Journal of Population Economics (2024) 37:10
https://doi.org/10.1007/s00148-024-00987-z

ORIGINAL PAPER

Empowering women through multifaceted interventions: long-term evidence from a double matching design

Stanislao Maldonado[1]

Received: 4 January 2022 / Accepted: 8 January 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
Empowering women is a policy goal that has received a lot of interest from policymakers in the developing world in recent years, yet little is known about effective ways to promote it sustainably. Most existing interventions fail to address the multidimensional nature of empowerment. Using a double matching design to construct the sampling frame and to estimate causal effects, I evaluate the long-term impact of a multifaceted policy intervention designed to improve women's empowerment in the Atlantic region in Colombia. This intervention provided information about women's rights, soft skills and vocational training, seed capital, and mentoring simultaneously. I find that this intervention has mixed results: there are improvements in incomes and other economic dimensions along with large political and social capital effects, but limited or null impacts on women's rights knowledge and control over one's body. Using a list experiment, I even find an increase in the likelihood of intra-household violence. The results highlight the importance of addressing the multidimensional nature of women's empowerment in policy innovations designed to foster it and incorporating men in these efforts.

Keywords Women · Empowerment · Multifaceted interventions · Matching

JEL Classification I38 · J16 · J24 · O17

Responsible editor: Alfonso Flores-Lagunes

This article is dedicated to the loving memory of my father, Humberto Zambrano. Te voy a extrañar mucho, papá.

✉ Stanislao Maldonado
sh.maldonadoz@up.edu.pe

[1] Universidad del Pacifico, Lima, Peru

Published online: 03 February 2024

1 Introduction

Women's empowerment is a critical component of economic development, yet very uneven progress has been observed in this regard (Doepke & Tertilt, 2011) and Jayachandran (2015). Following (Fox & Romero, 2017) and Bandiera et al. (2020), I define women's empowerment as a set of multidimensional outcomes to be approached by attitudes and behaviors in four critical dimensions: economic, political, social, and control over one's body.[1] Although women's empowerment is a relevant policy goal now everywhere, progress towards this goal has been very slow in the developing world (Duflo, 2012) and World Bank (2012).[2] The standard policy response to this dramatic situation is based on a set of interventions that aim to attack one particular barrier to women's empowerment, yet these efforts have been shown to have limited success.[3] A promising area of improvement can be based on the design of multifaceted interventions that simultaneously target different dimensions of women's disadvantage but little is known about the effects of this kind of program.[4] Very few interventions address empowerment in a multidimensional way, and there is limited evidence on their impacts. Therefore, whether a multifaceted policy intervention can sustainably break this low-level empowerment equilibrium remains largely unanswered.

In this paper, I study the effects of "Transformate Tu Mujer" (TTM), a multifaceted empowerment intervention implemented by the Government of the Atlantico in the Colombian Caribbean, on a large set of economic, political, and social outcomes. Because women's empowerment is multidimensional, complementary interventions are needed to break the low-level empowerment equilibrium women in the Atlantic Colombian face. This program was designed to provide training on soft skills (including training modules on women's rights and self-esteem), vocational training with an emphasis on technical skills (including the development of a business plan with experts' support), mentoring, and in-kind seed capital to women entrepreneurs older

[1] Empowerment was introduced into economics by Sen (1999) and is related to fundamental concepts such as power and agency. Common elements in existing frameworks (Fox & Romero, 2017) include aspirations and self-efficacy as well as the behavioral dimension of action, the fact that individuals operate in environments with formal and informal behavioral constraints, and the idea that empowerment is a process and an outcome. Although there are some important differences among them, these approaches are interpreted as complementary in the rest of the paper.

[2] The empirical evidence is clear in this regard. Female participation in the labor market is still low, with labor earnings typically below those of men for similar occupations (Klasen, 2019) and Jayachandran (2020). Many women have children at a younger age due to limited access to contraceptive methods, and lack of information (Upadhyay et al., 2014). Educational opportunities are still unequal (Aslam, 2013) and Heath and Jayachandran (2018), and violence towards women is more prevalent and acceptable due to social norms (Vyas & Watts, 2009). Gender gaps in political participation and voice are still prevalent, which has led to a large underrepresentation in political office and leadership positions (Milazzo & Goldstein, 2019). Although some progress in many economic, political, and social dimensions has been documented (World Bank, 2012) and Buvinic and Furst-Nichols (2016), there are many areas in which women still face considerable disadvantages with respect to men.

[3] A review paper of (mostly) unidimensional empowerment interventions by Baird and Ozler (2016) concludes that "...studies that exist generally do not point to meaningful and sustained effects on key indicators of economic empowerment." (p. 16)

[4] Few papers have explored women's empowerment using multifaceted interventions. Recent examples include (Austrian et al., 2020) (Bandiera et al., 2020), (Buehren et al., 2017), (Elsayed & Roushdy, 2017), and Prennushi and Gupta (2014).

than 18 years old with a business or business idea. This follows the same logic of the recent wave of graduation programs, where multiple interventions are simultaneously delivered to break a poverty trap (Banerjee et al., 2015) and Bandiera et al. (2017). I evaluate the long-term impacts of this program 4 and 5 years after its completion.

Studying the effects of this multifaceted intervention faces a non-trivial set of methodological challenges. As is common in many retrospective evaluation designs, baseline data were not collected before the intervention began in 2012. Therefore, this study's fundamental methodological challenge consists of constructing a valid control group to identify the counterfactual state of program participants. I deal with this challenge by exploiting pre-treatment data recovered from the National System of Social Programs' Beneficiaries (SISBEN). I secured access to the socio-economic information of more than 2.5 million individuals in the areas of the program's coverage. Although using SISBEN data allows me to solve data restrictions in this setting, it can only capture socioeconomic characteristics from TTM beneficiaries who are vulnerable to poverty. However, it could be argued that this is a relevant group from a policymaker's perspective.[5]

Using the program's administrative records, I identify a sub-group of the program's beneficiaries that are also registered in SISBEN records. For this sub-group of former program participants, I recovered a broad set of pre-treatment characteristics for 2011, a year before the start of the intervention. To construct the control group, I identify women of similar socio-economic characteristics as this previous group from SISBEN's records using a genetic matching algorithm (Diamond & Sekhon, 2013). This matching algorithm is used as a pre-processor tool in the sense of Ho et al. (2007). For each identified program participant in SISBEN's records, I identify at least three individuals with similar pre-treatment characteristics based on the matching algorithm. This process allowed me to identify a sampling frame for the study that is based on a set of individuals who are similar in a large set of socio-economic characteristics.

Using this sampling frame, a survey company carried out the fieldwork for collecting a household survey with information on a large number of empowerment outcomes and related information in March 2017. This company followed all standard statistical procedures to collect this survey but used a sampling frame designed to ensure that treated and comparison units are similar ex-ante in a relevant set of socio-demographic dimensions. To the best of my knowledge, this strategy has not been used before in the way proposed in this paper. Using this sample, genetic matching is used again, but this time to identify a causal relationship in combination with the nearest neighbor matching estimator proposed by Abadie and Imbens (2006) and Abadie and Imbens (2011). This combination is based on the advantages of both algorithms. On the one hand, genetic matching has been shown to have the best performance in achieving covariates balance among a large set of alternative matching and machine learning algorithms (Colson et al., 2016). On the other hand, the matching estimator proposed

[5] Unfortunately, data for the program's participants not included in SISBEN is not available. The information collected by the program was minimal; therefore, this group of participants was discarded when constructing the sampling frame. Consequently, I cannot provide details about this set of participants.

by Abadie and Imbens (2006) and Abadie and Imbens (2011) has the best performance among non-experimental estimators to recover experimental results (McKenzie et al., 2010).

The TTM program has mixed effects on women's empowerment. In the case of economic empowerment, I find a positive set of impacts on economic well-being indicators. TTM is causally related to a 31% increase in monthly income, a 26% reduction in a multidimensional poverty indicator, and a 50% increase in the likelihood of being employed. I also document large effects on labor supply (either measured in terms of occupation, working hours, or related measures) and an increase in savings (56% for women and 31% for other household members). On the other hand, I find no impact on household assets.

In terms of political and social empowerment, the evidence is mixed. I find no changes in terms of household decision-making power, but an increase in several political participation and social capital indicators. Unfortunately, I also find an increase of 15% in the likelihood of experiencing intimate partner violence (IPV) using a list experiment. Regarding dimensions of knowledge and attitudes on women's rights, I find the program plays no role in changing outcomes related to these dimensions. In sum, the program is related to valuable empowerment dimensions on the economic, political, and social sides but with more mixed evidence regarding impacts on decision-making within the household as well as attitudes and beliefs about women's rights. Even worse, the program is associated with an increase in IPV. Most of these results are robust to violations in the "selection on observables" assumption using the sensitivity analysis proposed by Rosenbaum (2002).

This paper is related to a growing literature on women's empowerment in developing countries. Existing evidence is mostly non-experimental and has been surveyed elsewhere (Duflo, 2012), (Buvinic & Furst-Nichols, 2016), (Buvinic & O'Donnell, 2019), and World Bank (2012). A small yet growing experimental literature has been produced in recent years, mostly focused on the effects of multifaceted interventions on young girls. The evidence is mixed, with some programs showing important effects on empowerment (Bandiera et al., 2020),[6] and others showing partial (Buehren et al., 2017)[7] or null evidence of impacts (Austrian et al., 2020).[8] Non-experimental evidence about multifaceted empowerment interventions is also scarce but it seems to be

[6] They study the Empowerment and Livelihood for Adolescent (ELA) program in Uganda, which provides vocational and soft skills training using social clubs. This program was implemented by the NGO BRAC and was expanded to several countries in Africa and Asia, including Liberia, Nepal, Sierra Leone, South Sudan, and Tanzania. This program successfully increased girls' participation in economic activities, reduced early pregnancy and marriage, and raised aspirations regarding marriage, childbearing, and fertility.

[7] They study a similar ELA program in Tanzania that included a microcredit component and explored similar outcomes as Bandiera et al. (2020) without finding improvements in any of these outcomes but some evidence in terms of increases in savings.

[8] They study the AGEP program's effects in Zambia on empowerment outcomes for young girls. This program provided vocational and socio-emotional skill training using clubs, vouchers for health services, and a savings account. The intervention did not lead to relevant changes in economic assets, educational, or fertility outcomes.

more positive as suggested by Elsayed and Roushdy (2017)[9] and Prennushi and Gupta (2014)[10] for Egypt and India, respectively.

This paper differs from this previous literature regarding the dimensions of empowerment addressed by TTM and the scope in terms of evaluation outcomes. Except for Prennushi and Gupta (2014), most existing multifaceted interventions are based on vocational and life skills training, sometimes bundled with financial services. TTM incorporates these dimensions and includes mentoring and seed capital (including inputs and equipment), which is expected to tackle multiple and interlinked constraints to women's well-being by exploiting potential complementarities between the program's components. In terms of outcomes, this paper studies a larger set of economic, social, political, and psychological outcomes, including effects on IPV. In this sense, this paper adds to the existing literature by analyzing the effects of a more complex multifaceted intervention on a more rich set of outcomes.

This paper is also related to literature that explores the effects of interventions that focus on partial aspects of women's empowerment regarding treatments and outcomes. These include papers that explore the effects of one dimension of women's empowerment on multidimensional or uni-dimensional outcomes, which is limited since improvements in some dimensions can be accompanied by reductions in others (Fox & Romero, 2017). For instance, a large literature has explored the effects of vocational and business training with mixed results.[11], and a recent scholarship has shown that life or socio-emotional skills can have large effects on economic empowerment.[12] Some recent evidence suggests that mentoring can be a powerful yet cost-effective strategy to promote women's economic empowerment.[13] At the same time, scholars

[9] They analyze the effect of an empowerment intervention in Egypt based on vocational, business, and life skills training (including some basic support for legal registration and opening bank accounts) on labor market and entrepreneurship outcomes. Using a matching-differences in difference design, the authors find medium-size effects of this intervention on labor outcomes.

[10] They study the effects of an empowerment program in India that provided seed funds, access to financial services, and training in business and life skills on a broad set of assets and expenditures, human development, and women's empowerment outcomes. The authors find positive impacts on asset accumulation, increases in household expenditures, and larger investments in education, along with positive changes in women's autonomy, mobility, and participation in social activities.

[11] The evidence for vocational training suggests that these interventions have some effects on labor market outcomes but they are typically modest (McKenzie, 2017) These results do not seem to be different for women. Regarding business training, the evidence suggests that they are effective in improving business practices but their effects on profits and business survival are small (McKenzie, 2021) and (Woodruff & McKenzie, 2012).

