2.517**	2.905***	1.676	3.506***	-0.068	3.828***
	(0.998)	(0.713)	(1.142)	(1.248)	(1.084)	(0.950)
Treatment:Age[16-17]	0.213					
	(1.721)					
Treatment:Age[18-28]	0.919					
	(1.525)					
Treatment:Has children		-0.841				
		(2.201)				
Treatment:Medium Expenditure			2.631			
			(1.617)			
Treatment:High Expenditure			1.111			
			(1.644)			
Treatment:Enrolled School				-0.881		
				(1.474)		
Treatment:Empowerment Score[14-43]					5.102***	
					(1.538)	
Treatment: Empowerment Score[44-100]					4.044**	
					(1.755)	
Treatmnent:Lives Rural						-1.881
						(1.332)
Observations	4,831	4,695	4,792	4,831	4,831	4,831
R^2	0.065	0.064	0.065	0.065	0.068	0.066
Adjusted R^2	0.062	0.061	0.062	0.063	0.065	0.063
Residual Std. Error	21.649 (df = 4816)	21.637 (df = 4680)	21.615 (df = 4775)	21.648 (df = 4816)	21.622 (df = 4814)	21.643 (df = 4817)
F Statistic	23.992*** (df = 14; 4816)	22.959*** (df = 14; 4680)	20.808*** (df = 16; 4775)	24.009*** (df = 14; 4816)	21.925*** (df = 16; 4814)	25.976*** (df = 13; 4817)

Note: *p<0.1;* ***p<0.05;*** p<0.01

Note. Estimated CATEs by linear regression on the three outcomes at two-year follow-up. Standard errors in brackets. All models control for branch number and age. The Gender Empowerment index cummulates answers regarding gender roles (scale 0 to 100). Higher values indicate more empowerment. See Table 1. for full specification of the outcomes. The comparison groups are: model 1 - Age[10-15], model 2 - Has no children, model 3 - Low Expenditure, model 4 - Not enrolled in school, model 5 - Empowerment Score[0-13], model 6 - Lives in urban area.

6.3.4.1. Expenditure on Goods and Gender Empowerment

We first report, as expected, the found variability in treatment effects depending on the BRAC branch girls attended.[45] This variability implies that the quality or other aspects of the implementation of the programme in different communities had considerable effects on the programme's success and thus the magnitude of the resulting positive effects on treated girls. For the outcome Expenditure on Goods, we checked for differences in different levels of the following covariates: how many hours they invested studying, consumption expenditure, GEI, age if they had a partner, children and if they lived in a rural area. For GEI, we chose potential moderator

[45] See Appendix F, Figure F. 1 and Figure F. 2.

variables that include the baseline of education,[46] level of consumption, level of gender empowerment before the programme and other social characteristics such as age, if they have a partner, children, and their household geographical location.

Our results for the two reported outcomes appear to signal heterogeneity in the treatment effects of the programme across urban and rural communities. The estimated CATEs for each of these levels of the covariate seem to deviate from each other, indicating higher estimated effects for girls from urban areas no matter the outcome observed. In addition, the age of the girls seems to affect treatment effects of the intervention. For both outcomes, economic and social empowerment, the results give the impression that older girls benefited more from the programme. In the same sense, the estimated CATE for girls with higher consumption of goods at baseline appear to be bigger than the CATE for girls with low expenditure within their household.

Adolescent girls with and without partners have very similar estimated CATEs. In that sense, it would appear that being in a relationship does not affect the effects of the treatment on the economic and social dimensions chosen to be improved. Also, it is relevant to note how effects on expenditures and gender empowerment diverge for girls with and without children. We notice a weak signal that having children hinders the program's impact on girls' score in the GEI, but it did not affect their Expenditure on Goods.

Moreover, it is interesting to note the opposite directions of the CATEs on both outcomes in relation to education. In general, we did not see signals of heterogeneity among girls enrolled in school and those who were not. Nevertheless, the results with regard to study hours seem to indicate heterogeneity of treatment effects in regards to girl's economic empowerment following the programme. The participants who invested less time studying in baseline had higher treatment effects. On the other hand, for GEI outcome, girls who invested more hours studying do not appear to show a different effect from girls who invested less hours studying.

[46] As for the outcome expenditure on goods, we present only the estimated CATEs for the levels of the variable study hours, as it was selected by the causal forest as important. We also checked for the variable enrolled in school, but the estimated CATEs did not seem to differ from another. In addition, this variable was not selected as important by the model.

Finally, in relation to both outcomes, the results also suggest that adolescent girls who believed tasks should be gender neutral or more equally split between genders, coded by the empowerment index at baseline, benefited more from the programme.

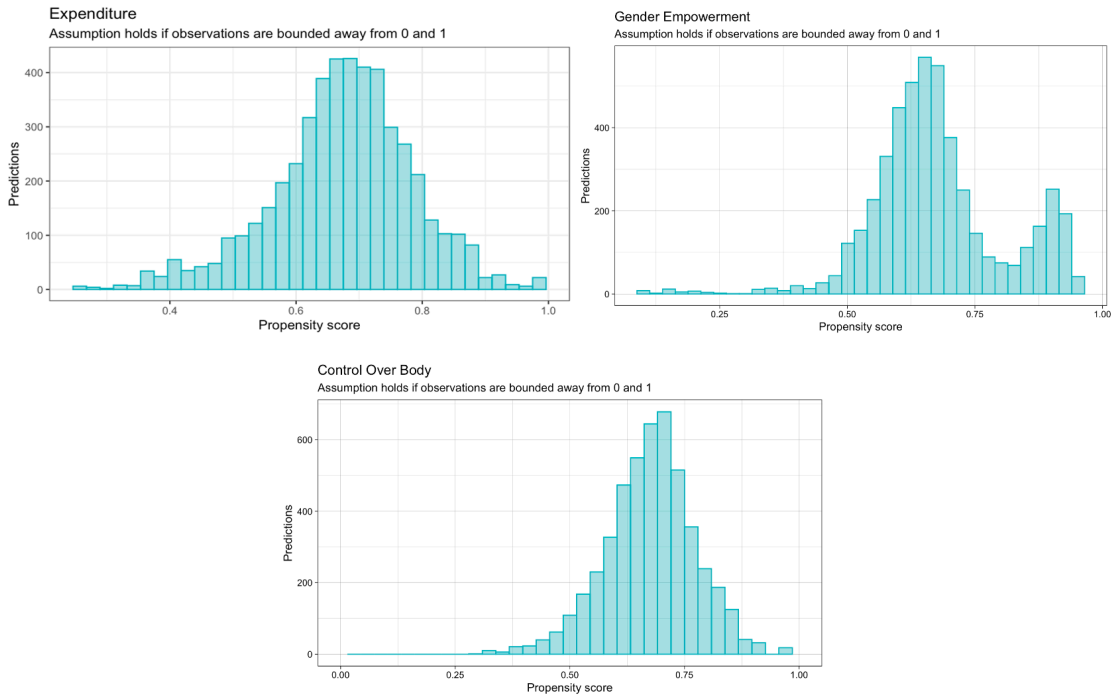
6.4. Evaluation Causal Forest Fit

Finally, as a way of validating our models, we discuss below the results of the evaluation of our main assumptions of the causal forest on the three outcomes.

6.4.1. Assessing the Overlap Assumption

Figure 7 presents histograms of the propensity scores for each outcome. The propensity scores for the three outcomes of interest are bounded away from 0 and 1 validating the overlap assumption which requires a positive probability of treatment for every individual in the sample.

Figure 7. Histogram of propensity scores by outcomes

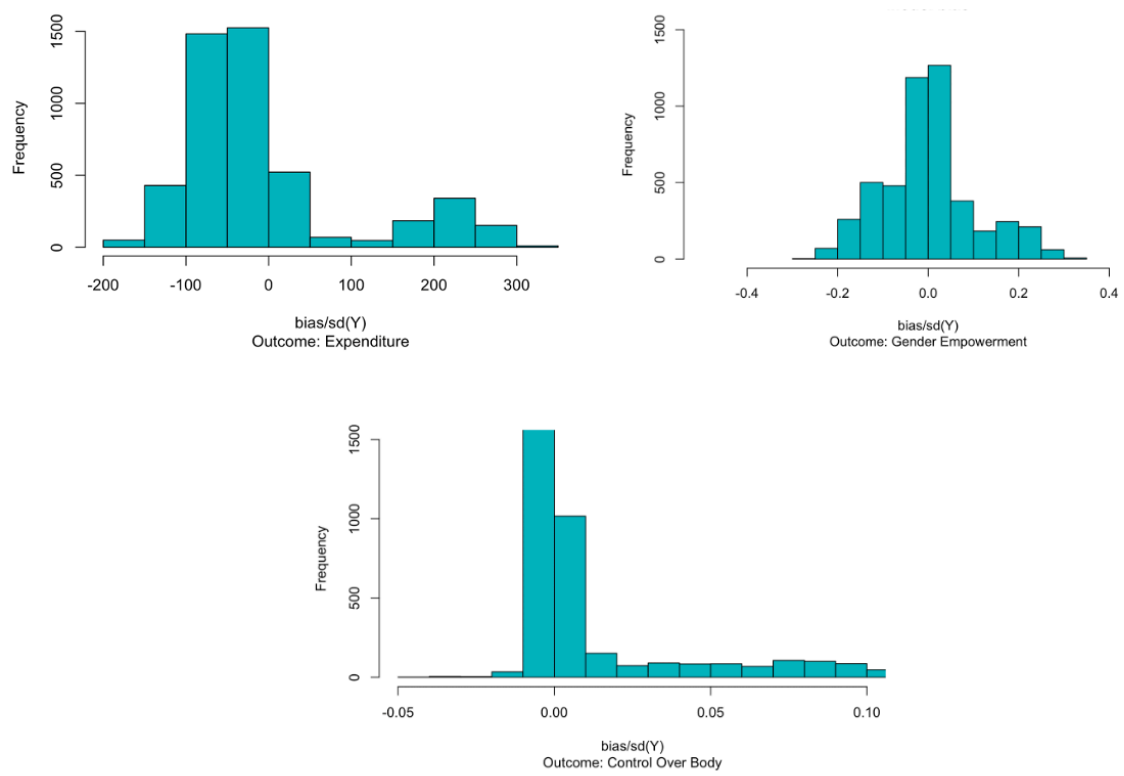


Note: Estimated propensity scores from the trained causal forest with selected variables. Each Xi should have a positive probability of treatment for the overlap assumption to hold.

6.4.2. Assessing bias of the causal forest predictions

The results of the bias estimates in Figure 8, scaled by the standard deviation of the outcome, suggest that there is not so much sparsity of the product of the propensity scores and the conditional expectations of being treated on the three outcomes (Athey et al., 2017).

Figure 8. Histogram of bias by outcomes



Note: Estimation of bias scaled by the standard deviation in each outcome.

7. Return on Investment (ROI)

Our analysis has proved the existence of heterogeneous treatment effects following the programme. However, these results do not yet provide evidence that would justify the targeting of the programme to subgroups who seem to have benefited more would actually result in higher productivity of the programme in the future. To make the argument that targeting practices should adapt to account for the heterogeneity detected, we calculated the ROI such a practice would have.