[12] A recent scholarship has shown that soft skills can boost profits, employment, and crop adoption for women (and also men). See, for instance, the evidence provided by Edmonds et al. (2021) for India, Ashraf et al. (2020) for Zambia, Campos et al. (2017) for Togo, Montalvao et al. (2017) for Malawi, and Croke et al. (2017) for Nigeria.

[13] Two recent papers have shown the positive impacts of mentoring on microenterprise-level outcomes. Brooks et al. (2018) study the effects of a mentorship program for inexperienced female microentrepreneurs in Kenya finding a 20% increase in profits. On the other hand, Lafortune et al. (2018) evaluate a similar intervention in Chile where personalized consulting sessions are compared to role models. These role models are former participants of a training program. They find that both interventions raised household income by 15%, being the role model one more cost-effective.

have shown mixed effects of providing capital to entrepreneurs.[14] This paper builds on the existing literature about the effects of these programs by analyzing the impacts of a bundled empowerment intervention that includes all these dimensions simultaneously.

The rest of the paper is organized as follows. Section 2 provides details about the institutional setting and the TTM program. Section 3 introduces the research design, and Section 4 presents the empirical results. Section 5 explores additional results, and Section 6 presents the results of the sensitivity analysis. Section 7 concludes by discussing the implications of this paper's findings.

2 Institutional setting

2.1 Women's empowerment in Colombia

According to the Global Gender Gap Report 2020 (World Economic Forum, 2019), Colombia ranks 22 out of 153 countries in terms of the global gender gap index, an indicator designed by the World Economic Forum to measure progress towards the gender equality goal. This good performance is driven by significant progress in terms of educational attainment and health conditions where gender gaps have disappeared. However, important gaps remain in terms of economic opportunities and political empowerment.

In terms of education, women's participation in tertiary education almost doubled in the past decade. Now, 6 out of 10 women between 17 and 21 years old are enrolled in a university or technical-level institution (ONU MUJERES, 2018). There are no gaps in terms of enrollment in primary and secondary education, and the literacy rate is close to 100% for both sexes, but some gaps persist in terms of learning as measured by standardized tests (World Economic Forum, 2019) and Sanchez (2018).

Regarding health, almost universal and gender-equal coverage was achieved by 2017, although quality issues remain a challenge for the country (ONU MUJERES, 2018). Healthy life expectancy is high (70 years), close to the standards of developed nations like Sweden (73.4) or Canada (74.3), and above most of its neighbors (World Economic Forum, 2019). This contrasts with women's exposure to IPV, where Colombia ranks second in a sample of Latin American countries, with a prevalence of physical violence of 17.5% and sexual violence of 3.8% (Bott et al., 2019). Even worse, 2.7 women per day were murdered in 2017 (IDEA et al., 2019).

In terms of economic empowerment, progress toward gender equality has been slow and very unequal. According to official statistics, the gender gap in labor market participation is still high (21%) and has remained almost the same for the last decade (ONU MUJERES, 2020). Unemployment is higher for women (14% over 8% for

[14] The evidence in this regard is mixed. de Mel et al. (2008) documents large effects of small infusions of capital on profits for male-headed small enterprises. The short-term effects are close to 6% and five years later are close to 12% (de Mel et al., 2012). The effect for female-headed small firms is indistinguishable from zero. Fafchamps et al. (2016) study a similar intervention in Ghana, finding differential effects by gender. In the case of female-owned microenterprises, only in-kind grants seem to have a positive effect, especially for businesses of larger size. However, if one focuses on smaller female-headed firms, the returns to capital disbursement are close to zero.

men), and a woman earns 88% of a man's wage (ONU MUJERES, 2018) and (Fonseca, 2019). Moreover, the fraction of women without income is 3 times the same proportion for men (ONU MUJERES, 2020). It is not surprising that the incidence of poverty among women is higher: 120.3 women are poor for every 100 men (ONU MUJERES, 2018). This indicator has been worsening in the last decade.

Progress in terms of political empowerment has also been limited (IDEA et al., 2019). Although 50% of ministerial cabinets are female, the fraction of women in the parliament is only 27% (World Economic Forum, 2019). In 2018, only 12.1% of elected mayors and 17.6% of elected city council members were female (ECLAC, 2020). Part of the gap is explained by the fact that few women decide to run for political office. For instance, the fraction of women running for Congress was only 34.5% (IDEA et al., 2019). Even though some progress has been observed in the past decade (World Economic Forum, 2019), Colombia's political empowerment performance is still below the regional average.

These numbers hide important heterogeneities and inequalities. Many of those indicators are worse for poor women in the Atlantico department, the typical TTM participant in the evaluation sample. For instance, in terms of economic empowerment, the participation rate, the unemployment, and the wage gap are larger in Atlantico than the national average (Fonseca, 2019).[15] This is similar to other dimensions of empowerment. Therefore, although Colombia shows areas of substantial progress in terms of women's empowerment along dimensions that required additional effort, it is crucial to keep in mind that the Atlantico department is behind in terms of economic and social opportunities for women in comparison with the country's averages.

2.2 "Transformate Tu Mujer" program

2.2.1 The intervention

TTM was implemented by the Women and Gender Equality Secretariat of the Government of Atlantico to improve women's empowerment along multiple dimensions. The program was designed to incorporate several components that are expected to simultaneously affect the psychological, economic, political, and social dimensions of women's empowerment by delivering training on business and socio-emotional skills, mentoring, and in-kind seed capital.

The program was open to women of legal age with a trade. With the support of local governments, social workers were hired to set up temporary offices across the department to reach the target population and help potential participants fill out application forms. This form collected applicants' personal information and details about the business or business idea. Then, applicants were interviewed by experts in entrepreneurship to classify them according to how innovative their business ideas were and by their level of commitment to their businesses. Applicants were ranked according to these two criteria, and a final list with the selected applicants was published. The same

[15] In the Atlantico department, women's participation rate is 52% versus 78% for men. The unemployment rate for women is 11% versus 5% for men, and the wage gap is 19%. See Fonseca (2019) for details.

procedure was used for the first two cohorts of the program that will be analyzed in this paper (2012 and 2013).

2.2.2 Selection into the program

Based on the qualitative evaluation (see Section 8.2 of CNC (2017)) and in interviews with former program officials, I identify two sets of factors that explain the selection into treatment: socioeconomic characteristics and skills/traits related to entrepreneurship. Regarding the first set of factors, in interviews reported in CNC (2017), most program officials emphasize "equality of opportunities" as a criterion to select participants (CNC, 2017: 120), along with being older than 18 and having a small business or a business idea. Therefore, they were interested in giving more weight to socioeconomic characteristics for choosing participants.[16]

The second set of factors captures skills and traits related to entrepreneurship. Program officials and external jurors run interviews and auditions among potential candidates to classify them according to their level of progress regarding the business activity. The qualitative study suggests that program officials and external jurors paid more attention to some traits and skills rather than the level of development of the business or business idea. In most cases, the level of knowledge about the business was basic. One interviewed official expressed this in the following way:

"*So, the women auditioned for about 3, 4, 5 minutes, brought a product and presented it, some brought their own products and others borrowed what they wanted to do, like flip-flops... but what one looked at was the attitude, the drive, the thing, and then you rated them.*" (Person 7, CNC, 2017: 122).

Therefore, although having a business or business idea was considered a prerequisite for joining the program, in practice, program officials were flexible and paid attention to socioeconomic characteristics and traits related to entrepreneurship. About 55% of program participants in the sample reported having a business in 2011, before the start of TTM.

2.2.3 Program phases

Once selected, program participants were exposed to the following program phases:

(i) **Phase 1: Training on soft skills, women's rights, and self-esteem**
The first phase of the intervention was designed to train participants on soft skills, emphasizing women's rights, agency, and self-esteem. Participants received a 90-h long training program delivered by specialized trainers of the Universidad del Norte, one of the most prestigious universities in the department. This program was composed of 5 modules that emphasized aspects such as self-control and autonomy (module 1), health and lifestyles (module 2), women's rights (module 3), empowerment and solidarity (module 4), and women as an agent of change and social transformation (module 5).

[16] This idea is reinforced by the fact that the program opened additional slots for beneficiaries affected by seasonal rains in 2010, with a clear goal of targeting women economically affected by natural disasters. Hence, socioeconomic conditions and economic vulnerability played an important role in the selection process.

(ii) **Phase 2: Vocational training with an emphasis on technical skills and the development of a business plan with experts' support**
The second phase was designed to improve technical skills related to participants' line of business. In this phase, participants were accompanied by instructors with large experience and knowledge in the business line selected by participants during the enrollment phase. It also included the development of a business plan with the instructor's support.
This phase was composed of 4 modules, of which the last two were focused on developing a business plan and support during its implementation. The first two modules emphasized entrepreneurship and business associativity. In total, 170 hours were assigned to these phases. Women who completed the first two phases received a certification from the Universidad del Norte.

(iii) **Phase 3: Mentoring**
In this phase, participants worked with mentors in strengthening their business plans and were exposed to specialized training in accounting and communication tools.

(iv) **Phase 4: In-kind seed capital**
In this final step, the program distributes inputs and equipment to participants according to the business plan and budget developed in the second phase.

All these phases were implemented over six months. The first and second cohorts of TTM covered 6,202 participants. In this period, 623 businesses were constituted among TTM participants. The total program cost was USD 2,916,060.38 for the first 2 years. It was funded by the Colombian National Government through the national oil royalties system, the Government of Atlantico, the Mario Santo Domingo Foundation, and the Universidad del Norte (SMEG, 2020). The cost per participant was USD 470.[17]

As it is common in many of these interventions, it is tough to disentangle the role of specific program components. Due to some implementation issues, some of these components faced implementation challenges and were partially delivered. In the evaluation sample, all participants reported to have completed the TTM's training component, and 96.06% reported to have graduated from the program. Regarding the training component, a large proportion of participants considered that training materials (88%), topics covered (98%), activities implemented (98%), facilitators (98%), class duration (97%), and class participation (92%) were good.[18] Hence, the training component in phases 1 and 2 was delivered to almost all participants, and their evaluation of them was positive.

Regarding phases 3 and 4, the results are mixed. A large fraction of participants worked with mentors (76%) during the development and implementation of their business plans. Most of them also received inputs and equipment for their businesses (90%), but participants' evaluation of this component is less optimistic. For instance,

[17] Own calculation based on program information. Unfortunately, detailed information about the program's cost structure is not available to perform a cost-benefit analysis. See SMEG (2020) for details.

[18] These and the following calculations in this section are based on the evaluation survey carried out by CNC in March, 2017.

qualitative interviews carried out as part of TTM evaluation (CNC, 2017) suggest that participants perceived that the distribution of inputs and equipment was insufficient.[19]

One critical element of the program's implementation was the goal of fostering associations between participants of the same line of business. Still, this goal was not effective because participants' businesses were highly heterogeneous, and they found little value in the kind of associations fostered by the program.[20] Therefore, it seems that these elements of TTM were less effective as perceived by participants.

It is important to emphasize that disentangling the effects of the program's components is not feasible. From a policy-maker perspective, it would be helpful to understand which elements of a multifaceted intervention are more effective in raising women's empowerment. However, a rigorous analysis of mechanisms requires variability in components' assignment across treated units, so the contribution of each component can be assessed. When every treated unit gets all the components of the treatment package, such variability does not exist, so applying approaches like this becomes unfeasible. Therefore, a formal analysis to study the mechanisms is impossible in this context.

3 Research design

In this section, I describe the methodological steps used to evaluate the effects of TMT and provide a discussion about the design's validity. This design is retrospective and evaluates the effects of this intervention for 2012 and 2013 cohorts, 5 and 4 years after the program's end.

The main components of the research design are the following: First, I construct the sampling frame using genetic matching as a pre-processor tool (Ho et al., 2007) based on pre-treatment data from SISBEN records; second, with the sample collected using this sampling frame, I then estimate the causal effects of TTM on empowerment outcomes using genetic matching for a second time along with the matching estimator proposed by Abadie and Imbens (2006) and Abadie and Imbens (2011); and third, I perform a sensitivity analysis a la Rosenbaum (2002) to evaluate how sensitive the results are to violations of the CIA assumption.

3.1 Matching techniques

Genetic matching is an ideal approach for the research problem at hand both for pre-processing data to construct the sampling frame and the estimation of causal effects.

[19] It is also important to mention that the development of the business plan created expectations regarding the amount that it was unfeasible for the program to cover. Almost all treated women reported a negative perception of this program's component.

[20] Unfortunately, the heterogeneity among participants was a significant barrier in this regard. It was hard for the program to create these associations, except in cases where participants were friends or neighbors before the start of the program. Many of these associations did not work in the long term. Because the delivery of the seed capital was planned to be at the level of these associations, this process encountered important challenges. The evaluation team reported encountering many women still waiting for the seed capital during the data collection process. In many cases, when seed capital was distributed, it was common that members of a given association decided to split this capital among its members.