The ELA programme costs 17.9$ per eligible adolescent girl participating, for a two-year period. followed the original researchers who chose to focus on the second-year total cost per participating adolescent girl, because we also measure the effect as captured in the follow-up survey, two years after the baseline survey. This cost corresponds to 21% of the average annual consumption expenditure on the basket of goods measured by this variable (Bandiera et al., 2014). Thus, from a policymaking perspective, the programme would be desirable for re-implementation if the overall effects of the programme were to increase girls' expenditure at least by the same percentage. We monetised benefits from the programme by looking at the girls' increased monthly consumption expenditure at follow up.[47] The monthly consumption expenditure of adolescent girls can be understood as a proxy for the girls' earning. As we did not have direct information about earnings, this variable was the second-best option.[48]

We wish to show the benefit of targeting when considering for different subgroups' treatment effects. With the results we have obtained, however, we can only do so to a limited extent. To be specific, we are able to show that in principle targeting to subgroups results in higher return on investment. Yet, because we did not manage to detect significant variation across specific covariates, it remains somewhat unclear what characteristics the subgroups with the diverging effects have. Nevertheless, and to emphasise the importance of estimating and targeting while considering heterogeneous effects, we will proceed in the following manner.

We based the estimates on conversion rate in January 2008, as done in the original survey where 1$ was worth about 1,700UGX. Based on this we can deducted the estimated consumption

[47] See Table 1 for variable description.
[48] Monthly expenditure is a proxy for earning as in low-income settings it is hard to measure self-employment activities (Bandiera et al., 2014).

expenditure in the three ntiles in dollar terms. Considering the conversion rate and the estimated effects, we calculated the following consumption expenditure by year in dollars for the three ntiles, as shown in Table 12.

To estimate the ROI we followed the steps taken by Campos et al. (2017). Firstly, we assumed a discount rate of 10 percent. This means that overall benefits would depend on the timeframe in which ROI is presented. We then proceeded to calculate three different scenarios, with different assumptions regarding how long the programme's positive effects would last. In this way, we were able to illustrate several plausible ROI. The results of this exercise are found in Table12. In the first scenario, we assumed that there exist no benefits besides the ones measured in the follow up survey. Even in this scenario we see that the ROI is positive across all ntiles. An alternative scenario would be to anticipate positive programme effects to persist in the third and fourth year, but to a lesser extent. That is, by assuming expenditure gains would be half in the third year of what they were in the first two years and a quarter in the fourth year. Lastly, we considered a scenario in which gains persist for 8 years, whereby from the third year onwards they do so with 20 percent depreciation rate.

Looking at the estimated ROI, we first see that the positive effects of the programme translate to an increase in girls' consumption expenditure that is higher than 21%, meaning that it is desirable to scale the programme and implement it again. Zooming into ROI by ntiles, the differences in the ROI are rather striking. In the first scenario, if the programme would have been targeted to ntile 2, the subgroup with the highest increase in expenditure following the programme, it would have resulted in 30.9% higher returns compared to the group with the second highest estimated effect, ntile 1, and in a 42.9% lower returns to ntile 3 compared to ntile 2. In the second scenario, returns would be higher by 60% when targeting the programme to those who benefited more (ntile 2) relative to those who benefited least (ntile 3). Similarly, in the third scenario, targeting the programme to subgroup ntile 2 would translate to 103% higher expected ROI than if the programme were to be allocated to the subgroup that benefited the least (ntile 3). Lastly, we calculated what the yield would be when directing the programme to the highest ntile 3 instead of to the full sample, and this would result in 67% higher returns. All in all, targeting those who are expected to benefit more from a programme results in a sizable increase in future gains.

Table 14. Return on Investment

	Program ATE	Ntile 1	Ntile 2	Ntile 3
TOTAL COST				
Cost per eligible girl	17.9$	17.9$	17.9$	17.9$
EXPENDITURE IN DOLLARS				
One-year Expenditure (based on measured effects at follow-up)	33.8$	31.3$	55.4$	21.9$
RETURN ON INVESTMENT – 3 SCENARIOS				
Assuming no benefits following the two-year follow-up	42.2%	39.1%	70%	27.1%
Assuming benefits persist but compared to the first 2 years, they are ½ in the 3rd year and a ¼ in the fourth year	57.3%	52.8%	95.1%	37.1%
Assuming gains persist for 8 years but from the 3rd year onwards there is a 20% depreciation rate	102.5%	95.1%	169.9%	66.7%
BENEFIT OF TARGETING				
Increase in ROI if targeting compared to the overall ATE	-	- 7.4%	+ 67.5%	- 35.8%

8. Discussion

This study utilised a causal forest to estimate the heterogeneous treatment effects of an ELA programme designed to empower adolescent girls in Uganda. Our research followed two key lines of inquiry: the first was a methodological one exploring the potential and capacity of the causal forest to produce accurate estimates of treatment heterogeneity, while the second was a substantive one that aimed at achieving a deeper understanding of the workings of the ELA programme and its capacity to empower girls.

8.1. Methodological Considerations

On the methodological level, we show that a causal forest has significant advantages as compared to traditional linear models in several regards. For one, while the linear model has to be carefully specified and presented with clear pre-determined relations between covariates and outcomes, including how covariates interact with one another, a causal forest is much more flexible as it detects those relations in the splitting procedure and allows the inclusion of as many variables as possible.[49] Another advantage of causal forest is the data-driven approach for the selection of subgroups based on the splitting procedure. In linear models, determination of subgroups is derived only from theory, which may mean that some important subgroups would be overlooked.[50]

In addition, the results also reveal the ability of causal forests to credibly detect different ATEs of subgroups, but the model is very imprecise when subgroups are small. Further research should also explore interactions between different covariates. Although in principle this is already possible, because of the size of the subgroups in our dataset, constructing interactions would have left us with too little observations which would have resulted in meaningless analysis. The poor capacity of causal forest to detect effects with small numbers of observations is not a new finding (Athey, 2019). Our study implies that a dataset with approximately 5,000 observations is still too small for ML. This might be one of the reasons as to why our causal forest results show higher standard errors than the linear models' estimates, indicating that linear models may perform better than ML models in small samples settings.

[49] Another advantage is how the model treats the missing values of the X matrix. GRF accounts for the fact that missing values may be informative in that they may be predictive of unobserved sources of heterogeneity.

[50] Nevertheless, it can be the case that a variable did not weigh much in the causal forest splitting procedure but varies strongly with treatment. Therefore, even with causal forest it is important to account for theory when deciding on the analytical strategy.

One element of this finding suggests that if we are to use ML, and specifically the GRF package, we also need to amend how and which data we choose to collect. This may mean that in the first stage, the cost of performing an impact evaluation may increase because of the need and aim to have larger subgroup construction, depending on our research questions. On the face of it, this may seem too ambitious, but when accounting for the estimated ROI targeting based on heterogeneous effects, the sizable positive return may justify the increased initial costs. What we are suggesting is, in a sense, a feedback loop between the data we collect and the analysis we perform.

8.2. Substantive Results

There is a clear and significant variability in treatment effects depending on the branch implementing the ELA programme. This effect must be seriously considered when targeting the communities and the respective teams implementing the programme. It would be beneficial to try and learn what elements of implementation were responsible for driving these effects. Furthermore, the results of our findings provide clues to the underlying causal mechanisms that drive treatment effects. In this sense, it would appear that the baseline characteristics of the adolescent girls treated influence the impacts of the intervention.

Our results also suggest that girls from rural areas experienced smaller treatment effects compared to their urban counterparts. These findings hint at a diverging direction from the test in the original paper, which found homogeneous impacts across rural and urban areas for the indexes related to outcomes of economic empowerment and to girls' control over their bodies (Bandiera et al., 2014).[51] However, the likely different treatment effects among urban and rural girls who took part in the programme are consistent with the literature on women's empowerment, but not with the surveyed experts. In general, rural women face specific economic and social challenges associated with their environment. Thus, it is important that implementers consider the barriers rural women face when trying to foster better economic and social outcomes. Policymakers and NGOs should support rural women more actively to overcome the challenges they face so they can benefit the same as girls who live in urban areas.

[51] IGA, social empowerment using COB, and education using enrolment in school, study hours and if they want to go back to school.

Moreover, although our theoretical framework established that education could drive more positive effects of the results in relation to economic empowerment, our findings show the opposite direction: girls who invested less in education experienced more benefits from the programme. One potential confounder for this is the time spent in clubs. Indeed, we observe that girls who studied less also self-reported they attended the club more than girls who studied more.[52]

It is interesting that one third of the experts predicted that the programme would be less successful for girls who were cohabiting or married.[53] We do not observe such results. Our finding is aligned with empirical research that does not detect statistical differences among the effects of gender interventions in girls with and without partners (Ngezahoguhora, 2013; Obayelu & Chime, 2020). Here it is important to note that only 300 girls in the treated group had a partner, which means that even if heterogeneity existed, our causal forest model would have not been able to detect it. One clear finding from these results is that confidence intervals grow when the subgroups size in the data are small. Therefore, we cannot conclude much from this estimate.

Although we did not expect to find heterogeneous treatment effects on women with and without children, it seems that having a child can affect women's social empowerment regarding traditional gender roles. It is possible that vulnerable women with children have internalised their attitudes towards their role in social and private spheres, making it more challenging for the programme to relax those attitudes. However, there are no clues suggesting different treatment effects in women's economic empowerment depending on whether they have children or not. These results might be specific to women in vulnerable communities, since empirical studies in developed countries suggest that when women have a child, their professional trajectories might get interrupted, thereby increasing the gender gap.

Interestingly, our findings are also different from the original analysis that found largely homogenous effects across rich and poor households and young and older girls on the IGA and COB indexes (Bandiera et al., 2014). Higher wealth in the household of the girls seems to be a driver of higher positive effects of the treatment, and one third of the experts also agree on this

[52] In treatment status, 102 girls who studied between 0 and 55 hours reported attending the club sometimes (at least once a week), while 49 girls from this subgroup reported they attended the club frequently (more than three times a week). In contrast, only 74 girls who studied between 55 and 68 hours reported attending the club sometimes and 39 attended frequently. Girls who studied over 68 hours attended even less - 58 girls attended sometimes and 40 attended frequently.
[53] See Appendix C, Figure C. 6.

matter. As explained in our theoretical framework, girls with more resources in general are in a better position to make better strategic choices. On the other side, regardless of the uncertainty expressed by experts on how age would interact with treatment effects, we anticipated heterogeneity without indicating its direction. The results of several studies suggest that older girls can benefit more from training and skills associated with vocational programmes since they have more agency and knowledge about their preferences and professional trajectory. Our results signal that treatment effects were more positive as age increases, where the estimated CATE was smaller for girls between 10 and 15 years old relative to older girls.

Finally, because the treatment's data we analysed was based on voluntary participation, the extent to which baseline characteristics play a determinative role should be questioned. As one of the experts we surveyed said, "since this programme is 'opt in' - some of the criteria like having children will not necessarily lead to more or less positive effects from the programme, but it may lead girls not to opt in. For example, it is likely that a girl, after the age of 16 who is married is unlikely to attend, due to competing family duties and sense of appropriateness in certain contexts." Nevertheless, it is still valuable to assert heterogeneity among the participating girls. This comment is just admitting to the fact that in such a research design, it could be that another source of heterogeneity is unobserved and driven by the programme design.

9. Lessons Learned and Limitations

What can we deduce from the fact that our results seem to suggest the presence of heterogeneity, and show sizable expected returns if targeting was directed to those who benefited more? Especially because we conducted a causal analysis exercise and not a prediction task, the generalisability of our findings is not obvious (Williams, 2020). In any case, mapping the mechanisms in which the programme has worked to produce positive effects for the Ugandan girls is beneficial to decision makers. Nevertheless, this input can be important when considering re-implementation in a different setting, which necessitates programme adaptations (Williams, 2020), even within Uganda, and potentially in East Africa as well. One encouraging point is that the ELA programme and similar programmes were repeatedly positively evaluated across different contexts, which allows us to more freely discuss their value to policymakers operating in these countries (Bandiera et al., 2014).

Although we managed to find signals into the direction of heterogeneity in subgroups, these results fall short in producing clear cut evidence that can be transferable to other settings. Indeed, the estimates for heterogeneity across covariates are too imprecise. It should not, however, be discarded. For once, personnel in BRAC or other organisation running the ELA can use our finding as signals as to which mechanisms of their programme they should further improve and adapt. Here, the results are informative to set the grounds for future testing of assumptions on how the programme works. For example, the programme assumes that targeting young adolescent girls is crucial as they are transitioning from dependence on their parents to independence or reliance on men (Bandiera et al., 2014). Seeing that the trend of heterogeneity suggests more benefits to older girls, the assumption that younger girls would benefit more should be put to the test. However, as we did not perform subgroup analysis with interactions, it is hard to tell with certainty if when considering the whole profile of the participating girls, the results would stay the same.

Another important finding is that targeting to subgroups who are likely to have larger treatment effect results in a significantly better return on investment, even in the most conservative scenario where benefits of the programme would be seen after two years. Knowing that provides clear incentive to try and further develop this type of analysis. As the causal forest enables causal inference also for observational studies, further research can try and leverage large quantities of administrative data for the task of detecting or predicting heterogeneous effects. Based on our

analysis we perceive this to be a desirable practice, as it has the potential to improve targeting and to yield larger returns, both to taxpayers and to participating individuals.

To conclude, the evidence we present in this thesis, contributes to the growing evidence on the promising capacity of causal forest to produce credible, data-driven estimates of heterogeneous treatment effects. However, it also shows that it is still underdeveloped, in particular with regard to applications on RCT data with smaller numbers of observations. Therefore, we believe that applications should focus on program evaluations with larger datasets, perhaps using administrative data and observational type inference.

References

Adams, A. V. (2008). Skills Development in the Informal Sector of Sub-Saharan Africa. In *Skills Development in the Informal Sector of Sub-Saharan Africa*. https://doi.org/10.1596/28120

Athey, S. (2019, November). *Estimation of Heterogeneous Treatment Effects and Optimal Treatment Policies*. Continuing Education Webcasts - American Economic Association (AEA). https://www.aeaweb.org/conference/cont-ed/2018-webcasts

Athey, S. (2020, October 9). *Estimating Heterogeneous Treatment Effects*.

Athey, S. (2015). *Machine Learning and Causal Inference for Policy Evaluation*. 5–6. https://doi.org/10.1145/2783258.2785466

Athey, S., & Imbens, G. (2015a). Machine Learning Methods for Causal Effects. *ArXiv Preprint ArXiv:1504.01132*.

Athey, S., & Imbens, G. (2015b). Machine Learning Methods for Estimating Heterogeneous Causal Effects. *ArXiv*, *October 2013*.

Athey, S., & Imbens, G. (2015c). *Recursive Partitioning for Heterogeneous Causal Effects*. http://arxiv.org/abs/1504.01132

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges. *American Economic Review*, *107*(5). https://doi.org/10.1257/aer.p20171042

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, S., Imbens, G. W., & Wager, S. (2016). Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing. *ArXiv*, *June*.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *80*(4). https://doi.org/10.1111/rssb.12268

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Athey, S., & Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. In *ArXiv*.

Baiardi, A., & Naghi, A. (2021). The Value Added of Machine Learning to Causal Inference: Evidence from Revisited Studies. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3759867

Baird, S., Chirwa, E., de Hoop, J., & Ozler, B. (2014). Girl power: Cash transfers and adolescent welfare. Evidence from a Cluster-randomized Experiment in Malawi. *African Successes: Health and Gender*.

Bandiera, O., Buehren, N., Burgess, R., Goldstein, M., Gulesci, S., Rasul, I., & Sulaiman, M. (2014). *Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa*. https://doi.org/10.1257/app.20170416

Bandiera, O., Buehren, N., Burgess, R., Goldstein, M., Gulesci, S., Rasul, I., & Sulaiman, M. (2020). Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa. *American Economic Journal: Applied Economics*, *12*(1). https://doi.org/10.1257/app.20170416

Bantebya, G. K., Muhanguzi, F. K., & Watson, C. (2014). Adolescent Girls in the Balance: Changes and Continuity in Social Norms and Practices around Marriage and Education in Uganda. *Overseas Development Institute: London, UK.*

Battioui, C., Shen, L., & Ruberg. (2014). *A Resampling-based Ensemble Tree Method to Identify Patient Subgroups with Enhanced Treatment Effect*.

Bhaskaran, K., & Smeeth, L. (2014). What is the Difference Between Missing Completely at Random and Missing at Random? *International Journal of Epidemiology*, *43*(4). https://doi.org/10.1093/ije/dyu080

Bhattacharya, D., & Dupas, P. (2012). Inferring Welfare Maximizing Treatment Assignment under Budget Constraints. *Journal of Econometrics*, *167*(1). https://doi.org/10.1016/j.jeconom.2011.11.007

Breiman, L. (2001). *Random Forests* (Vol. 45).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and Regression Trees. In *Classification and Regression Trees*. https://doi.org/10.1201/9781315139470

Bruce, J., & Hallman, K. (2008). Reaching the Girls Left Behind. *Gender and Development*, *16*(2). https://doi.org/10.1080/13552070802118149

Buchwald, D. (2019). *Comparing Methods to Estimate Personalised Treatment Effects from Observational Data*.

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D., & Mensmann, M. (2017). Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa. *Science*, *357*(6357). https://doi.org/10.1126/science.aan5329

Carbo-Valverde, S., Cuadros-Solas, P., & Rodríguez-Fernández, F. (2020). A Machine Learning Approach to the Digitalization of Bank Customers: Evidence from Random and Causal Forests. *PLoS ONE*, *15*(10 October). https://doi.org/10.1371/journal.pone.0240362

Carnegie, N., Dorie, V., & Hill, J. (2019). *Examining Treatment Effect Heterogeneity using BART*.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. In *Observational Studies* (Vol. 5).

Chang, W., Díaz-Martin, L., Gopalan, A., Guarnieri, E., Jayachandran, S., & & Walsh, C. (2020). *What Works to Enhance Women's Agency: Cross-Cutting Lessons from Experimental and Quasi-Experimental Studies*.

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring. *Biometrics*, *73*(4). https://doi.org/10.1111/biom.12676

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-val, I. (2018). *Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments*.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian Ensemble Learning. *Advances in Neural Information Processing Systems*. https://doi.org/10.7551/mitpress/7503.003.0038

Coady, D., Grosh, M. E., & Hoddinott, J. (2004). *Targeting of Transfers in Developing Countries Review of Lessons and Experience*.

Coles, A., Gray, L., & Momsen, J. (2015). The Routledge Handbook of Gender and Development. In *The Routledge Handbook of Gender and Development*. https://doi.org/10.4324/9780203383117

Cornwall, A. (2016). Women's Empowerment: What Works? *Journal of International Development*, *28*(3). https://doi.org/10.1002/jid.3210

Davis, J., & Heller, S. (2018). *Rethinking the Benefits of Youth Employment Programs- The Heterogeneous Effects of Summer Jobsú*.

Dehejia, R. H. (2005). Program Evaluation as a Decision Problem. *Journal of Econometrics*, *125*(1-2 SPEC. ISS.). https://doi.org/10.1016/j.jeconom.2004.04.006

Do, M., & Kurimoto, N. (2012). Women's Empowerment and Choice of Contraceptive Methods in Selected African Countries. *International Perspectives on Sexual and Reproductive Health*, *38*(1). https://doi.org/10.1363/3802312

Duflo, E. (2012). *Women Empowerment and Economic Development*.

Duflo, E. (2020). Field Experiments and the Practice of Policy. *American Economic Review*, *110*(7), 1952–1973. https://doi.org/10.1257/aer.110.7.1952

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, *101*(5). https://doi.org/10.1257/aer.101.5.1739

Elek, P., & Bíró, A. (2021). Regional Differences in Diabetes Across Europe – Regression and Causal Forest Analyses. *Economics and Human Biology*, *40*. https://doi.org/10.1016/j.ehb.2020.100948

Fang, A. (2020, May 23). *10 Things to Know About Heterogeneous Treatment Effects*. EGAP.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup Identification from Randomized Clinical Trial Data. *Statistics in Medicine*, *30*(24). https://doi.org/10.1002/sim.4322

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local Linear Forests. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2020.1831930

Frölich, M., & Melly, B. (2013). Unconditional Quantile Treatment Effects Under Endogeneity. *Journal of Business and Economic Statistics*, *31*(3). https://doi.org/10.1080/07350015.2013.803869

Glennerster, R., Walsh, C., Diaz-martin, L., FAO and The World Bank, Cook, J., Kabubo-mariara, J., & Kimuyu, P. (2018). A Practical Guide to Measuring Women's and Girls' Empowerment in Impact Evaluations. *J-Pal*.

Green, D. P., & Kern, H. L. (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, *76*(3), 491–511. https://doi.org/10.1093/poq/nfs036

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, *20*(1). https://doi.org/10.1198/jcgs.2010.08162

Hirano, K., & Jack R. Porter. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, *77*(5). https://doi.org/10.3982/ecta6630

Hunt, A., & Samman, E. (2016). Women's Economic Empowerment: Navigating Enablers and Constraints. *Overseas Development Institute Research Report*, *September*.

Ibanez, M., Khan, S., Minasyan, A., Sahoo, S., & Balasubramanian, P. (2017). PROTOCOL: The Impacts of Interventions for Female Economic Empowerment at the Community Level on Human Development: a Systematic Review of the Evidence in Low- and Middle-Income Countries. *Campbell Systematic Reviews*, *13*(1). https://doi.org/10.1002/cl2.183

Imai, K., & Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Annals of Applied Statistics*, *7*(1). https://doi.org/10.1214/12-AOAS593

Imbens, G. W., & Rubin, D. B. (2015). Causal inference: For Statistics, Social, and Biomedical Sciences an Introduction. In *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. https://doi.org/10.1017/CBO9781139025751

Johansson, F., Shalit, U., & Sontag, D. (2016). Learning Representations for Counterfactual Inference. *International Conference on Machine Learning*.