This matching technique is based on a genetic search algorithm that automatizes the process of checking and improving overall covariate balance (Sekhon & Mebane, 1998).[21] Genetic matching extends existing matching techniques by generalizing the Mahalanobis metric in the following way:

$$GMD(X_i, X_j, W) = \sqrt{(X_i - X_j)'(S^{-1/2})'W(S^{-1/2})(X_i - X_j)}; \quad (1)$$

where X is a matrix of covariates for units i and j, S is the sample covariance matrix, $S^{-1/2}$ is the Cholesky decomposition of S, and W is positive definite weight matrix. The genetic search algorithm chooses W such that it optimizes a pre-specified loss function to maximize the balance of observed covariates across matched treated and control units. Compared to other matching and machine learning approaches, genetic matching has the best performance in achieving covariates balance (Colson et al., 2016).[22]

Once covariates balance is achieved, the next step consists of comparing outcomes across treatment and control groups in the matched sample constructed with the genetic search algorithm. This can be done in several ways. The simplest one is just to perform a simple difference in means in the matched sample but this approach is unattractive. A more powerful approach consists of using the nearest neighbor matching estimator proposed by Abadie and Imbens (2006) and Abadie and Imbens (2011). In particular, I use the bias-corrected matching estimator that is robust to biases when matching is not exact in finite samples.

This matching estimator recovers the average treatment effect under standard assumptions. The first assumption is the so-called "stable unit treatment value assumption" (SUTVA). This assumption restricts the possibility that one unit's treatment assignment affects another unit's potential outcomes. The second assumption is "unconfoundedness" or CIA, which requires the treatment assignment to be independent of potential outcomes after controlling for a set of covariates. The third assumption is overlap or common support. This assumption implies that the distribution of the propensity scores in the treatment and control groups are similar.[23]

Whereas the two first assumptions are not directly testable, I will provide evidence about the feasibility of the third assumption in my setting. I will show that my matching strategy guarantees it by design. The first assumption is standard in matching designs, and one feature of my design allows me to explore the presence of spillovers. Although the sample size is not large enough to be conclusive, I will provide evidence suggesting

[21] A genetic algorithm solves complex optimization problems using heuristics inspired by the natural selection process. A population of potential solutions is evaluated iteratively. In each iteration or generation, the fitness of every individual in the population is evaluated. The more fit individuals are selected from the current population, and each individual's genome is modified to form a new generation. The new generation of potential solutions is used in the next iteration of the genetic algorithm. See Sekhon and Mebane (1998) for details.

[22] According to the simulations implemented by Colson et al. (2016), genetic matching has the best covariate balance among 7 alternative methods even when there is poor support in terms of the propensity score. This advantage is consistent across alternative balance metrics.

[23] Alternative formulations of these assumptions also include some regularity conditions, whether the set of covariates X has continuous distributions and restrictions about the sampling process. See Abadie and Imbens (2011) for details.

that spillover effects may be relevant in this context. The second assumption is likely to be violated in my setting. As explained above, selection into treatment was partially explained by unobservable traits and skills associated with entrepreneurship, which may bias my results upward. I will relax this assumption using two strategies. First, I include a propensity score as an additional control in multivariate matching, so I can accommodate a large set of variables in my design without falling into the "curse of dimensionality." Second, I implement the sensitivity analysis described in Section 3.4

In the next subsections, I describe how these matching techniques are used with the goal of pre-processing data and causal effect estimation.

3.2 Constructing the sampling frame and sample

3.2.1 Sampling frame

To construct the sampling frame, I first use all the records of the program participants for the 2012 and 2013 cohorts. In total, I have access to 3670 participants' records. This information was provided by TTM for this evaluation under standard protocols for data protection. I then merge this information with SISBEN records for 2011. This information was provided by the National Planning Department (DNP) for this study. I am able to identify 2445 participants in both the program and SISBEN's records. In this way, I can identify pre-treatment data for a subset of the program participants. However, it is important to emphasize that these are participants who are vulnerable to poverty, so the results of this study are relevant for this type of population.

Recovering pre-treatment socioeconomic data for the program's beneficiaries helps identify non-beneficiaries with similar socioeconomic characteristics using SISBEN's records. I take advantage of 2,552,455 available records for the area of influence of this intervention. This includes 26 municipalities in the department of Atlantico, as well as the departments of Bolivar and Magdalena.[24] Figure 1 presents a map with all the municipalities included in this study.

I use a multivariate genetic matching strategy to identify three non-beneficiaries in SISBEN's records for each program's participant. As with any matching design, a critical step consists of selecting the pre-treatment variables to be used to match individuals. I use a large set of household characteristics, assets, and other socioeconomic characteristics available in SISBEN's records. A full list of the variables and their definitions are included in Table S1 in the Online Appendix. I also use the SISBEN score, a welfare index used by the Colombian government to select social programs' beneficiaries. Each household receives a score from 0 to 100 (from poorest to richest). The score is calculated using 24 variables across four dimensions: health, education, housing, and vulnerability. Given its importance, I include this variable along its square in a multivariate matching design.

As it is well-known, multivariate matching may suffer from the "dimensionality curse." One way to address this issue is by combining multivariate matching with

[24] These departments were included in the sampling frame because of potential concerns of within-municipality spillovers. As mentioned later, the sample size is not large enough to exploit this feature of the sampling design in the analysis.

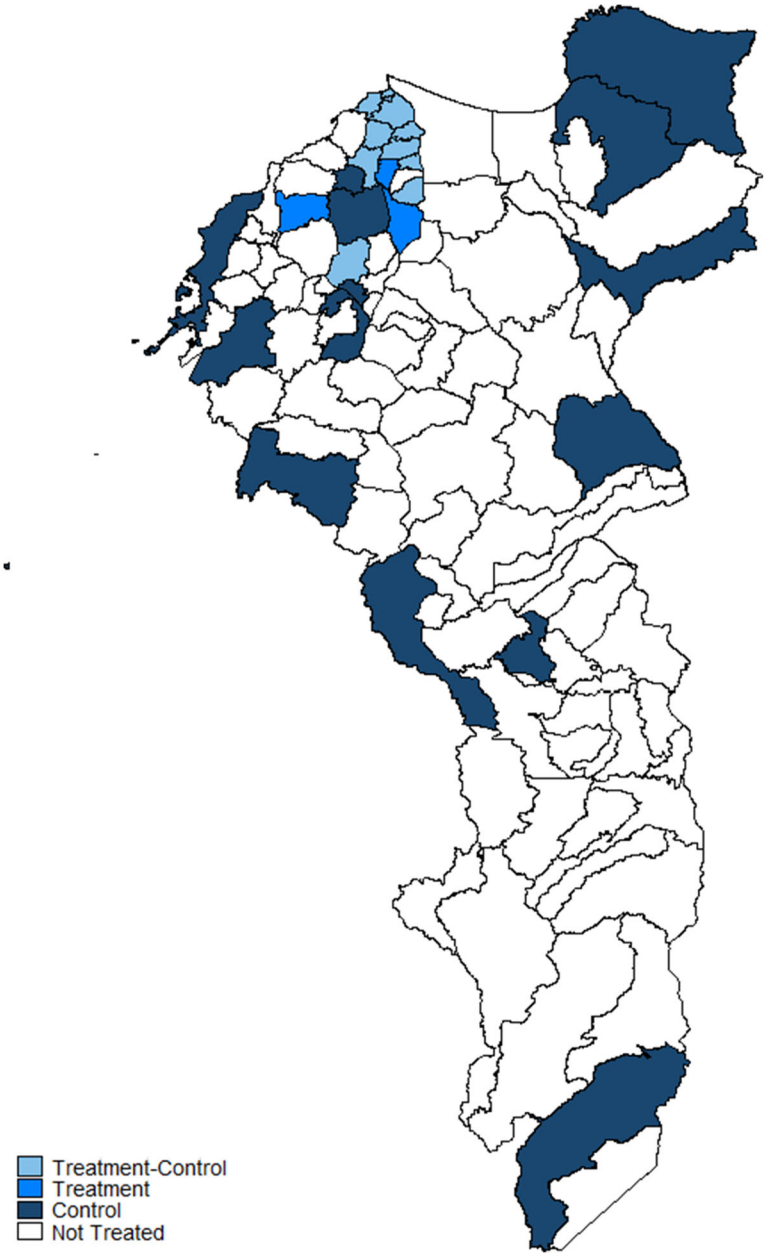


Fig. 1 Municipalities in the sampling frame. Note: Author's elaboration. "Treatment" refers to municipalities where only program participants were identified for the sampling frame. "Control" refers to municipalities where only control units were identified. "Treatment-control" refers to municipalities where both participants and non-participants were identified. The map covers the departments of Atlantico, Bolivar, and Magdalena

propensity score matching. Following (Rosenbaum & Rubin, 1985), it is recommended to include a propensity score as one additional variable in a multivariate matching design. This would allow the design to include relevant information in the matching design without concerns about the "dimensionality curse." Given that this part of the design is descriptive, I use a simple strategy to estimate the propensity score. 18 variables were used to estimate the propensity score. Table S2 in the Online Appendix presents the variables used in the estimation.

It is important to notice that this matching strategy was applied to keep the reference labor market constant. This implies applying this matching design in each municipality where treatment and control units were available in the Department of Atlantico. Each treated woman in a given municipality is then matched with three non-treated women in the same municipality. The final sampling frame for Atlantico is just the collection of all the matches from these municipalities.

Concerns about potential spillover effects led to the inclusion of municipalities in the departments of Bolivar and Magdalena. Due to a government's request, it was decided that a fraction of the final sample should include controls from these departments as a way to address potential within-municipality spillovers. Therefore, the procedure for these municipalities was based on applying the matching strategy using all the treated women in the department of Atlantico and matching them with women of similar characteristics in Bolivar and Magdalena.[25] Due to budget constraints, the final sample size is not large enough regarding observations from these two departments.[26] Moreover, this strategy is only valid if there are no department-specific shocks that differently affect control groups in each department.[27] Therefore, any evidence about spillover effects in this context should be taken with a grain of salt.

Table 1 presents descriptive and balance statistics for the sampling frame obtained after applying this matching strategy. Besides means and variances for the treatment and control groups, I also report the standardized difference, the variance ratio, and the *p*-values for *t*-student, Wilcoxon, and Fisher tests of difference in means between treatment and control groups. In general, means and variances between treated and control units are pretty similar. This is confirmed by standardized differences in means, which are—in all the cases—well below the 0.3 threshold suggested by Imbens (2015) or the 0.2 proposed by Cohen (1988).[28] Following Imai et al. (2008), I also report variance ratios to capture other moments of the empirical distribution of pre-treatment variables beyond the mean. Results are closer to 1 in most cases, which suggests that covariates are balanced in means and variances.

[25] One implication of this feature of my research design is the existence of repeated treated observations in the sampling frame. This characteristic is not present in the final sample. As it will be shown later, this does not affect the quality of the final sample.

[26] As it is well-known, power calculations for measuring spillover effects are typically very demanding. Therefore, the sampling design was designed to provide only suggestive evidence about externalities.

[27] It is unclear whether the size of within-municipality spillovers is relevant in this setting, but some international evidence is available. For the case of young girls in Uganda, Bandiera et al. (2020) present suggestive evidence of spillovers for aspirations and control over the body indices, but weaker effects for economic empowerment. Some evidence for within-household spillovers will be discussed later.

[28] Standardized differences compare the difference in means in units of the pooled standard deviation, and it has the advantage of not being influenced by sample size. See Imbens (2015) for details.