Kabeer, N. (1991). The Quest for National Identity: Women, Islam and the State in Bangladesh. *Feminist Review*, *37*. https://doi.org/10.2307/1395470

Kelly, J. M. (1992). *A Short History of Western Legal Theory CLARENDON PRESS • OXFORD*. http://www.w3.org/TR/images/latin1.gif

Kent, D. M., Steyerberg, E., & van Klaveren, D. (2018). Personalized Evidence-based Medicine: Predictive Approaches to Heterogeneous Treatment Effects. In *BMJ (Online)* (Vol. 363). https://doi.org/10.1136/bmj.k4245

Kleven, H., Landais, C., & Søgaard, J. E. (2019). Children and Gender Inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, *11*(4). https://doi.org/10.1257/app.20180010

Klugman, J., Hanmer, L., Twigg, S., Hasan, T., McCleary-Sills, J., & Santamaria, J. (2014). Voice and Agency: Empowering Women and Girls for Shared Prosperity. In *Voice and Agency: Empowering Women and Girls for Shared Prosperity*. https://doi.org/10.1596/978-1-4648-0359-8

Knaus, M. C., Lechner, M., & Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. In *arXiv*. https://doi.org/10.1093/ectj/utaa014

Kuenzel, S. R. (2019). *Heterogeneous Treatment Effect Estimation using Machine Learning*.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. In *arXiv*.

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple Hypothesis Testing in Experimental Economics. *Experimental Economics*, *22*(4). https://doi.org/10.1007/s10683-018-09597-5

Luo, X., Lu, X., & Li, J. (2019). When and How to Leverage E-commerce Cart Targeting: The Relative and Moderated Effects of Scarcity and Price Incentives with a Two-stage Field Experiment and Causal Forest Optimization. *Information Systems Research*, *30*(4). https://doi.org/10.1287/isre.2019.0859

Malhotra, A., Schuler, S. R., & Boender, C. (2002). Conceptualizing and Measuring Women's Empowerment as a Variable in International Development. *Paper Prepared for the World Bank Workshop on Poverty and Gender: New Perspectives*.

Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, *72*(4). https://doi.org/10.1111/j.1468-0262.2004.00530.x

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91. https://doi.org/10.1257/app.20170299

Montanari, B., & Bergh, S. I. (2019). A Gendered Analysis of the Income Generating Activities under the Green Morocco Plan: Who Profits? *Human Ecology*, *47*(3). https://doi.org/10.1007/s10745-019-00086-8

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Ngezahoguhora, H. (2013). *Women Empowerment Through Income Generating Activities. The Promotion of Women's Rights In Rwanda*.

Nie, X., & Wager, S. (2017). Quasi-oracle Estimation of Heterogeneous Treatment Effects. In *arXiv*. https://doi.org/10.1093/biomet/asaa076

Obayelu, O. A., & Chime, A. C. (2020). Dimensions and Drivers of Women's Empowerment in Rural Nigeria. *International Journal of Social Economics*, *47*(3). https://doi.org/10.1108/IJSE-07-2019-0455

Okumo, A. A. (1992). *Participation of women in income-generating Activities: a Case Study of Dandora, Nairobi, Kenya*.

Osler, S. W. (1939). The Principles and Practice of Medicine. Designed for the Use of Practitioners and Students of Medicine. *The American Journal of the Medical Sciences*, *197*(3). https://doi.org/10.1097/00000441-193903000-00013

Panda, S. M. (2000). Women's Empowerment Through NGO Interventions: A Framework for Assessment. *Social Change*, *30*(3–4). https://doi.org/10.1177/004908570003000405

Panigrahi, S., & Taylor, J. (2018). Scalable Methods for Bayesian Selective Inference. *Electronic Journal of Statistics*, *12*(2). https://doi.org/10.1214/18-EJS1452

Pearl, J., & Mackenzie, D. (2018). *The Book of Why. The New Science of Cause and Effect*.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions. *Statistics in Medicine*, *37*(11). https://doi.org/10.1002/sim.7623

Qian, M., & Murphy, S. A. (2011). Performance Guarantees for Individualized Treatment Rules. *The Annals of Statistics*, *39*(2). https://doi.org/10.1214/10-aos864

Razavi, S. (2016). The 2030 Agenda: Challenges of Implementation to Attain Gender Equality and Women's Rights. *Gender and Development*, *24*(1), 25–41. https://doi.org/10.1080/13552074.2016.1142229

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients when Some Regressors are Not Always Observed. *Journal of the American Statistical Association*, *89*(427). https://doi.org/10.1080/01621459.1994.10476818

Schuck, P. H., & Zeckhauser, R. J. (2005). *Targeting in Social Programs Avoiding Bad Bets, Removing Bad Apples*.

Schwab, P., Linhardt, L., & Karlen, W. (2018). Perfect match: A simple Method for Learning Representations for Counterfactual Inference with Neural Networks. In *arXiv*.

Sen, A. K. (1999). Introduction - Development as Freedom. *Development as Freedom*.

Sheheli, S. (2012). *Improving Livelihood of Rural Women through Income Generating Activities in Bangladesh*.

Stoffi, F. J. B., & Gnecco, G. (2020). Causal Tree with Instrumental Variable: an Extension of the Causal Tree Framework to Irregular Assignment Mechanisms. *International Journal of Data Science and Analytics*, *9*(3). https://doi.org/10.1007/s41060-019-00187-z

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, *109*(508). https://doi.org/10.1080/01621459.2014.951443

Tibshirani, R. (2011). Regression Shrinkage and Selection via the Lasso: A Retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *73*(3). https://doi.org/10.1111/j.1467-9868.2011.00771.x

UNESCO. (2014). UNESCO Roadmap for Implementing the Global Action Programme on Education for Sustainable Development. In *Unesco*.

United Nations. (2021). *Goal 5| Department of Economic and Social Affairs*. United Nations. https://sdgs.un.org/goals/goal5

Varadhan, R., & Seeger, J. D. (2013). Estimation and Reporting of Heterogeneity of Treatment Effects. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*.

Vollmer, S., Khan, S., Tu, L. T. N., Pasha, A., & Sahoo, S. (2017). PROTOCOL: The Effect of Interventions for Women's Empowerment on Children's Health and Education: A Systematic Review of Evidence from Low- and Middle-Income Countries. *Campbell Systematic Reviews*, *13*(1). https://doi.org/10.1002/cl2.172

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Williams, M. J. (2020). External Validity and Policy Adaptation: From Impact Evaluation to Policy Design. *The World Bank Research Observer*, *35*(2), 158–191. https://doi.org/10.1093/wbro/lky010

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, *14*(1). https://doi.org/10.1007/s10115-007-0114-2

Young, K. (1993). Planning Development with Women: Making a World of Difference. *Choice Reviews Online*, *30*(11). https://doi.org/10.5860/choice.30-6252

Zhao, Q., Small, D. S., & Ertefaie, A. (2017). *Selective Inference for Effect Modification via the Lasso*.

Appendix A: The Empowerment and Livelihood for Adolescents (ELA) Programme

We based our analysis on impact evaluation of the ELA, powered by the NGO BRAC and implemented in Uganda. This programme aimed to "break the vicious circle between low labour force participation and high fertility by kick-starting human capital accumulation along two dimensions through the provision of i) vocational skills to enable adolescent girls to start small-scale income-generating activities; ii) life skills to build knowledge enabling girls to make informed choices about sex, reproduction and marriage" (Bandiera et al., 2014, p.3).

The programme was made available to adolescent development clubs, meaning that young female drop-outs could also attend, as well as women enrolled in school. In general, these clubs offered a protected space where girls could socialise, take part in various recreational activities and privately raise issues of concern. The clubs were usually open five afternoons a week in such a way that girls who attended full-day school could nevertheless participate in the club's activities. The clubs had female mentors who were slightly older than the target participants. Importantly, participation was voluntary and independent of participation in other initiatives of BRAC. The take-up rate was 21% of the target population of adolescent girls aged between 14 and 20.

The vocational skills training aimed at strengthening the economic empowerment of the participants. It focused on helping the adolescent girls to set up their own small-scale enterprise, but it also taught the participants a range of income-generating skills, such as "hair-dressing, tailoring, computing, agriculture, poultry rearing and small trades operation" (Bandiera et al., 2014, p.6). The matching of the girls to a respective activity considered their education level and the local surroundings' demand for such services and business practices. To complement this, the programme integrated financial literacy courses about budgeting, negotiations, accounting and more (p.7). The vocational skills part of the programme were instructed by "hired professionals as well as BRAC's own agriculture and livestock programme staff" (p.7).

The life skills training aimed at empowering girls by developing the control they hold over their lives and bodies. Thus, training sessions covered topics such as "sexual and reproductive health, menstruation and menstrual disorders, pregnancy, sexually transmitted infections, HIV/AIDS awareness, family planning and rape." In addition to that, other sessions focused on negotiations and conflict resolution, as well as leadership and management skills and legal information relating to women's issues (e.g., child marriage, violence against women and bride prices). The life skills training was delivered by pre-trained female mentors and/or professional staff of the NGO.

Appendix B: Survey to Experts

Dear Expert,
You are invited to participate in this survey aimed at gaining your insights on how you would estimate which adolescent girls are likely to benefit from an after-school club programme to improve their vocational and life skills. Specifically, you will be asked to estimate which profiles of adolescent girls are more likely to be receptive to a vocational and life skills programme. The outcomes of this survey will be used to compare experts' predictions with machine learning predictions, with the goal of improving policymakers' capacity to allocate such programmes in an informed manner. The research aims to improve the capacity of policymakers to target skill programmes. It will take 5 minutes to complete the questionnaire. Your participation in this study is completely voluntary and there are no foreseeable risks associated with this project. However, if you feel uncomfortable answering any questions, you can withdraw from the survey at any point. Your personal information will not be published and data from this research will be reported only in the aggregate and according to the General Data and Protection Regulation (GDPR). If you have questions at any time about the survey or the procedures of our Master thesis at the Hertie School, please don't hesitate to contact Ofer Dotan at o.dotan@mpp.hertie-school.org or Mariana Saldarriaga at m.saldarriaga@mpp.hertie-school.org. Thank you very much for your time and support. Please start the survey by clicking on the Continue button below.

Please read the description of the programme and answer the following questions based on this information.