Table 1 Descriptive and balance statistics of sampling frame

Variable	Treated		Control		Standardized Difference	Variance ratio	*t*-test *p*-value	Wilcoxon *p*-value	Fisher *p*-value
	Mean (1)	Variance (2)	Mean (3)	Variance (4)	(5)	(6)	(7)	(8)	(9)
House	0.852	0.126	0.861	0.120	−0.024	1.051	0.118	0.118	0.120
Gas	0.831	0.141	0.838	0.136	−0.020	1.037	0.202	0.202	0.203
Telephone	0.127	0.111	0.108	0.096	0.058	1.148	0.000	0.000	0.000
Aqueduct	0.909	0.082	0.908	0.083	0.004	0.988	0.790	0.790	0.806
Fridge	0.583	0.243	0.571	0.245	0.025	0.992	0.119	0.119	0.123
Washing machine	0.190	0.154	0.184	0.150	0.015	1.024	0.348	0.348	0.353
Television	0.831	0.141	0.827	0.143	0.009	0.984	0.548	0.548	0.560
Cable TV	0.081	0.075	0.087	0.079	−0.021	0.940	0.190	0.190	0.195
Computer	0.037	0.035	0.034	0.033	0.017	1.088	0.286	0.286	0.282
Stereo	0.176	0.145	0.178	0.146	−0.005	0.992	0.759	0.759	0.773
Motorcycle	0.043	0.041	0.042	0.040	0.006	1.027	0.715	0.715	0.724
Health	0.755	0.185	0.761	0.182	−0.014	1.018	0.361	0.362	0.367
Basic education	0.923	0.071	0.910	0.082	0.047	0.868	0.003	0.003	0.003
Inactive	0.901	0.089	0.904	0.087	−0.010	1.028	0.516	0.516	0.524
Propensity score	0.074	0.012	0.053	0.006	0.222	2.049	0.000	0.000	NA
SISBEN score	44.321	280.082	43.388	260.017	0.057	1.077	0.000	0.001	NA
Married	0.726	0.199	0.721	0.201	0.009	0.991	0.565	0.565	0.574
Age	38.609	99.778	38.161	115.363	0.043	0.865	0.007	0.000	NA
Sampling frame size	5656	14,145							

Note: This table reports balance statistics for the sampling frame based on SISBEN's records for 2011. Columns 1 and 3 present the mean, and columns 2 and 4 present the variance of the pre-treatment variables for treated and control groups. Column 5 shows the standardized difference in means and column 6 presents the variance ratio between treated and control groups. Columns 7, 8, and 9 show the *p*-value for a *t*-test, Wilcoxon test, and Fisher test of differences in means for treated and control groups, respectively

For the sake of completeness, I also report the *p*-values for the *t*-test with similar results, except for three unbalanced variables. Due to the results for the standardized difference in means, this seems just to be driven by sample size. Because *t*-tests are based on a particular distribution, I also use the non-parametric Wilcoxon test to account for the fact that data are matched and to relax the parametric assumptions. Results do not change. Finally, I use the Fisher's exact test because many variables in my design are categorical. This test is valid regardless of the sample size. The results are similar. In sum, there is strong evidence regarding the balance of pre-treatment characteristics in the sampling frame.

A final piece of evidence on balance in the sampling frame is reported in Fig. 2. As it was mentioned above, the SISBEN score is a summary measure of socioeconomic characteristics. I assess differences in the distribution of this variable by treatment status, finding that the distributions are the same. This suggests, along with all the previous evidence, that the matching strategy to build the sampling frame was highly successful in achieving pre-treatment covariates balance.

3.2.2 Final sample

Using the matched sampling frame, the National Consulting Center (CNC) carried out the fieldwork to collect the final sample, using the standard statistical procedures for fieldwork and data collection. A sample of 1738 women was finally collected with 715 women in the treatment group. Few observations were discarded after trimming

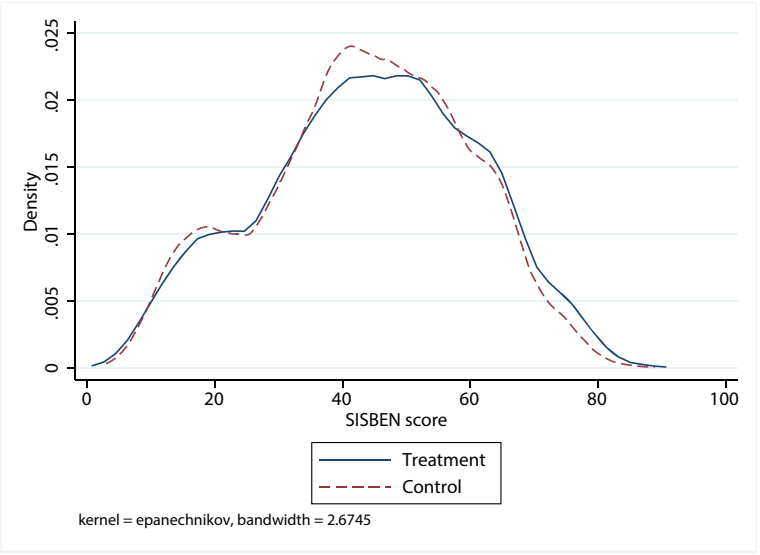


Fig. 2 Kernel density for the SISBEN score by treatment status in the matched sampling frame. Note: Author's elaboration. The SISBEN score is a welfare index used by the Colombian government for the selection of social programs' beneficiaries. Each household receives a score from 0 to 100 (from poorest to richest). The score is calculated using 24 variables across four dimensions: health, education, housing, and vulnerability

observations with extreme values of a propensity score in a procedure described in the next section.[29] The final sample for the primary analyses includes 1723 observations.

It is important to emphasize that the final sample is not representative of the population in the project's influence area. A common approach in retrospective evaluations is to collect data based on descriptive statistical designs, similar to the ones used in household surveys designed to estimate a population parameter like the poverty rate. Because this study seeks to establish a causal relationship between a treatment and a set of outcomes, the corresponding statistical design should be analytical, which implies that a set of hypotheses regarding differences between two groups is being tested. Rather than using a descriptive sample to apply techniques for analytical designs (a typical mistake in retrospective impact evaluation), this paper builds the sampling frame to produce a sample that is ex-ante balanced. Therefore, the sample was collected to construct a valid counterfactual for the treated units without any intention to be representative of a particular population.[30]

Table 2 replicates the analysis in Table 1 to check the quality of the sample.[31] In general, I find that the collected sample preserves the good properties in terms of pre-treatment balance in covariates that were achieved with the matched sampling frame. I include additional pre-treatment covariates that were collected in the survey and check whether these are balanced between treatment and control groups. Among the list of 18 covariates, only one variable is unbalanced in terms of standardized differences in means. Five variables are slightly unbalanced in terms of variance ratios (larger than 1.2), and the results for the *t*-student, Wilcoxon, and Fisher tests suggest some imbalances but, as was mentioned before, only in one case this is consistent with standardized differences in means larger than 0.3 in absolute value. Therefore, the evidence presented in Table 2 is consistent with an outstanding level of balance in pre-treatment covariates between treatment and control groups.

3.3 Estimating causal effects

The sample collected in the previous step is used in this section to estimate the causal effects of interests using the combination of Abadie and Imbens (2006) and Abadie and Imbens (2011)'s nearest neighbor matching estimator along with genetic matching to optimize the process of finding covariates balance. The methodological steps are described below.

3.3.1 Estimating the propensity score using the Imbens and Rubin's (2015) algorithm

In the first step, I estimate a propensity score using the algorithm proposed by Imbens and Rubin (2015). This algorithm selects covariates for inclusion in the propensity score's specification by choosing a set of linear terms as well as interactions and

[29] Trimming of the sample did not affect the balance in terms of pre-treatment characteristics between treatment and control groups (results not shown).

[30] For a discussion of descriptive versus analytical designs, see Roy et al. (2016).

[31] The full definition of these variables is the same as the ones in the sampling frame and is presented in Table S2 in the Online Appendix.

Table 2 Descriptive and balance statistics of trimmed sample

Variable	Treated		Control		Standardized Difference	Variance ratio	*t*-test *p*-value	Wilcoxon *p*-value	Fisher *p*-value
	Mean (1)	Variance (2)	Mean (3)	Variance (4)	(5)	(6)	(7)	(8)	(9)
House	0.880	0.106	0.882	0.104	−0.005	1.012	0.922	0.922	0.940
Gas	0.865	0.117	0.861	0.120	0.011	0.919	0.829	0.829	0.887
Telephone	0.134	0.116	0.097	0.087	0.117	0.978	0.016	0.016	0.016
Aqueduct	0.928	0.067	0.921	0.073	0.026	1.330	0.592	0.592	0.645
Fridge	0.630	0.233	0.558	0.247	0.148	0.945	0.003	0.003	0.003
Washing machine	0.207	0.165	0.162	0.136	0.118	1.213	0.015	0.016	0.018
Television	0.848	0.129	0.844	0.132	0.010	0.982	0.844	0.844	0.892
Cable TV	0.094	0.086	0.071	0.066	0.085	1.298	0.078	0.078	0.088
Computer	0.038	0.037	0.024	0.023	0.083	1.586	0.082	0.082	0.085
Stereo	0.189	0.153	0.173	0.143	0.043	1.074	0.382	0.382	0.407
Motorcycle	0.048	0.046	0.046	0.044	0.008	1.033	0.877	0.877	0.908
Health	0.760	0.183	0.792	0.165	−0.076	1.107	0.119	0.119	0.125
Basic education	0.938	0.058	0.937	0.059	0.004	0.985	0.929	0.929	1.000
Inactive	0.883	0.104	0.900	0.090	−0.056	1.153	0.248	0.248	0.267
Pscore IR	0.444	0.015	0.390	0.011	0.487	1.370	0.000	0.000	NA
SISBEN score	47.237	255.020	43.245	236.638	0.255	1.078	0.000	0.000	NA
Married	0.708	0.207	0.753	0.186	−0.102	1.113	0.036	0.036	0.040
Afro	0.237	0.181	0.208	0.165	0.069	1.098	0.155	1.155	0.157
Sample size	709	1014							

Note: This table reports balance statistics for the sample collected for the retrospective evaluation of TTM in 2017. The sample was trimmed using the propensity score following Crump et al. (2009). Columns 1 and 3 present the mean, and columns 2 and 4 present the variance of the pre-treatment variables for treated and control groups. Column 5 shows the standardized difference in means and column 6 presents the variance ratio between treated and control groups. Columns 7, 8, and 9 show the *p*-value for a *t*-test, Wilcoxon test, and Fisher test of differences in means for treated and control groups, respectively

quadratic terms based on a data-driven approach.[32] Based on this algorithm, 22 variables were used to estimate this propensity score, which I also used as a covariate for the matching process as in Rosenbaum and Rubin (1985). I also use this propensity score to trim observations with extreme values following Crump et al. (2009).[33] In particular, I drop observations with propensity scores larger than 0.9 and smaller than 0.1. Tables S3 and S4 in the Online Appendix present the definitions of variables used to estimate the propensity score and the results of this exercise applying the Imbens and Rubin (2015) algorithm, respectively.

A critical element in evaluating my research design's validity consists of checking whether the estimated propensity score fulfills the overlap condition. The evidence in Fig. 3 suggests that this is the case. Therefore, this important assumption for matching designs is valid in this setting.

3.3.2 Genetic matching for covariate balance

The second step consists of using a multivariate genetic matching design to find balance in covariates across treatment and control units in the sample. Table S5 in the Online Appendix presents the variables used to match treated and control women. I use 11 pretreatment covariates, including the SISBEN score and the propensity score previously estimated. I set the number of neighbors at 2 per treated woman, following Imbens and Rubin (2015).[34] I also use a population size of 1000 for the genetic algorithm to search for alternative weights to achieve overall balance, following a recommendation by Diamond and Sekhon (2013).[35]

Of course, controlling by only 11 covariates may seem insufficient to believe in the CIA assumption. Using the estimated propensity score as an additional covariate, as suggested by Rosenbaum and Rubin (1985) and Diamond and Sekhon (2013), allows us to incorporate information from 22 additional variables flexibly. In this way, this matching strategy accounts for the socioeconomic characteristics related to selection into the treatment but does not account for the traits and skills associated with entrepreneurship. Therefore, conditioning on a large set of covariates does not prevent unobservable variables from playing a role in explaining selection into the

[32] A detailed discussion about the steps involved in this algorithm can be found in Appendix A of Imbens (2015).

[33] One important issue is that the propensity score is a generated regressor and, as such, standard errors should account for this fact. The recommended solution for this problem is bootstrapping standard errors, but Abadie and Imbens (2008) have highlighted the problems with this alternative for matching. Abadie and Imbens (2016) study the case where matching is done exclusively on the propensity score without additional covariates—as in this paper—and show that, for the case of ATET, ignoring the estimation error has ambiguous effects on the size of confidence intervals. Because there is no clear suggestion in the technical literature about how this issue should be addressed, I proceed like the rest of the empirical literature, by ignoring this issue.

[34] Using more than one match can improve precision but also can increase bias because it increases the average covariate discrepancy between matches. Intuitively, adding a match will decrease the match quality. These authors suggest that "...going beyond two or three matches can only lead to small improvements in the sampling precision…" (Imbens & Rubin, 2015): 427. Therefore, I use two matches as suggested by these authors.

[35] Recall that the population size in genetic algorithms represents the set of potential solutions for the optimization problem of interest.