The programme is implemented mostly in East Africa in after-school youth clubs. It aims to combat low labour force participation and high fertility by promoting human capital accumulation focusing on two dimensions: i) Vocational skills to enable adolescent girls to start small-scale income-generating activities; The programme focuses on helping the adolescent girls to set up their own small-scale enterprise. Additionally, it teaches the participants a range of income-generating skills, such as hairdressing, tailoring, computing, agriculture, poultry rearing and small trades operation ii) Life skills to build knowledge enabling girls to make informed choices about sex, reproduction and marriage Training sessions covered topics such as sexual and reproductive health, menstruation and menstrual disorders, pregnancy, sexually transmitted infections, HIV/AIDS awareness, family planning and rape. Next to that, other sessions focused on negotiations and conflict resolution, as well as leadership and management skills and legal information relating women's issues (e.g. child marriage, violence against women and bride prices). Descriptions modified from source 1 (listed at the end of the survey)

The programme measures impact on three main outcomes:
1. Economic empowerment.
Measured by the acquisition of the business and entrepreneurial skills, engagement in income-generating activities, scores on GEI.
2. Control over body.
Measured by marriage behaviour, childbearing, sexual violence, knowledge on contraception
3. Aspirations related to marriage and childbearing.
Attitudes toward family planning, e.g., appropriate age for marriage, childbearing

How do you anticipate each baseline characteristic would influence the effectiveness of the skills programme on a treated adolescent girl's (1) economic empowerment? ((1) Measured by the acquisition of the business and entrepreneurial skills, engagement in income-generating activities, scores on gender empowerment index)

	Negative effect	**No effect**	**Positive effect**	**Not sure**
Married or cohabitating	❑	❑	❑	❑
Has one or more children	❑	❑	❑	❑
Lives in a rural area	❑	❑	❑	❑
Enrolled in school	❑	❑	❑	❑
Subjectively satisfied with her earnings	❑	❑	❑	❑
Older than 16 when programme starts	❑	❑	❑	❑
Attends school regularly	❑	❑	❑	❑
Rarely attends school	❑	❑	❑	❑
Has more traditional gender roles attitudes	❑	❑	❑	❑
Relatively rich	❑	❑	❑	❑

How do you anticipate each baseline characteristic would influence the effectiveness of the skills programme on a treated adolescent girl's (2) Aspirations Related to Marriage and Child-Bearing? ((2) Measured by marriage behaviour, childbearing, sexual violence, knowledge on contraception)

	Negative effect	**No effect**	**Positive effect**	**Not sure**
Married or cohabitating	❑	❑	❑	❑
Has one or more children	❑	❑	❑	❑
Lives in a rural area	❑	❑	❑	❑
Enrolled in school	❑	❑	❑	❑
Subjectively satisfaction with her earnings	❑	❑	❑	❑
Older than 16 when programme starts	❑	❑	❑	❑
Attends school regularly	❑	❑	❑	❑

Rarely attends school	❑	❑	❑	❑
Has more traditional gender roles attitudes	❑	❑	❑	❑
Relatively rich	❑	❑	❑	❑

How do you anticipate each baseline characteristic would influence the effectiveness of the skills programme on a treated adolescent girl's (3) control over their bodies? ((3) Attitudes toward family planning, e.g. appropriate age for marriage, childbearing)

	Negative effect	**No effect**	**Positive effect**	**Not sure**
Married or cohabitating	❑	❑	❑	❑
Has one or more children	❑	❑	❑	❑
Lives in a rural area	❑	❑	❑	❑
Enrolled in school	❑	❑	❑	❑
Subjectively satisfied with her earnings	❑	❑	❑	❑
Older than 16 when programme starts	❑	❑	❑	❑
Attends school regularly	❑	❑	❑	❑
Rarely attends school	❑	❑	❑	❑
Has more traditional gender roles attitudes	❑	❑	❑	❑
Relatively rich	❑	❑	❑	❑

On a scale from 1-Strongly disagree to 100-Strongly agree (50-Undecided), how fair is it to prioritise those who are likely to benefit most from a skill programme compared to those who are most in need?

	❑

Personal Information
What is your name?

What is your profession?

What organisation/s are you affiliated with?

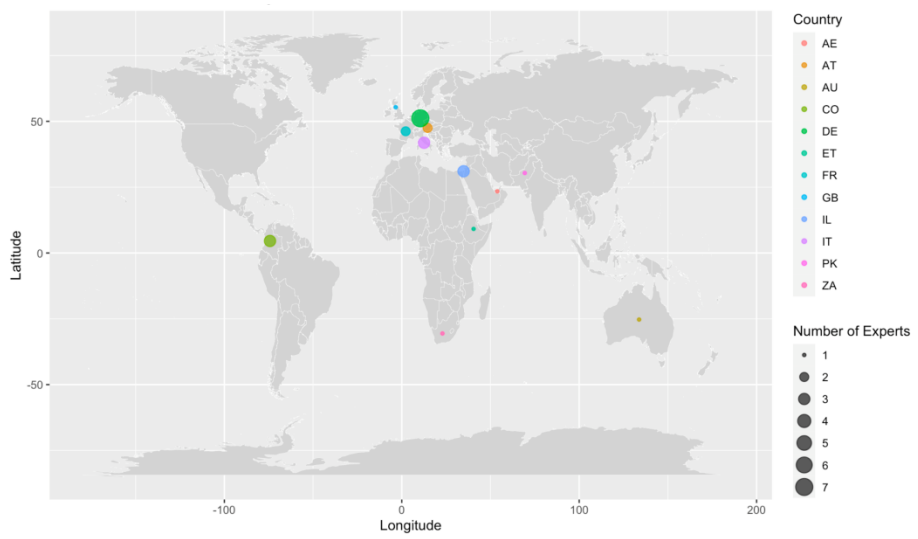
What is your job position?

You have answered all the questions. Do you have any additional comments? If not, please press "submit"

Appendix C: Survey Analysis

In sum, we surveyed 26 experts from twelve different countries, the majority of them residing in Europe.

Figure C. 1. Experts' countries of residence



Most experts provided us with information about their professional affiliation. The variety of organisations they represent confirms that we managed to reach relevant institutions that have direct experience with the topics discussed in the survey. This supports our approach of attempting to elicit high-quality answers from top practitioners rather than large numbers of perhaps less specialised responses.

Figure C. 2. Experts' professional affiliation

AFD	1
Colombia's Presidential Council for the Youth	1
GIZ	6
Google	1
Government organization	1
Hebrew university jerusalem	1
Hertie school	2
J-PAL Africa	3
Jurisdicción especial para la paz - Colombia	1
Peaceplayers International	1

General remarks

Experts expressed caution and noted that a key component that may be overlooked is participants' motivation to engage with the opportunities associated with the programme. They also expressed considerable uncertainty as to how characteristics will affect the treatment's success, saying that specific programme design and implementation capacity are sometimes more decisive to the programme's success than participants' profiles. For example, one expert from the World Bank DIME department noted that effects are dependent on "whether it is a programme based on cognitive versus behavioural models, whether problem-solving tools are provided and whether the programme has a follow-up". One expert from J-PAL, for example, pointed out that girls may not opt in to the programme if they have competing family obligations or very traditional social norms. Further, on the principled question about the merits of targeting, one practitioner commented that although targeting those who are likely to benefit most is desirable, vulnerability should also be considered "so as to avoid both excessive social homogeneity in groups and exclusion of vulnerable people from the programme".

Key findings

- While the majority of experts (30%) answered that **having children** would not have any effect on programme outcomes, others anticipated both negative and positive effects.
- **Being married** was perceived by most experts (30%) to hinder the treatment effect, as well as **holding traditional values,** although 24% were uncertain what the effect would be and some 20% actually thought it could result in a more positive effect of the treatment.
- Most experts (33%) also predicted that girls from **rural areas** would have more positive treatment

effect, while a considerable number of them (29%) was of the opinion that whether girls come from rural or urban locality would not impact programme effects at all.

- Across all outcomes, 40% of the experts agreed that **school enrolment** would have a positive impact on programme effectiveness. However, regarding school attendance, there was more variation, with 30% of experts' anticipating that rarely attending school would have either positive or no effect on programme outcomes.
- Regarding the effect of girls' **economic status**, experts had diverging views that reflect uncertainty. On the one hand, 28% of experts anticipated the positive impact of being rich or being satisfied with one's income. On the other hand, a comparable number of experts were not sure what the effect would be or predicted that having relatively more wealth would have no effect on the programme's outcomes.
- Finally, 33% of experts thought that relatively **older adolescent girls** were likely to benefit more from the programme. However, here too we see that 28% anticipated that age would not change treatment effect.