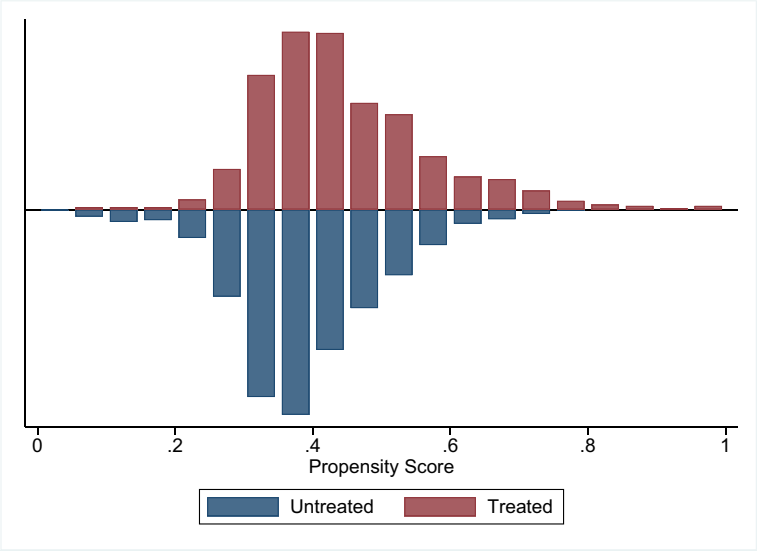


Fig. 3 Assessing overlap in the sample-based propensity score estimated using Imbens and Rubin's (2015) algorithm. Note: Author's elaboration. The propensity score was estimated using the methodology suggested by Imbens and Rubin (2015). The variables used in the estimation are described in Table S3 in the Online Appendix

treatment. By relaxing the CIA assumption using the sensitivity analysis proposed by Rosenbaum (2002), I can account for the role of the unobservable variables, as I discuss in Section 3.4.

Figure 4 assesses pre-treatment balance before and after matching. I check for balance for a list of 26 covariates for both the full and trimmed samples. I use a standardized difference in means of 0.2 in absolute value to assess balance.[36] In general, a large set of covariates was balanced even before matching and remained in this way after. This suggests that the sample was already well-balanced before using the genetic algorithm in this phase. This speaks well about the quality of the first matching process to build the sampling frame. Only three characteristics are unbalanced before applying matching in the sample. After matching, only the variable years of education is slightly above the 0.2 standard, but well below the common 0.3 standard suggested by Imbens (2015). Table 3 additionally provides comprehensive balance statistics for the trimmed sample both before and after matching. The findings are in alignment with those delineated in Fig. 4. Therefore, it can be safely concluded that matching in this phase successfully achieved balance in pre-treatment characteristics beyond the original 11 variables included in the genetic algorithm.

This good balance may be only present for differences in means but not for differences in distribution. Figure 5 assesses the distributional balance for the SISBEN score before and after matching. Again, there are very minor differences between treatment and control units before matching, implying a good quality of the sampling

[36] Using this threshold is stricter than the standard proposed by Imbens (2015). I follow this approach to be more demanding with this element of my research design.

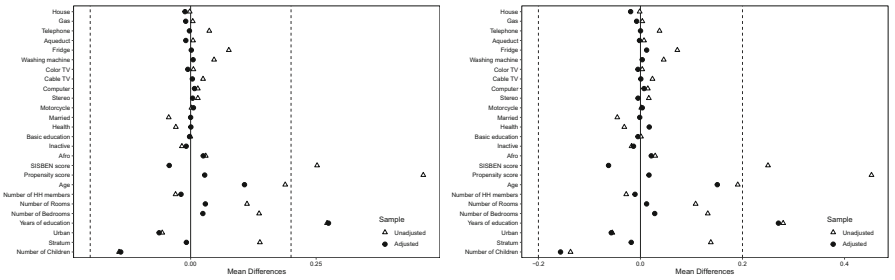


Fig. 4 Assessing balance before and after matching. Note: Author's elaboration. The variables used in this exercise are described in tables S1, S2, and S3 in the Online Appendix

frame in terms of balance. Although by a small margin, balance is improved after matching, being the distribution of the SISBEN score practically indistinguishable across treatment and control groups.

Figures S1 to S4 in the Online Appendix display, side by side, quantile-quantile plots for both the unmatched and matched samples for a set of (continuous) pre-treatment covariates in the trimmed sample. Results suggest that matching does improve distributional balance, but the gains are small because these variables were already well-balanced. Consistent with the results above, only the variable "years of education" seems to be slightly unbalanced. Figures S5 to S10 in the Online Appendix display kernel densities for a similar set of (continuous) pre-treatment covariates before and after matching, finding again that the covariates' distribution across treatment and control are almost the same, with marginal improvements after matching. In sum, the distributions of covariates seem to be very well-balanced between treatment and control women.

Once pre-treatment balance has been achieved, the next step consists of estimating the causal effects with any arbitrary matching technique. I use the bias-corrected nearest neighbor matching estimator proposed by Abadie and Imbens (2006) and Abadie and Imbens (2011) to compute the ATT given its better properties compared to alternative matching estimators (McKenzie et al., 2010). Standard errors are computed following the formula for the variance derived by Abadie and Imbens (2006). Due to a large set of outcomes in this study, I first need to discuss how multiple hypothesis testing will be handled in this paper.

3.3.3 Addressing multiple hypothesis testing

Due to women's empowerment's multidimensional nature, a large set of outcomes will be analyzed in this study. I use the correction for multiple hypothesis testing proposed by Benjamini and Hochberg (1995) to control for the false discovery rate.[37] I apply this correction for each of the sets of outcomes explored in this study.

[37] This procedure ranks the number m of hypotheses $H_{(i)}$ according to their uncorrected p-values $P_{(i)}$ and sets a proportion q of rejected null hypotheses that are erroneously rejected. Then, letting k be the largest i for which $P_{(i)} \leq \frac{i}{m}q$; the procedure rejects all $H_{(i)}, i = 1, 2, ..., k$ that violate this weak inequality. See Benjamini and Hochberg (1995) for details.

Table 3 Descriptive and balance statistics of trimmed sample—before and after matching

Variable	Before matching					After matching				
	Mean treatment (1)	Mean control (2)	Standardized difference (3)	Var. ratio (4)	*t*-test *p*-value (5)	Mean treatment (6)	Mean control (7)	Standardized difference (8)	Var. ratio (9)	*t*-test *p*-value (10)
House	0.880	0.882	−0.475	1.012	0.923	0.880	0.903	−7.087	1.206	0.049
Gas	0.865	0.861	1.066	0.978	0.828	0.865	0.876	−3.433	1.080	0.454
Telephone	0.134	0.097	10.955	1.330	0.018	0.134	0.132	0.483	1.011	0.916
Aqueduct	0.928	0.921	2.693	0.919	0.589	0.928	0.926	0.818	0.974	0.860
Fridge	0.630	0.558	14.964	0.945	0.003	0.630	0.631	−0.146	1.001	0.930
Washing machine	0.207	0.162	11.240	1.213	0.017	0.207	0.204	0.869	1.013	0.286
Color TV	0.848	0.844	0.971	0.982	0.843	0.848	0.852	−1.307	1.026	0.623
Cable TV	0.094	0.071	8.026	1.298	0.085	0.094	0.092	0.723	1.021	0.584
Computer	0.038	0.024	7.525	1.586	0.095	0.038	0.030	4.050	1.246	0.044
Stereo	0.189	0.173	4.190	1.074	0.385	0.189	0.182	1.740	1.029	0.634
Motorcycle	0.048	0.046	0.750	1.033	0.877	0.048	0.042	2.639	1.127	0.102
Married	0.708	0.753	−9.981	1.113	0.037	0.708	0.708	0.000	1.000	1.000
Health	0.760	0.792	−7.417	1.107	0.122	0.760	0.760	0.000	1.000	1.000
Basic education	0.938	0.937	0.438	0.985	0.929	0.938	0.938	0.000	1.000	1.000
Inactive	0.883	0.900	−5.427	1.153	0.254	0.883	0.881	0.658	0.985	0.891
Afro	0.237	0.208	6.784	1.098	0.158	0.237	0.211	6.022	1.085	0.196
SISBEN score	47.237	43.245	25.001	1.078	0.000	47.237	48.198	−6.014	1.039	0.044
Propensity score	0.445	0.390	45.281	1.370	0.000	0.445	0.442	1.675	1.112	0.044

Note: This table reports balance statistics before and after matching for the sample collected for the retrospective evaluation of TTM in 2017. The sample was trimmed using the propensity score following Crump et al. (2009). Columns 1, 2, 6, and 7 present the mean of the pre-treatment variables for treated and control groups before and after matching. Columns 3 and 8 show the standardized difference in means before and after matching multiplied by 100. Columns 4 and 9 present the variance ratio between treated and control groups before and after matching. Columns 5 and 10 show the *p*-value for a *t*-test of differences in means for treated and control groups before and after matching

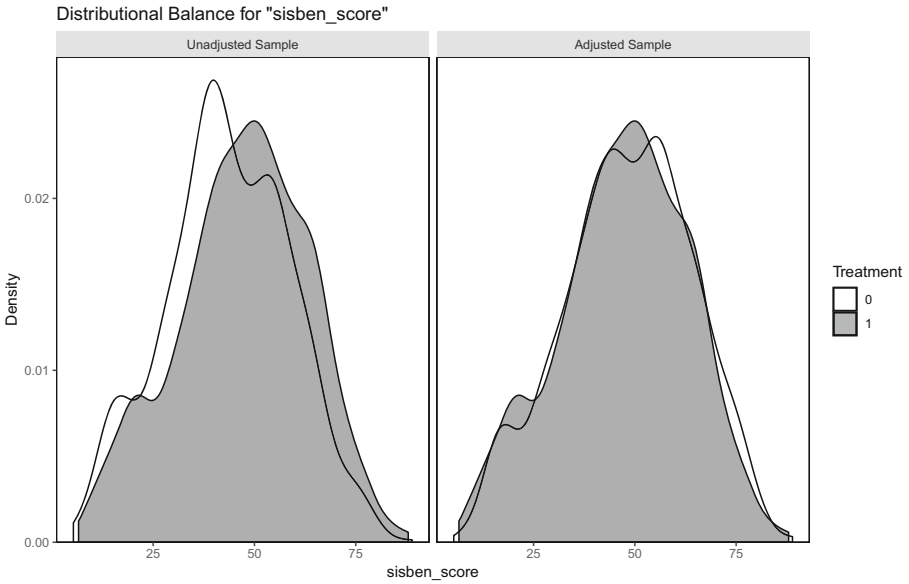


Fig. 5 Assessing distributional balance of the SISBEN score before and after matching. Note: Author's elaboration. The figure displays the difference in the distribution of the SISBEN score before and after matching for the trimmed sample

3.4 Sensitivity analysis

A final piece of this paper's research design is a sensitivity analysis to check the validity of the CIA assumption. An extensive literature shows the limitations of matching to recover causal effects because of violations of this assumption (Arceneaux et al., 2006) and Smith and Todd (2005), so it is important to relax it using a sensitivity analysis as in Rosenbaum (2002).

As discussed in Section 2.2.2, the selection into the program is based on socioeconomic characteristics and skills/traits associated with entrepreneurship. These include willingness to take risks, talent, grit, patience, and related skills that are hard to observe and measure in an ex-post evaluation design. Because these skills and traits are non-observables, the CIA assumption is violated. This violation introduces an upward bias in estimating the treatment effect because these traits and skills are usually positively associated with economic and social outcomes. Sensitivity analysis allows us to identify how large a deviation from the CIA assumption needs to be for a result to disappear.

The basic intuition behind this approach is that two matched individuals with the same observable characteristics X should have the same probability of receiving treatment e. Therefore, the ratio of probabilities should be equal to 1 between one person in the treatment group and another person in the control group. If the CIA is violated, then this ratio should be higher than 1. Rosenbaum (2002) derived bounds of the following

form using a logistic function:

$$\frac{1}{\epsilon^{\gamma}} \leq \frac{e_i(1-e_j)}{e_j(1-e_i)} \leq \epsilon^{\gamma} \tag{2}$$

where e is the propensity score for individual i or j, and γ is a sensitivity parameter. An alternative derivation based on assignment probabilities allows us to establish that $\Gamma = \epsilon^{\gamma}$ since $\frac{e_i(1-e_j)}{e_j(1-e_i)} = \epsilon^{\gamma(v_i - v + j)}$.[38] Therefore, Γ can be interpreted as the size of log of the coefficient γ for the unobserved covariate v. If $\Gamma = \epsilon^{\gamma} = 1$, two matched individuals with the same X will have the same probability of participation. On the other hand, if $\Gamma = \epsilon^{\gamma} > 1$, two matched individuals with the same X will have a different probability of participation, implying that there are unobserved factors. This exercise allows us to measure how large the deviation from the CIA needs to be to qualitatively change the original results under the assumption that the CIA is valid.[39]

This procedure is applied via a Wilcoxon sign rank test for continuous outcomes and a McNemar test for dichotomous outcomes. A set of values for Γ from 1 to 2 are chosen following Keele (2010). The goal is to evaluate how p-values change due to bias related to an unobserved confounder; in other words, how inference regarding a causal estimate is affected once confounding factors are accounted for. Upper and lower bounds on the p-value are calculated based on the test statistic for different values of Γ. A value of $\Gamma = \Gamma_0$ implies that a treated person is Γ_0 times more likely to participate in the program due to unobserved covariates than a non-treated person. Then, the procedure evaluates whether the original results in terms of p-values remain unaffected. Because the lower bound is always lower than the estimated p-value, only the upper bound is relevant. Values of Γ associated with p-values larger than the usual significance standards reflect a qualitative change regarding the original inference about the effect of TTM for a given outcome.

4 Empirical results

I organize the results according to the dimensions of empowerment that are emphasized by the academic and policy literature. Due to space constraints, the sensitivity analysis results are presented in the Online Appendix and will be briefly discussed in the next section.

[38] The equivalence of the derivation based on assignment probabilities and the original derivation based on unobserved covariates is proved in Proposition 12 of Rosenbaum (2002).

[39] See Rosenbaum (2002), Chapter 4, for details. There are three main assumptions behind this approach. First, a logit form linking treatment assignment to observed and unobserved covariates is assumed. Second, the unobserved covariates are constrained to be in the (0,1) range. Finally, there are no interaction terms between observed and unobserved covariates. The second assumption seems to be restrictive, but it has many interpretations. The most straightforward interpretation assumes the unobserved factors as dummy variables. However, an alternative interpretation is to see this assumption as a restriction regarding the scale of the unobserved covariates to provide meaning to the numerical value of the γ parameter. As suggested by Rosenbaum (2002), this restriction can be relaxed.

4.1 Economic empowerment

Table 4 presents the main ATT impacts on economic empowerment. I consider a set of economic outcomes and assets. To put results in perspective, column 1 presents the outcome mean for the control group, and column 2 the causal effect estimate. Columns 3, 4, and 5 report the standard error, the t-statistics, and the associated p-value respectively. Column 6 shows the q-value to adjust for multiple outcomes, and column 7 informs whether the null hypothesis is rejected after this adjustment. Column 8 reports the effect size.

Among the economic outcomes, I study monthly per-capita income and expenditure,[40] a multidimensional poverty indicator,[41] an indicator of employment, and social protection measured by health and pension coverage. Regarding assets, I use indicator variables for whether the household has a fridge, DVD, motorcycle, car, and washing machine.

I find important increases in monthly income and expenditure per-capita. In the case of income (row 1), the point estimate is USD 45.7 (Column 2).[42] With respect to a control group mean of USD 144.8, this effect is large (31% increase) and represents 18% of the minimum wage. This coefficient is statistically significant at the usual standards (p-value of 0.002) and robust to multiple comparisons. The effect for the monthly expenditure per-capita (row 2) is also important (coefficient of USD 30.3) and statistically significant without controlling for multiple hypothesis testing (p-value of 0.018), but it is not robust to multiple comparisons.[43]

Row 3 shows the impact of TTM on poverty reduction. Column 1 reports an estimated coefficient of -10.7 percentage points (t-statistic of -3.76). Given a mean poverty rate for the control group of 40.2%, this represents a reduction of 26.7% (column 8). This reduction is robust to accounting for multiple comparisons. Therefore, the program was highly successful in reducing poverty.[44]

Evidence regarding impacts on labor market outcomes is mixed. On the one hand, I find large impacts in the likelihood of being employed (point estimate of 21.6 percent-

[40] All monetary measures were converted from Colombian Pesos to US dollars using the exchange rate of February 2017 (USD 1= COP 2,879.57)

[41] This multidimensional poverty indicator was developed using the methodology proposed by Schreiner (2014), and it has been used in around 60 countries. It considers ten indicators of family size, education, labor supply, access to electricity, source of energy for food preparation, and household assets.

[42] The minimum wage in Colombia in 2017 was COP 737,717, about USD 256.

[43] These results are consistent with the effects of multifaceted programs for the poor. For instance, effects for income (0.20 standard deviations (SDs) for TTM) are smaller than the ones reported by Banerjee et al. (2015) (0.38 SDs). However, these effects do not distinguish women from men. A more similar case is Bandiera et al. (2017). They report a 21% increase in earnings with respect to the control group, smaller than the 31% increase in incomes estimated for TTM. Regarding other women's empowerment programs, Bandiera et al. (2020) find no effects on expenditures after four years, and Buehren et al. (2017) report null effects on labor incomes. Of course, these comparisons are suggestive since differences in categories across studies may be related to differences in data collection methods and methodological differences for capturing incomes and expenditures.

[44] This reduction was larger than the observed in some of the multifaceted programs for the poor. For instance, in Bandiera et al. (2017), this reduction was 14% after four years for poor women in Bangladesh. However, they use a monetary poverty line, so their results are not strictly comparable with those reported here.

Table 4 Effects of TTM on economic empowerment. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
Panel A: Main outcomes								
1. Monthly income	144.800	45.730	15.084	3.032	0.002	0.014	1	0.316
2. Monthly expenditure	378.400	30.306	14.135	2.144	0.032	0.018	0	0.080
3. Poverty	0.402	−0.107	0.029	−3.765	0.000	0.009	1	−0.267
4. Employed	0.414	0.216	0.029	7.324	0.000	0.005	1	0.521
5. Health coverage	0.985	−0.021	0.011	−1.894	0.058	0.023	0	−0.021
6. Pension coverage	0.115	−0.023	0.018	−1.240	0.215	0.036	0	−0.199
Panel B: Assets								
7. Fridge	0.858	0.033	0.018	1.776	0.076	0.027	0	0.038
8. DVD	0.238	0.039	0.027	1.455	0.146	0.032	0	0.164
9. Motorcycle	0.156	−0.012	0.021	−0.583	0.560	0.050	0	−0.079
10. Car	0.019	−0.006	0.010	−0.611	0.541	0.045	0	−0.331
11. Washing machine	0.559	−0.026	0.030	−0.880	0.379	0.041	0	−0.047
Observations								1723
Treated observations								709
Matched observations								709
Matched observations (unweighted)								1457

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

age points with an effect size of 52%), robust to multiple comparisons (row 4). On the other hand, there is no evidence regarding increases in health and pension coverage (rows 5 and 6).[45]

Rows 7 to 11 show the impact of the intervention on the household's asset accumulation. Although the program seems to increase the likelihood that a household has a fridge (a critical appliance in a tropical area), this change is not robust to multiple comparisons.

As discussed in Section 3.2.1, the final sample includes control units from municipalities in Atlantico (within the project influence area) and from municipalities in Bolivar and Magdalena (outside the project influence area). One obvious robustness check is to verify whether the main results are maintained when only controls from the project influence area are considered. The paper's main results are robust to restricting the sample to controls from the same municipality. For instance, this subsample's estimated effect on monthly income is 49.14 US$, slightly higher than the original estimate (45.73). All the other coefficients are similar, except for poverty and monthly expenditure (Results available upon request).

Table 5 presents the ATT estimates for labor market outcomes. Row 1 shows that the TTM caused an increase of 29% in the likelihood of being part of the economically active population (18.2 percentage points, significant at the 1% level). Rows 2 and 3 present evidence of a 20% increase in household labor supply, but this is not explained by the formal sector (rows 4 and 5). The program is also related to a reduction of 36.8% in the likelihood of having unpaid family workers (row 6) and increases the probability of doing a paid activity by 45.6% (row 7). There is no evidence of changes in the likelihood of doing a business activity due to the program (row 8).[46] Finally, the program caused a reduction of 43% in the likelihood of searching for a job (row 9), and an increase of 7.5 in the number of working hours (row 10). These results are robust to multiple comparisons. Because of the importance of employment opportunities for women's autonomy and empowerment,[47] these effects are of first order from a policy perspective.[48]

Table 6 presents evidence regarding credit and saving effects. Regarding savings, rows 1 and 2 show that TTM increases savings for participants and other household members. In the first case, the point estimate is 7.3 percentage points, with an effect size of 31.5%. The effects are larger for the participant, with a point estimate of 21.8 percentage points, associated with an effect size of 55.5%. These results are robust to multiple comparisons. Row 5 shows that these savings are not in the formal banking

[45] Notice that health coverage is almost universal in Colombia because of the existence of a subsidized regime.

[46] This result is consistent with Alfonsi et al. (2020), who also found effects in labor market outcomes without effects in entrepreneurship. One potential explanation consistent with Alfonsi et al. (2020) is that the certification provided by the program after completing phases 1 and 2 allowed participants to signal their skills.

[47] Anderson et al. (2009) propose and test a model where female autonomy depends on access to the labor market, mainly when employment opportunities are outside husbands' farms.

[48] Elsayed and Roushdy (2017) also find similar large effects in employment outcomes for women's empowerment intervention in Egypt, primarily driven by the informal sector. A key difference between their results and the ones reported in this paper is that their setting is mostly rural, whereas TTM operated in urban settings.

Table 5 Effects of TTM on labor market outcomes. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean Control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
1. Belonging to the EAP	0.629	0.182	0.026	6.930	0.000	0.010	1	0.289
2. Number of workers in the HH	1.568	0.313	0.061	5.101	0.000	0.030	1	0.200
3. Fraction of workers in the HH	0.380	0.079	0.014	5.575	0.000	0.020	1	0.208
4. Numer of formal workers in the HH	0.725	−0.014	0.050	−0.279	0.780	0.045	0	−0.019
5. Fraction of formal workers in the HH	0.170	−0.001	0.013	−0.106	0.915	0.050	0	−0.008
6. Unpaid family worker	0.516	−0.190	0.030	−6.408	0.000	0.015	1	−0.368
7. Realization of paid activity	0.495	0.226	0.028	7.996	0.000	0.005	1	0.456
8. Realization of business activity	0.043	−0.004	0.013	−0.334	0.738	0.040	0	−0.100
9. Job search	0.091	−0.039	0.015	−2.563	0.010	0.035	1	−0.431
10. Work hours	19.040	7.527	1.469	5.126	0.000	0.025	1	0.395
Observations								1723
Treated observations								709
Matched observations								709
Matched observations (unweighted)								1457

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value, respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

Table 6 Effects of TTM on credit and saving outcomes. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean Control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
Panel A: Savings								
1. Household savings	0.233	0.073	0.027	2.759	0.006	0.017	1	0.315
2. Savings (interviewed woman)	0.218	0.121	0.027	4.550	0.000	0.003	1	0.555
3. Savings in banks or corporations	0.020	0.008	0.010	0.785	0.432	0.033	0	0.376
4. Savings in microcredit institution	0.006	−0.001	0.005	−0.119	0.905	0.047	0	−0.094
5. Saving at home	0.188	0.112	0.025	4.497	0.000	0.006	1	0.596
6. Savings in informal schemes	0.003	0.004	0.004	0.906	0.365	0.031	0	1.201
Panel B: Credit								
7. Debt holdings	0.394	0.037	0.030	1.229	0.219	0.025	0	0.094
8. Lender- Bank	0.197	0.038	0.026	1.487	0.137	0.022	0	0.194
9. Lender- Microcredit	0.029	0.018	0.011	1.651	0.099	0.019	0	0.635
10. Lender-Producers	0.006	−0.001	0.005	−0.222	0.824	0.039	0	−0.197
11. Lender- Suppliers	0.053	0.003	0.015	0.189	0.850	0.042	0	0.052
12. Lender- Lenders	0.107	0.003	0.019	0.148	0.882	0.044	0	0.026
13. Lender - Friends	0.069	−0.018	0.016	−1.138	0.255	0.028	0	−0.260
Observations								1723
Treated observations								709

Table 6 continued

Variable	Mean Control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
Matched observations								709
Matched observations (unweighted)								1457

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

system. Therefore, women save more, but they keep these savings out of the formal saving system.[49] Regarding credit, there is no evidence that TTM is causally associated with changes in the likelihood of having credit. As a consequence, no relationship is found between TTM and the type of lender.

Overall, these results suggest that a multifaceted intervention such as TTM has significant impacts on economic empowerment. I have documented important effects on incomes, expenditures, and poverty. These effects are in line with the evidence of multifaceted interventions for the poor (Banerjee et al., 2015) and Bandiera et al. (2017). These effects are mainly explained by improvements in labor market access, which is consistent with the fact that TTM mostly targeted urban women. No evidence on the realization of business activities is found, which suggests that entrepreneurship is less relevant in this setting.[50] This can be consistent with a scenario where most TTM participants were "necessity entrepreneurs" who took advantage of TTM components to develop skills for a wage job, but this is hard to verify with the available information. The effects on savings are relevant because of the recent scholarship that shows that saving constraints are important for microenterprises' development (Dupas & Robinson, 2013). Interestingly, no effects were found for credit.[51]

4.2 Political and social empowerment

Table 7 presents the ATT impacts on decision-making and political empowerment. Rows 1 and 2 in Panel A present the effect of TTM on self-reported measures of household decision-making. In the first case, there is an increase in the likelihood of a woman reporting having the responsibility of making decisions in the household. The effect is 5.1 percentage points, a 17.2% change with respect to the control mean. However, this effect is not robust to multiple comparisons. In the second case, no effect is found for the likelihood of both women and men sharing decision-making in the household.