Figure C. 3. Estimated effect of Income

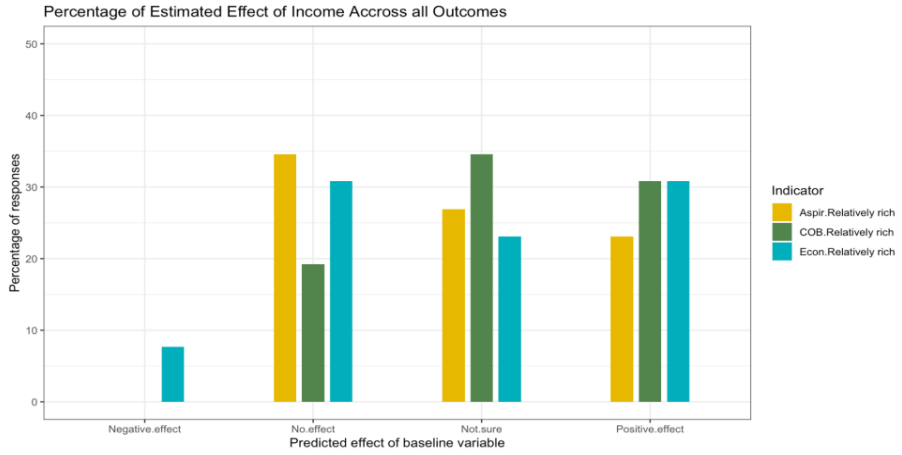


Figure C. 4. Estimated effect of satisfaction with income

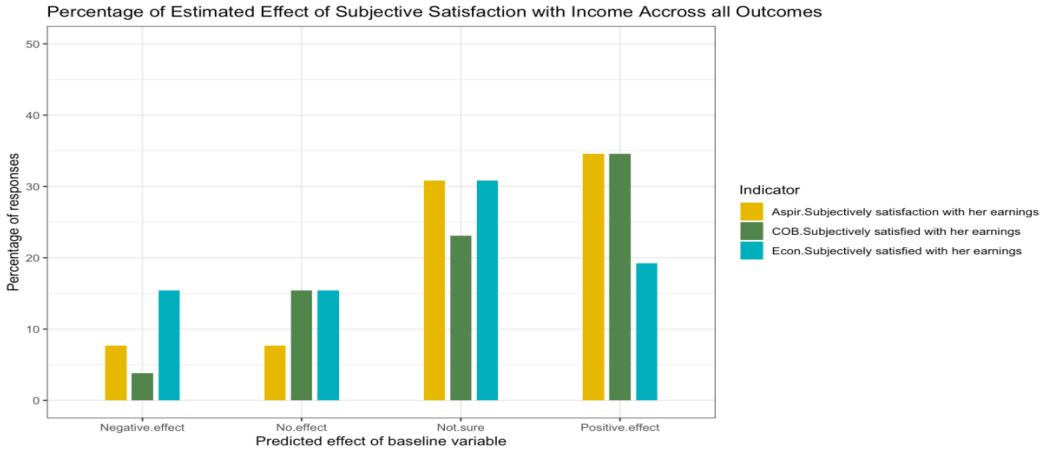


Figure C. 5. Estimated effect of school enrollment

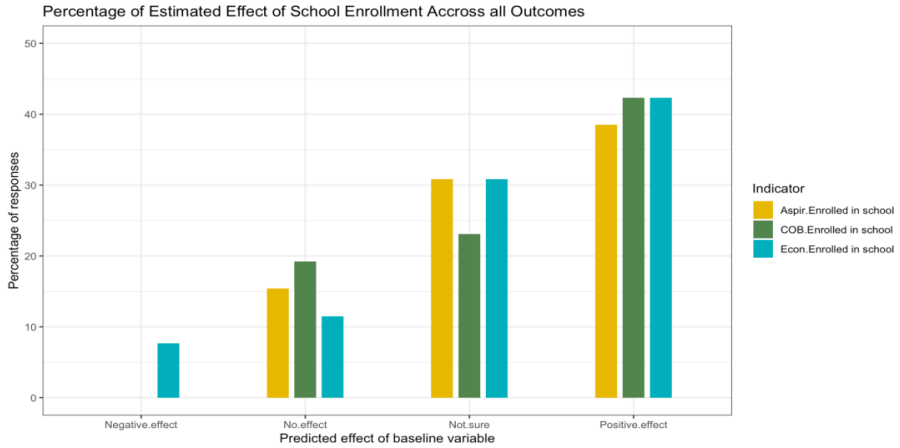


Figure C. 6. Estimated effect of Marriage

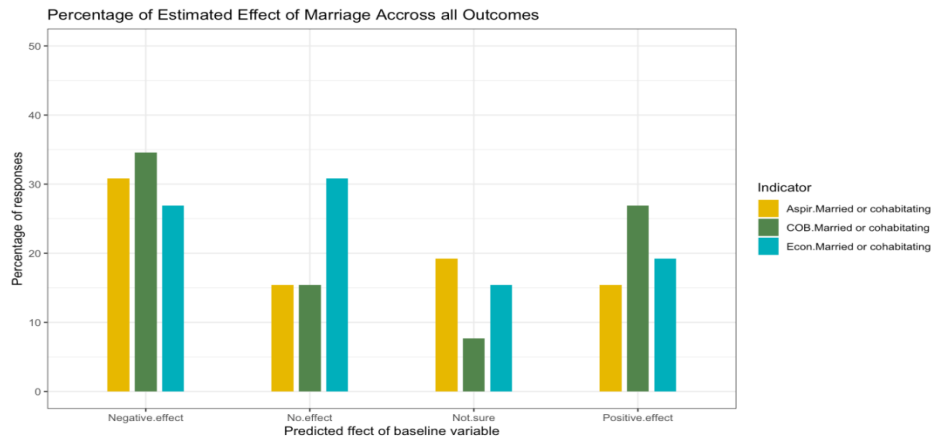


Figure C. 7. Estimated effect of having children

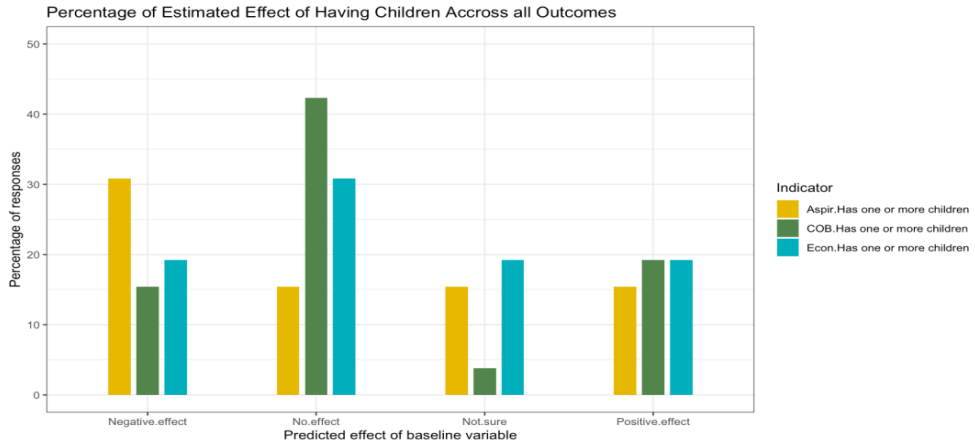


Figure C. 8. Estimated effects of Age

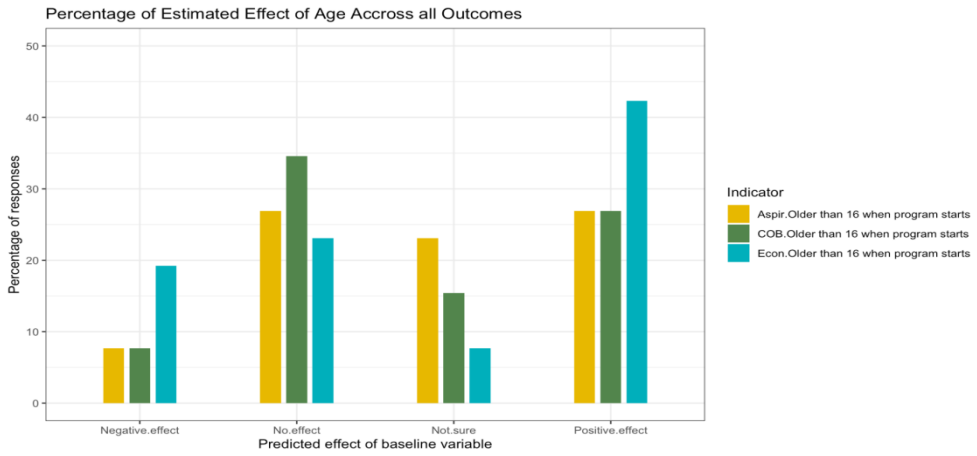


Figure C. 9. Estimated effect of living in rural area

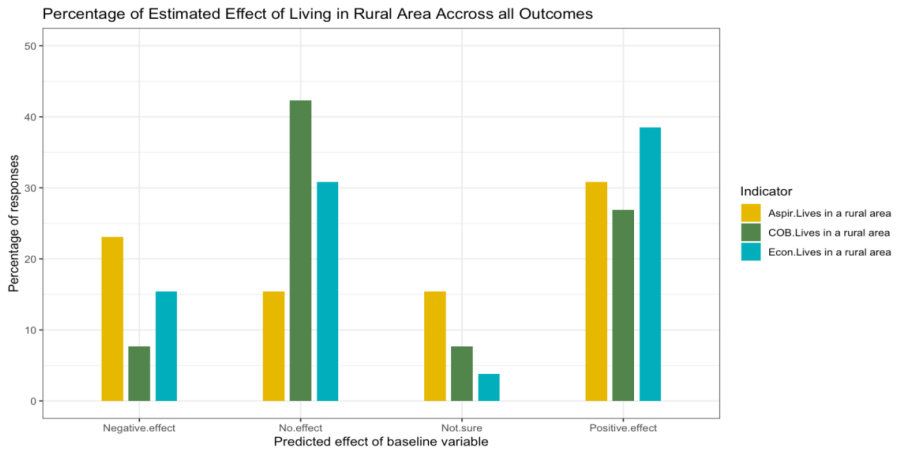
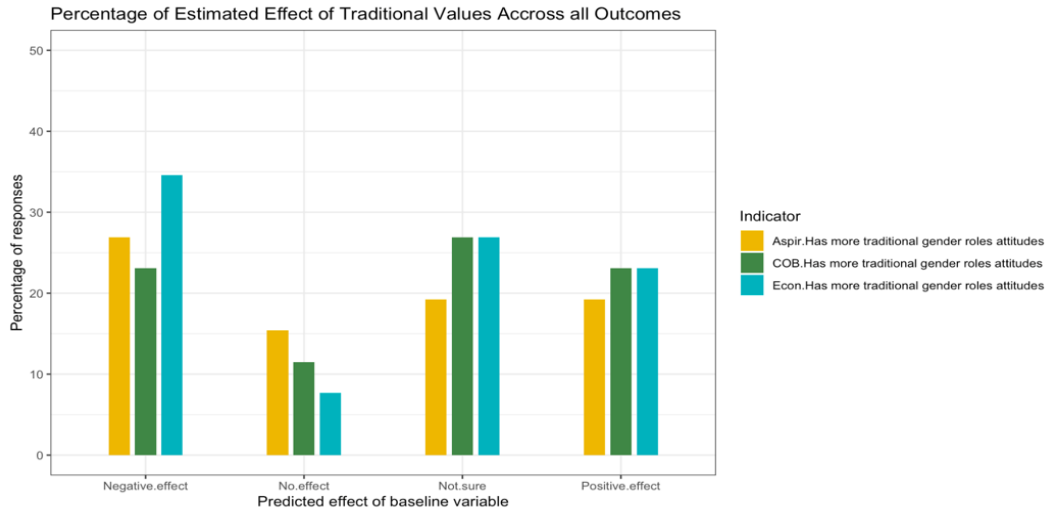
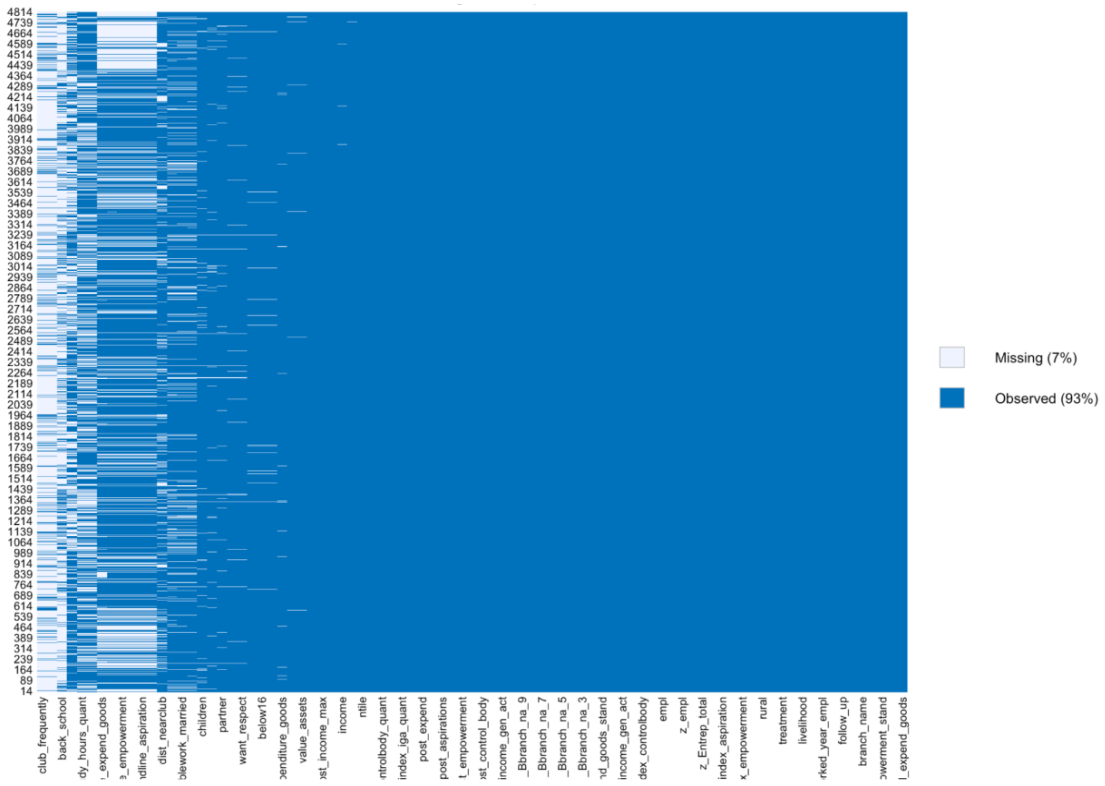


Figure C. 10. Estimated effect of having traditional gender roles attitudes



Appendix D: Missing Observations Map

Figure D. 1. Amalia missingness map



Appendix E: Tables of Selected "Important" Variables

Table E. 1 List of Selected Variables. Outcome: Expenditure on Goods

	Importance
children1	40
branchno3	38
age_quant18 to 28	24
expenditure_goods	23
partner1	22
rural1	18
rich1	17
study_hours_quant(55,68]	15
study_hours_quant(68,166]	14
worry_job1	13
index_empowerment_quant(14.3,42.9]	4
satisf_income1	2

Note:
Causal forest determines variable importance by a weighted sum capturing the number of splits a feature had in every depth layer in the model. Presented in descending order are variables that were above the mean weight of all the variables used for splitting in the model.The outcome Expenditure is coded as consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$.

Table E. 2 List of Selected Variables. Outcome: Gender Empowerment

	sort_var_imp
children1	42
branchno8	40
age_quant(17,28]	29
rural1	28
rich1	26
expenditure_goods	24
club_frequently1	23
club_sometimes1	22
partner1	19
study_hours_quant(55,68]	18
study_hours_quant(68,166]	17
worry_job1	16
ablework_married1	15
who_decidehusband2	14
who_decidehusband3	13
index_empowerment_quant(14.3,42.9]	9
satisf_income1	2

Note:
Causal forest determines variable importance by a weighted sum capturing the number of splits a feature had in every depth layer in the model. Presented in descending order are variables that were above the mean weight of all the variables used for splitting in the model.The outcome Gender Empowerment is based on multiple questions relating to gender roles in labor markets, education and household chores.

Table E. 3 List of Selected Variables. Outcome: Control Over Body

	sort_var_imp
children1	40
age_quant(15,17]	38
age_quant(17,28]	36
rural1	33
expenditure_goods	25
partner1	24
study_hours_quant(55,68]	23
study_hours_quant(68,166]	22
worry_job1	19
work_married1	17
index_aspiration_quant(0.0553,0.612]	14
index_iga_quant(-0.216,0.041]	13
index_empowerment_quant(14.3,42.9]	12
satisf_income1	2

Note:
Causal forest determines variable importance by a weighted sum capturing the number of splits a feature had in every depth layer in the model. Presented in descending order are variables that were above the mean weight of all the variables used for splitting in the model. 'The outcome index COB is based on indicators for reported motherhood, being married or cohabiting, having had sex unwillingly in the past one year and always using a condom (if sexually active) as well as a
pregnancy and HIV related knowledge scores' (Bandiera et al., 2014, p.42).

Appendix F: Additional Analysis

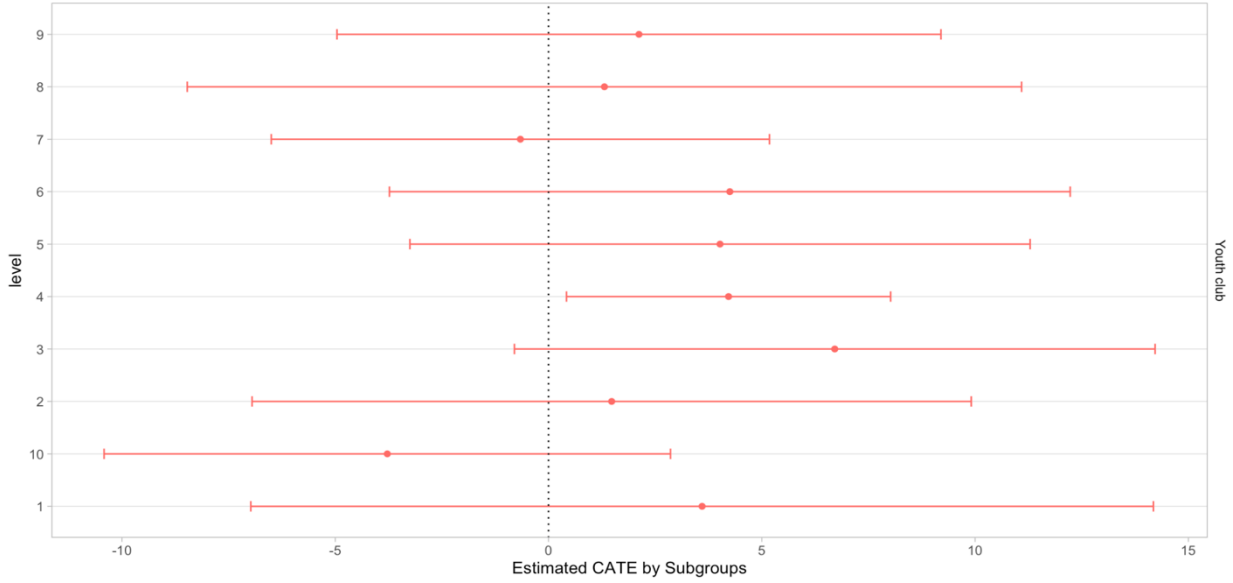
Table F. 1. Casual forest estimated Average Treatment Effects using only variables introduced to the linear models

	Expenditure	Gender Empowerment	Control Over Body
Estimated ATE	5811.271	2.827715	0.6615215
Standard error	1248.184	1.136856	0.0747113

Note:

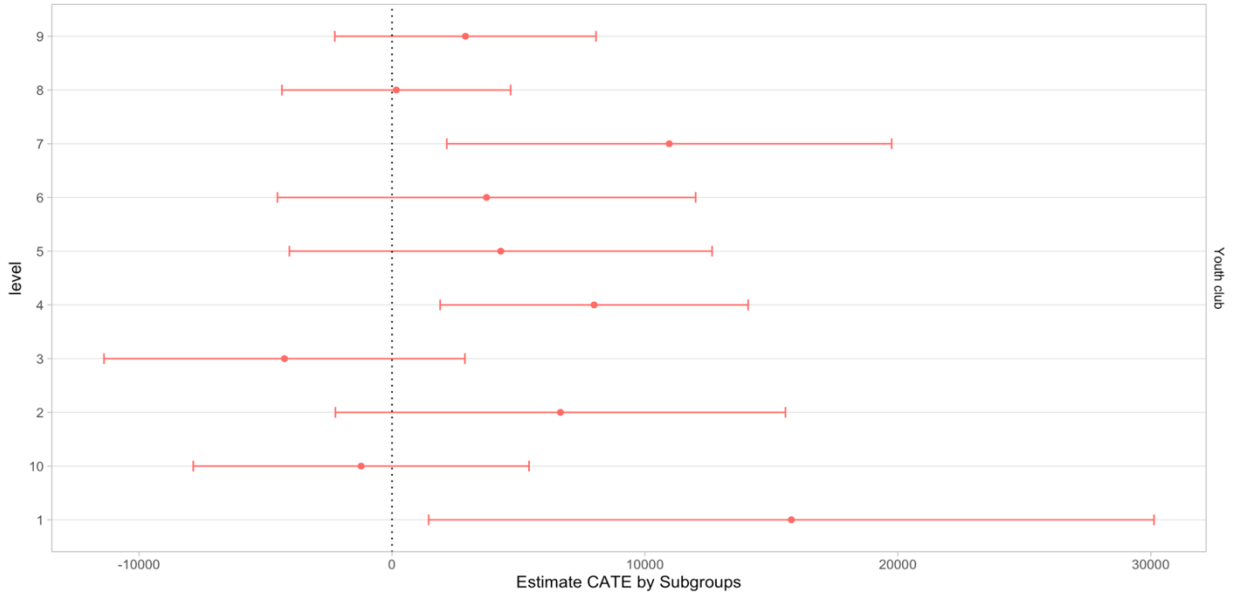
Estimated ATEs by Causal forest using only baseline variables selected by theory for the linear models. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The Gender Empowerment index cummulates answers regarding gender roles and Control Over Body represents the extent of reproductive health literacy and family planning. See Methodology chapter for full specification of variables

Figure F. 1. Estimated CATE by club attended. Outcome: Gender Empowerment



Note. 95% CI are calculated based on the standard error of each covariate level.
Each dot indicates the estimated ATE for the subgroup.
The vertical black dotted line designates the boundary of zero impact.
The outcome Gender Empowerment index is based on multiple questions relating to gender roles in labor markets, education and household chores (scale 0 to 100).
See Table 1. for full specification of the outcome.

Figure F. 2. Estimated CATE by club attended. Outcome: Expenditure on Goods



Note: 95% confidence intervals are calculated based on the standard error of each covariate level estimated ATE effect which can be seen by the colored dots.
The black dotted line show the boundary for a zero impact.
For expenditure variable, the goods categories are jewelry/ornaments, cosmetics/makeup, clothes, hairdressers, shoes/footwear, going to restaurants/bars/teashop/cafe, talk time for your mobile phone and presents/gifts.The top 1% outliers of the expenditure variable have been removed. copied from Bandiera et al., 2014

Appendix G: Estimated CATES by Ntiles

Table G. 1 Estimated CATEs of outcomes by Ntiles

Outcome: Expenditure on Goods [UGX].

	Sample ATE
ntile1	4435.59
	(1371.5)
ntile2	7859.59
	(1335.18)
ntile3	3108.25
	(1352.98)
P-Value	0.0358

Outcome: Gender Empowerment.