Panel B in Table 7 presents the results for political empowerment. Row 3 shows that TTM led to an increase in the likelihood of voting in the 2016 elections with respect to the control group mean. The point estimate is 7.2 percentage points, representing

[49] Austrian et al. (2020) and Buehren et al. (2017) also report large effects of empowerment intervention on savings. In the first case, the authors report an estimated coefficient of 19.3 percentage points for the treatment on the treated on the likelihood of saving (16% for the control mean at the baseline). In the second case, the authors estimate an effect of 2.8% percentage points (2% for the control mean at baseline). In the context of multifaceted programs for the poor, Banerjee et al. (2015) also find large effects for savings (156% increase with respect to the control mean).

[50] According to the evaluation survey, 64% of the program's participants reported having the same project or business activity supported by TTM in 2017. The other 36% of the program's participants reported having failed projects. About 42% of the program's participants with failed projects reported conflicts with other members of their associations as the main reason for project failure. On the other hand, about 44% reported issues related to lack of resources, credit, inputs, mentoring, and poor business environment as the main determinants for project failure.

[51] I am not the first to find positive effects on savings but no effects on credit. Buehren et al. (2017) find a similar result in Tanzania. They rationalize this result by exploring the role of spillovers in participants' social networks.

Table 7 Effects of TTM on decision-making and political empowerment. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean Control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
Panel A: Decision-making at home								
1. Woman	0.294	0.051	0.027	1.903	0.057	0.025	0	0.172
2. Both	0.556	−0.042	0.029	−1.459	0.145	0.040	0	−0.076
Panel B: Political outcomes								
3. Vote in 2016	0.570	0.072	0.029	2.462	0.014	0.020	1	0.126
4. Membership in political parties	0.057	0.051	0.015	3.300	0.001	0.010	1	0.894
5. Candidate	0.043	0.074	0.017	4.377	0.000	0.005	1	1.719
6. Political participation	0.864	0.055	0.018	3.068	0.002	0.015	1	0.063
7. Women's participation in the community	0.942	0.022	0.013	1.724	0.085	0.030	0	0.023
Observations								1723
Treated observations								709
Matched observations								709
Matched observations (unweighted)								1457

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value, respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

an increase of 12.6% with respect to the control group mean.[52] The program also led to important increases in membership in political parties (5.1 percentage points in row 4), the likelihood of being a political candidate (7.4 percentage points in row 5), and a self-reported measure of political participation (5.5 percentage points in row 6). All these effects are robust to multiple comparisons. No effect is found for a self-reported measure of women's participation in the community, although the baseline levels of this measure were already too high even in the control group (row 7).[53]

Table 8 presents the results of a list experiment implemented during the data collection process to elicit the level of exposure to IPV in this setting. This is an ideal approach to deal with sensitive issues like IPV where interviewed women may feel uncomfortable reporting episodes of violence.[54] Interviewed women were randomly assigned to one of two conditions. In condition 1, they are presented a list of five items about situations they may have experienced the last year including one sensitive item about whether their partners physically beat them. Condition 2 includes the same set of items, but excludes the sensitive one. Table S5 in the Online Appendix shows the exact question used to elicit IPV.

To estimate the effect of TTM on IPV, I run a linear regression of the total number of list outcomes reported on TTM participation and whether the surveyed woman was assigned to the sensitive item (condition 1), including an interaction between these two dummy variables on the matched sample. The coefficient of interest is this latter interaction. TTM is found to have a positive effect on IPV. The estimated coefficient is 18.3 percentage points, significant at the 10% level. Considering a mean control response in the total number of list outcomes of 0.99 in the matched sample, this represents an increase of 18.5% in the likelihood of experiencing physical violence,[55] This result suggests the existence of backlash effects, presumably associated with the positive effects of TTM on incomes and labor market participation.[56]

Table 9 shows TTM impacts on social empowerment. The outcomes of interest are dummy variables of whether the surveyed woman belongs to a set of social organizations. A list of the 10 most common forms of social organizations in the Colombian

[52] Notice that voting is voluntary in Colombia.

[53] Banerjee et al. (2015) also finds large effects on political participation in their evaluation of several graduation programs for the poor. They do not find effects on voting, but document impacts on political party membership and attendance in village meetings.

[54] This method provides additional confidentiality with respect to alternative elicitation methods, which is expected to create incentives for truthful reporting. See Blair et al. (2015) for a technical discussion. Cullen (2020) compares this technique against face-to-face questions and audio computer-assisted self-interviews. Using data from Nigeria and Rwanda, she finds that IPV rates are severely underestimated under the alternative methods. Once a list experiment is used, IPV rates increase by 100% in Rwanda and 39% in Nigeria.

[55] It is important to emphasize that this result only captures a specific form of violence. Likewise, it is estimated using a linear regression for the matched sample. This particular choice is consistent with the brand of the matching literature that interprets this method as a pre-processor (Ho et al., 2007) but I am not aware of previous uses in the empowerment literature in the way presented here.

[56] Using data for 31 developing countries, Bhalotra et al. (2020) find that an increase in the probability of employment for women is associated with a 3% increase in the probability of experiencing IPV.

Table 8 Effect of TTM on intimate partner violence. List experiment

Variable	Estimate	S.E	*t*-value	Pr(> |t|)
Treatment	−0.035	0.071	−0.499	0.618
Condition 1	−0.141	0.078	−1.810	0.071
Treatment × Condition 1	0.183	0.099	1.836	0.067
Matched observations				709

Note: This table reports the results of a list experiment to elicit sensitive behaviors, in this case, the exposure to intimate partner violence (IPV). Surveyed women were randomly exposed to one of two conditions: Condition 1 is a five-item list containing a sensitive question about IPV and condition 2 is the same list but excluding this sensitive question. The original question used in the list experiment is shown in Table S6 in the Online Appendix. A regression model with interactions between treatment status and condition 1 in the list experiment was estimated using the matched sample. The dependent variable is the total number of list outcomes reported

Caribbean was included.[57] TTM is found to have large effects on the likelihood of being a member of several of these social organizations. In particular, the program increases the probability of being a member of community action boards (8.3 percentage points), producer associations (3.7 percentage points), citizen veedurias (1.9 percentage points), volunteering organizations (7.2 percentage points), sports or cultural groups (4.8 percentage points), and women organizations (14.5 percentage points). All these effects are robust to multiple comparisons. The effect sizes associated with these effects are pretty large, suggesting increases in the likelihood of being a member of these organizations of more than 100%. Hence, the program was effective in improving participants' social capital levels.

Overall, these results suggest a mixed picture. On the one hand, the program was very successful in improving political and social participation. These are critical components of empowerment as pointed out by the literature (Milazzo & Goldstein, 2019) and Duflo (2012). Increasing women's political and social participation has been associated with a provision of public goods that better reflect women's preferences (Chattopadhyay & Duflo, 2004), raises girls' educational aspirations (Beaman et al., 2012), and increases in labor force participation (Deininger et al., 2020) and entrepreneurship (Ghani et al., 2014). Women's political participation also leads to more empowerment (Bargain et al., 2018), and involvement in community affairs (Beath et al., 2013). On the other hand, the increase in exposure to IPV implies that TTM was not able to modify the unequal structure of power within participants' homes. These results contrast with recent evidence that conditional (Buller et al., 2018) and unconditional (Haushofer et al., 2019) cash transfers reduce IPV prevalence. However, the effect of empowerment on IPV is theoretically ambiguous (Angelucci & Heath, 2020), and these results can be consistent with extractive and status threat theories of IPV.

[57] I acknowledge the limitation of measuring social capital using membership in social organizations, although there are several examples in the literature. There is no consensus about what dimensions to take into account for social capital measurement but some of the proxies used in the past include trust, norms of reciprocity, engagement in public affairs, and participation in voluntary organizations.

Table 9 Effects of TTM on social capital. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
1. Community action board	0.082	0.083	0.020	4.169	0.000	0.020	1	1.015
2. Producers' associations	0.009	0.037	0.010	3.654	0.000	0.025	1	4.151
3. Citizen veedurias	0.004	0.019	0.007	2.560	0.010	0.030	1	4.717
4. Displaced population organizations	0.056	−0.007	0.014	−0.499	0.618	0.045	0	−0.121
5. Volunteering	0.038	0.072	0.015	4.762	0.000	0.010	1	1.883
6. Parents' groups	0.088	−0.003	0.017	−0.175	0.861	0.050	0	−0.035
7. Sports or cultural groups	0.008	0.048	0.011	4.491	0.000	0.015	1	5.986
8. Women groups	0.037	0.145	0.019	7.453	0.000	0.005	1	3.909
9. Youth groups	0.006	0.006	0.006	1.015	0.310	0.040	0	1.041
10. Syndicates	0.006	0.007	0.005	1.382	0.167	0.035	0	1.176
Observations								1723
Treated observations								709
Matched observations								709
Matched observations (unweighted)								1453

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value, respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

4.3 Control over the body and psychological empowerment

Table 10 shows the effects of TTM on several proxies of women's control over the body and psychological empowerment. In Panel A, the outcomes of interest are measures of the use of different contraceptive methods and some preventive health practices like cytology and breast exam. In Panel B the outcomes are beliefs about tolerance of violence towards women, women's role as an agent of change, and a measure of self-confidence. These latter measures are interpreted as proxies of psychological empowerment.

There is no evidence of the effects of TTM on control over the body. No effect is found regarding the use of contraceptives. If something, it seems that the program reduced the use of contraceptive injections (row 7). The point estimate is −2.6 percentage points, a 65% reduction with respect to the control mean (4.1%), and robust to multiple comparisons. However, the program was effective in increasing the use of breast exams. The point estimate is 9.7 percentage points, representing an increase of 17% with respect to the control group mean. This effect is robust to multiple comparisons.

Regarding psychological empowerment, no evidence of impacts is found in any of these dimensions (Panel B). Although this does suggest the intervention was not effective in changing these beliefs, it has to be observed that the margin of improvement was small because the control group mean was already high.

These results seem to be consistent with weak or null effects of TTM on control over the body and psychological empowerment. Table S7 in the Online Appendix provides more evidence in this regard. This table presents results about the effects of TTM on attitudes towards women's rights. There is no evidence that the program affected these attitudes, although it is important to emphasize that the margin of improvement was minimal because the control mean for many of these attitudes was beyond the 95%.

However, these high levels of positive attitudes towards women's rights in this setting seem to be unrelated to changes in the allocation of responsibilities within the household. To see that, Table S8 in the Online Appendix studies the effect of TTM on time use for household chores. There is no evidence of changes in the time assigned for household chores as a consequence of the program. If something, it seems the program increased the time treated women assigned to family care (row 10). The estimated coefficient is 7.8 percentage points, representing an increase of 72% with respect to the control mean. Therefore, it seems TTM did very little or nothing in terms of changing gender inequalities within the household.

Overall, these results are consistent with the null effects of TTM on these dimensions. The evidence of the international literature is mixed in this regard. Bandiera et al. (2020) find large effects on behavior and outcomes about childbearing, marriage, and sex for girls in Uganda. On the other hand, (Elsayed & Roushdy, 2017), (Sieverding & Elbadawy, 2016), and Buehren et al. (2017) find no effect on related indicators for young girls in Egypt and Tanzania. It is possible that the difference in ages (40 years old on average for TTM participants), and the more conservative environments

Table 10 Effects of TTM on control over body and psychological empowerment. Bias-adjusted nearest neighbor matching estimator for the ATT in the trimmed sample

Variable	Mean control (1)	Estimate (2)	S.E (3)	*t*-stat (4)	*p*-value (5)	*q*-value (6)	Reject Ho (7)	Effect size (8)
Panel A: Control over body								
1. Use of contraceptive methods	0.434	−0.031	0.031	−1.004	0.315	0.022	0	−0.071
2. Use of natural method	0.002	0.001	0.003	0.218	0.827	0.050	0	0.334
3. Use of pills	0.027	−0.012	0.010	−1.285	0.199	0.017	0	−0.453
4. Use of condom	0.008	−0.005	0.006	−0.803	0.422	0.033	0	−0.631
5. Use of intrauterine device	0.007	−0.003	0.005	−0.546	0.585	0.044	0	−0.391
6. Female sterilization	0.319	0.022	0.029	0.784	0.433	0.039	0	0.070
7. Use of contraceptive injections	0.041	−0.026	0.010	−2.571	0.010	0.011	1	−0.645
8. Cytology	0.904	0.015	0.017	0.849	0.396	0.028	0	0.016
9. Breast exam	0.571	0.097	0.028	3.406	0.001	0.006	1	0.170
Panel B: Psychological empowerment								
10. Disagree with violence against women	0.956	−0.024	0.015	−1.597	0.110	0.035	0	−0.025
11. Woman as agent of change	0.869	−0.002	0.021	−0.075	0.940	0.050	0	−0.002
12. Self-confidence	0.760	0.036	0.026	1.375	0.169	0.045	0	0.047
Observations								1723
Treated observations								709
Matched observations								709
Matched observations (unweighted)								1449

Note: This table reports the results of implementing the bias-adjusted nearest neighbor matching estimator proposed by Abadie and Imbens (2006, 2011) to estimate the average treatment effect on the treated (ATT). Balance was achieved using the multivariate genetic matching algorithm proposed by Diamond and Sekhon (2013) using the variables described in Table S5 in the Online Appendix. The genetic matching algorithm was based on a population size of 1000 and a number of neighbors equal to 2. The sample was trimmed following Crump et al. (2009). Multiple outcomes are corrected by adjusting for the false discovery rate as suggested by Benjamini and Hochberg (1995). Dependent variables are reported in rows. Column 1 presents the outcome mean for the control group and column 2 is the causal effect estimate. Columns 3, 4, and 5 report the standard error, the *t*-statistics, and the associated unadjusted *p*-value, respectively. Column 6 shows the *q*-value and column 7 informs whether the null hypothesis is rejected after correcting for multiple comparisons. Column 8 reports the effect size

in these countries, can help to explain the differences in results with Colombia. As mentioned before, the control group's baseline levels in many of these indicators were already high, which helps to explain the lack of significant differences. This issue deserves more research in the future.

5 Additional results: consequences of empowerment

In this section, I explore some consequences of empowerment. In particular, I pay attention to the potential effects of TTM on proxies of investments and economic vulnerability. First, I study whether TTM participation is related to housing quality improvements, an important dimension of investment for poor households. Table S9 in the Online Appendix presents results for different house amenities. There is evidence in terms of better walls (row 2) and water (row 6) quality for treated women's households, but these effects are not robust to multiple comparisons.

Second, I explore the role of TTM in explaining two dimensions of vulnerability: food security and reliance on social programs. As a consequence of improvements in empowerment, less exposure to food insecurity is expected. The evidence reported in Table S10 in the Online Appendix suggests that some dimensions of food insecurity are reduced as a consequence of TTM. For instance, there is a reduction of 5.4 percentage points in the likelihood of reporting being worried about the possibility of no having food in the last 30 days (row 1), a reduction of 6.4 percentage points in the likelihood of reporting no consuming healthy foods in the last 30 days (row 3), and a reduction of 4.5 percentage points in the likelihood of no having a meal in the last 30 days (row 5). However, these effects are weak and not robust to multiple comparisons.

Regarding the dependency on social programs, one possibility is that TTM participants are less likely to depend on social programs. On the other hand, it could be possible that TTM can improve the information that participants have regarding access to social programs, leading to vulnerable participants being more likely to enroll in social programs. Recall that the sample is composed of women who are considered vulnerable because they are part of the SISBEN system. Table S11 in the Online Appendix presents results regarding access to social programs. TTM increased the likelihood of being a beneficiary of the flagship conditional transfer program "Families in Action." The estimated coefficient is 5.5 percentage points (a 15.6% increase with respect to the control group mean), but it lacks robustness when controlling for multiple comparisons. TTM is also causally associated with a reduction of 15.4 percentage points in the likelihood of receiving subsidies for displaced populations. This represents a reduction of 33.2% with respect to the control group mean. This effect is robust to multiple comparisons.

In sum, TTM seems to have mixed effects regarding the consequences of empowerment. There is some evidence in terms of better housing conditions and less exposure to vulnerability. However, these effects are weak and not robust to multiple comparisons in some cases.

6 Sensitivity analysis

This section discusses the sensitivity of the main results to violations of the CIA assumption by applying the sensitivity analysis proposed by Rosenbaum (2002) for the large set of outcomes analyzed in this study. This exercise aims to assess whether results that are statistically significant under the validity of the CIA remain in this way once a degree of violation of the CIA is allowed. A result would be considered robust if it remains statistically significant after this degree of CIA violation is introduced otherwise results are interpreted as very sensitive to relaxing the CIA assumption.

Tables S12-S31 in the Online Appendix show the results of this exercise. Because non-significant results are not affected by this analysis, the rest of the section will focus on those outcomes that were found statistically significant under the validity of the CIA assumption. Recall that a value of $\Gamma = 1$ implies that a treated person is equally likely to receive the treatment than someone in the control group. In other words, unobserved covariates play no role in explaining differences in treatment participation. On the other hand, a value of $\Gamma = 1.5$ implies that a treated person is 50% more likely to participate in the program due to unobserved covariates with respect to someone in the control group with the same observed covariates.

Tables S12 and S13 in the Online Appendix report the sensitivity analysis results for the economic empowerment outcomes originally discussed in Table 4. Lower and upper bounds for p-values for values of Γ from 1 to 2 are reported for each outcome variable. Because lower-bound p-values are always below the unconfounded p-value, the emphasis will be placed on the upper-bound ones. In the case of the monthly income variables, the unconfounded p-value is 0.00. Then, for values of Γ larger than 1.4, the estimated coefficient is no longer significant. In other words, when matched treated units are 1.4 times more likely to receive the treatment due to unobserved covariates than control units, the monthly income variable is no longer statistically significant. Because this violation of the CIA is large (40% higher chance of receiving treatment due to unobservables for treated units), this result suggests that the estimated effect is fairly robust.

The result for the monthly expenditure variable is less robust than the case of income. The original coefficient is no longer significant for values of Γ larger than 1.1, a minor violation of the CIA assumption. Not surprisingly this variable, although significant using the uncorrected p-value, was not robust to multiple comparisons. For the poverty and employed indicator variables, violations of CIA associated with 40% and 100% larger probabilities of receiving treatment due to unobservables are needed to modify the original inference about the statistical significance of these coefficients. These results suggest fairly and strongly robust estimated effects, respectively. The evidence for health and pension coverage shows that both variables are very sensitive to departures of unconfoundedness.

Due to space constraints, no detailed discussion about the sensitivity of the effects of TTM on economic, political, and social empowerment outcomes (Tables S14-S21 in the Online Appendix) as well as the control over the body and psychological empowerment outcomes (Tables S22-S27 in the Online Appendix) is provided. In summary, those outcomes for which effects are statistically strong are typically related to larger values of Γ for results to become no longer significant due to violations of the CIA

assumption. This speaks well about the validity of this paper's research design. For the sake of completeness, the sensitivity analysis for the additional results on the consequences of empowerment is also reported (Tables S28-S31 in the Online Appendix).[58]

7 Conclusion

This paper has analyzed the long-term effects of TTM, a multifaceted intervention designed to address the multidimensional nature of women's empowerment in the Colombian Caribbean. This intervention addressed several economic, political, social, and psychological barriers via the provision of training on soft and vocational skills, self-esteem, and women's rights. It also included mentoring and in-kind seed capital. Compared to the average intervention studied in the empowerment literature, this program offered a more integral approach that tackled interlinked constraints to women's well-being.

I find that TTM caused large effects on many relevant economic, political, and social dimensions of women's empowerment, but failed to affect the structure of power within the participants' households. TTM increased household incomes and labor market outcomes, causing a reduction in multidimensional poverty. The program also led to an increase in savings. Regarding political and social empowerment, the program led to more political participation and involvement in social organizations and public affairs. However, the program did not affect the use of contraceptives, self-confidence, beliefs about women's rights, or the allocation of time for household chores. Even worse, TTM is associated with a sizable positive effect on IPV. In a way, the program could not enter participants' homes despite its large effects on outside-home economic, social, and political activities.

These results are consistent with related interventions that have found important economic, political, and social effects with limited or null impacts in terms of beliefs, the within-household decision-making, and allocation of power (Beath et al., 2013), (Elsayed & Roushdy, 2017), (Austrian et al., 2020), and Sieverding and Elbadawy (2016). From a policy perspective, this implies that multifaceted programs might not be enough in settings where social norms and attitudes do not favor gender equality. Complementary interventions targeted to men can be useful in this regard. For instance,

[58] Alternative parametric methodologies in the context of linear regressions are available to explore the sensitiveness of my main results. However, these methods are inappropriate for a fully non-parametric technique like matching. One common approach is based on the inclusion of additional covariates to explore how their inclusion changes the estimated effects in linear regressions. As emphasized by Oster (2019), one limitation of this approach is that coefficient movements alone are not sufficient statistics to approximate the role of omitted variables. It is also important to consider how much of the variance in the outcome is explained by the inclusion of covariates. In other words, when controls are included, omitted variable bias is proportional to coefficient movements, but only if these movements are scaled by the change in the R-squared. Oster (2019) offers a methodology to address these issues. This method can be complemented with an approach that selects these covariates using machine learning methods like double lasso (Urminsky et al., 2019) so I can address concerns about ad-hoc variable selection. I implemented these methods for monthly incomes in the final sample and found that my main results are robust to alternative sets of control variables. This result is robust even when the double lasso selects covariates. In all these cases, I found after implementing the methods in Oster (2019) that the unobservable variables should be between 5.68 and 8.01 times as important as the observables to obtain a null effect. Although these results depend on a parametric approach, the evidence is consistent with my sensitivity analysis. Results of these exercises are available upon request.

programs that tackle the support of restrictive gender norms among boys can be a profitable long-term investment (Dhar et al., 2020). On the other hand, innovations to modify adult males' beliefs about gender equality and conformity with discriminatory norms are of first-order relevance. However, there is very little written about the effectiveness of this kind of approach.[59] Smarter policy design motivated by behavioral science can also modify gender norms, as some recent evidence has shown (Bursztyn et al., 2020).

In sum, women's empowerment interventions and gender-equalizing policy reforms require carefully thinking about innovative ways to incorporate men if their goal is to obtain sustainable results. Designing and evaluating such interventions constitutes a fundamental area of future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00148-024-00987-z.

Acknowledgements The authors would like to thank the anonymous referees for helpful comments and suggestions. I thank editor Alfonso Flores-Lagunes, two anonymous reviewers, Darwin Cortes, Lorena Trujillo, and seminar audiences at the Universidad del Rosario and Universidad del Pacifico for helpful comments and advice. I especially thank Cristina Querubin and Carlos Castro as well as the CNC's team for amazing field work. Juliana Aragon, Maria Paula Medina, and Julian Peña provided excellent research assistance.

Funding Data collection for this project was funded by the National Planning Department (DNP) and carried out by the Centro Nacional de Consultoria (CNC).

Declarations

Conflict of interest The author declares no competing interests.

References

Abadie A, Imbens G (2006) Large sample properties of matching estimators for average treatment effects. Econometrica 74(1):235–267

Abadie A, Imbens G (2008) On the failure of bootstrap for matching estimators. Econometrica 76(6):1537–1557

Abadie A, Imbens G (2011) Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics 29(1):1–11

Abadie A, Imbens G (2016) Matching on the estimated propensity score. Econometrica 84(2):781–807

Alfonsi L, Bandiera O, Bassi V, Burguess R, Rasul I, Sulaiman M, Vitali A (2020) Tackling youth unemployment: evidence from a labor market experiment in Uganda. Econometrica 88(6):2369–2414

Anderson S, Eswaran M (2009) What determines female autonomy? Evidence from Bangladesh. J Dev Econ 90:179–191

Angelucci M, Heath R (2020) Women empowerment programs and intimate partner violence. AEA Papers and Proceedings 110:610–614

Arceneaux K, Gerber A, Green D (2006) Comparing experimental and matching methods using a large-scale voter mobilization experiment. Polit Anal 14(1):37–62

Ashraf N, Bau N, Low C, McGinn K (2020) Negotiating a better future: how interpersonal skills facilitate intergenerational investment. Quart J Econ 135(2):1095–1151

[59] In Colombia, the Government of Atlantico developed a TTM's sister program for men called "Transformate Tu Hombre" in 2013. The program has a similar structure as TTM and includes a training component designed to address gender stereotypes and IPV. To date, this program has not been evaluated.

Aslam M (2013) Empowering women: education and the pathways of change. Background paper for the Education for All Global Monitoring Report

Austrian K, Soler-Hampejsek E, Behrman J, Digitale J, Jackson N, Bweupe M, Hewett P (2020) The impact of the Adolescent Girls Empowerment Program (AGEP) on short and long-term social, economic, education and fertility outcomes: a cluster randomized