	Sample ATE
ntile1	0.398247
	(1.15305)
ntile2	1.34984
	(1.14548)
ntile3	5.04412
	(1.25341)
P-Value	0.0173

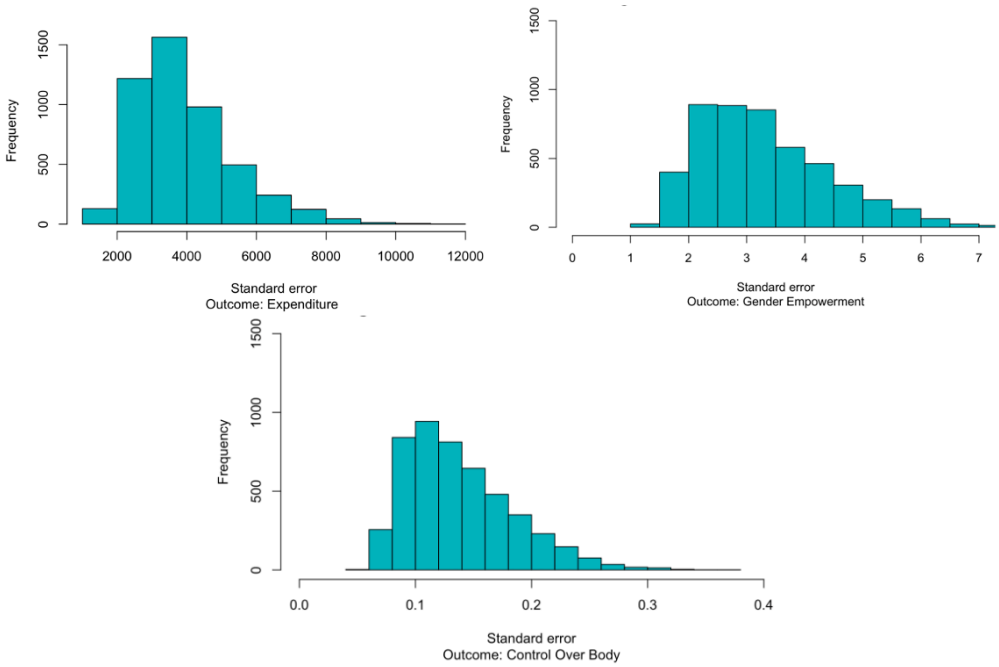
Outcome: Control Over body.

	Sample ATE
ntile1	0.636016
	(0.047705)
ntile2	0.485653
	(0.046651)
ntile3	0.53278
	(0.046082)
P-Value	0.071

Note. ATE per subgroup by out-of-bag CATE. P-value is testing Ho: ATE is constant across n-tiles. Sample ATE uses and F-test. The outcome Expenditure refers to consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The Gender Empowerment index cumulates answers regarding gender roles and Control Over Body represents the extent of reproductive health literacy and family planning. See Table 1. for full specification of variables.

Appendix H: Standard Errors of the Trained Causal Forest

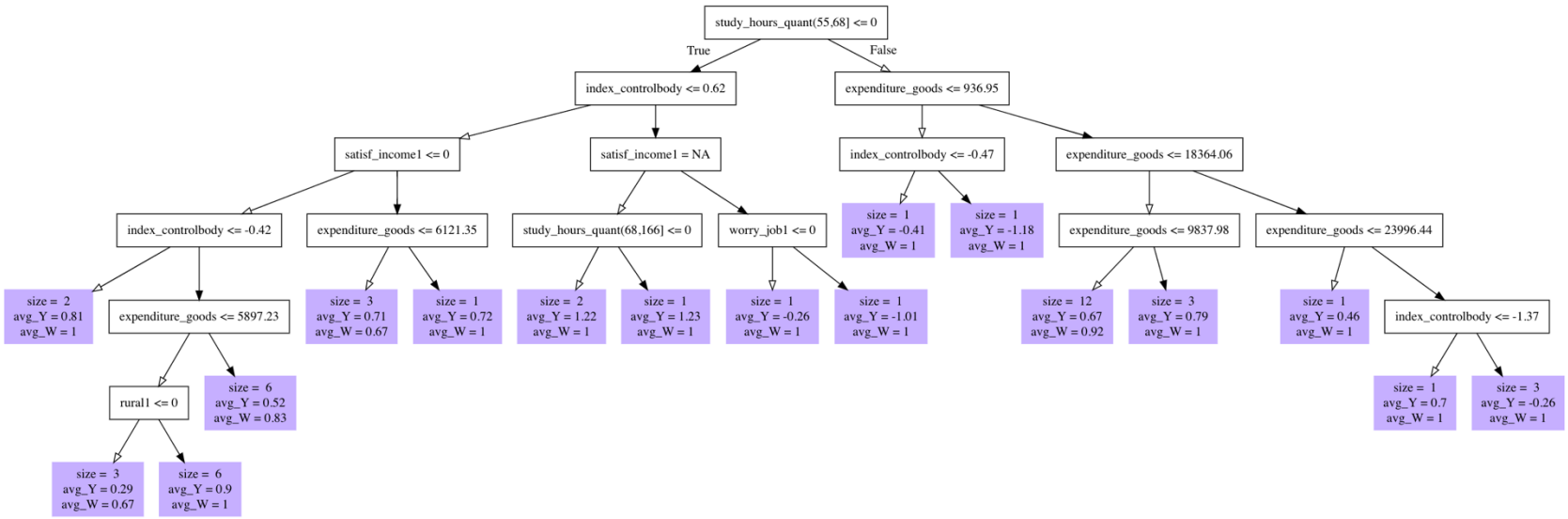
Figure H. 1 Histogram of estimated standard errors of trained causal forest



Note: The outcome Expenditure codes consumption of goods in several goods categories. The variable is measured in UGX currency. Exchange rate was 1,700UGX/1$. The Gender Empowerment index cumulates answers regarding gender roles, and it is rescaled from 0 to 100 where higher values indicate more empowerment. The index COB is composed by z-score standardised variables. It captures the extent of reproductive health literacy and family planning where higher values mean more COB. See Table 1 in Methodology chapter for the complete specification of these outcomes.

Appendix I: Example Tree from the Trained Causal Forest

Figure I. 1. Tree from a trained causal forest. Outcome: Control Over Body



Appendix J: Return on Investment Calculations

First scenario:

ATE: (12*33.8+ (12*33.8)/1.1)/17.9-1 = 42.25%

ntile 1: (12*31.3 + (12*31.3)/1.1)/17.9-1= 39.1%

ntile 2: (12*55.47 + (12*55.47)/1.1)/17.9-1 = 70%

ntile 3: (12*21.93+ (12*21.93)/1.1)/17.9-1 = 27.1%

Second scenario:

ATE: (12*33.8 + (12*33.8)/1.1 + (12*16.5)/1.1 + (12*8.25)/1.1)/17.9-1 = 57.34%

ntile 1: (12*31.3 + (12*31.3)/1.1 + (12*15)/1.1 + (12*7.5)/1.1)/17.9-1 = 52.8%

ntile 2: (12*55.47 + (12*55.47)/1.1 + (12*27.5)/1.1 + (12*13.7)/1.1)/17.9-1 = 95.1%

ntile 3: (12*21.93+ (12*21.93)/1.1 + (12*11)/1.1 + (12*5.5)/1.1)/17.9-1 = 37.1%

Third scenario:

ATE: (12*33.8 + (12*33.8)/1.1 + (12*26.5)/1.1 + (12*21.5)/1.1 + (12*17)/1.1 + (12*14)/1.1 + (12*11)/1.1+(12*9)/1.1)/17.9-1 = 102.5%

ntile 1: (12*31.3 + (12*31.3)/1.1 + (12*25)/1.1 + (12*20)/1.1 + (12*16)/1.1 + (12*13)/1.1 + (12*10)/1.1+(12*8)/1.1)/17.9-1 = 95.1%

ntile 2: (12*55.47 + (12*55.47)/1.1 + (12*44)/1.1 + (12*36)/1.1 + (12*28)/1.1 + (12*23)/1.1 + (12*18)/1.1+(12*15)/1.1)/17.9-1 = 169.9%

ntile 3: (12*21.93 + (12*21.93)/1.1 + (12*18)/1.1 + (12*14)/1.1 + (12*11)/1.1 + (12*9)/1.1 + (12*7)/1.1+(12*6)/1.1)/17.9-1 = 66.7%

MIA/MPP Master Thesis group form

Name of student 1	Ofer Dotan
Name of student 2	Mariana Saldarriaga
Thesis Advisor	Leonardo Iacovone

Master Thesis Title

Using Causal Machine Learning to Explore Heterogeneous Treatment Effects:

A case-study application to girls' empowerment programme in Africa

Please indicate which part(s) have been elaborated by whom, go into detail and be specific*:

Student 1

I initiated the idea for the thesis topic and Mariana then helped develop it and to choose the case study.

I wrote the following sections, but Mariana has reviewed all of them and added input.

We discussed all analysis of results before writing this and the discussion section. Moreover, we edited the thesis together and all sections were reviewed by both of us.

Executive summary

Introduction -
1.1. Choice of Case Study
1.2. Structure of the paper
Literature review
2.1. Causal Inference and Heterogeneous Effects Estimation
Theoretical framework
3.3. Research Hypotheses (together)
Methodology
4.1. Data
4.2. Pre-processing of the data
4.2.1. Operationalisation of the Outcomes
4.2.2. Selected Variables for the Analysis
4.3. Applying Causal Forest to the Data
4.4. Model Specification
4.5. Analytical Strategy
Results
6.3. Assessing Heterogeneous Treatment Effects
6.3.1. Variation in CATE estimates
6.3.2. "Best linear predictor" method
6.3.3. Assessing Heterogeneity Across Subgroups
Return on Investment (ROI) - Ofer
Discussion
8.1. Methodological Considerations
8.2. Substantive Results
8.3 Lessons Learned and Limitations

Code:
We developed the code together and most code was written during programming days in which we worked simultaneously.
Segments of code that were separated:
Survey analysis
descriptives of discrete variables

Student 2

While Ofer came with the idea for the thesis, I suggested the case study which is the focus of our application.

I wrote the following sections, but Ofer has reviewed all of them and added input.

We discussed all analysis of results before writing this and the discussion section. Moreover, we edited the thesis together and all sections were reviewed by both of us.

Literature review
2.2. Heterogeneous Effect Estimation with Machine Learning
2.2.1. The Causal Forest
2.2. Heterogeneous Effect Estimation with Machine Learning
2.2.1. The Causal Forest
Theoretical framework
3.1. Women's Empowerment
3.2. Enablers and Constraints Driving Women's Empowerment
3.3. Research Hypotheses (together)
Methodology
4.6. Model Evaluation
4.6.1. Assumptions
Descriptive statistics/Baseline characteristics
Results
6.1. The Average Treatment Effect (ATE)
6.2. Replications of the results using OLS
6.3.4. Assessing Heterogeneity Within Subgroups
6.3.4.1. Expenditure on Goods and Gender Empowerment
6.4. Evaluation Causal Forest Fit
6.4.1. Assessing the Overlap Assumption
6.4.2. Assessing bias of the causal forest predictions

Code:
We developed the code together and most code was written during programming days in which we worked simultaneously.

Segments of code that were separated:

Replication of linear models

Descriptives of continuous variables

* Please note that individual input has to be marked off and should be listed in this form; according to BerlHG §32 (5) individual, differentiated grading has to be possible.

We confirm that we both have equally invested our time and effort in the master thesis and that we followed good academic conduct in group work (see exam rules *§14 master thesis (5)*, and *§15 good academic conduct*).

Ofer Dotan
_____.

Mariana Saldarriaga

Date, signature student 1

Date, signature student 2

Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on 26/05/2021 is identical to the printed version I submitted to the Examination Office on 26/05/2021.

DATE: 26/05/2021

NAME: Ofer Dotan

SIGNATURE: Ofer Dotan

Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on 26/05/2021 is identical to the printed version I submitted to the Examination Office on 26/05/2021.

DATE: 26/05/2021

NAME: Mariana Saldarriaga

SIGNATURE: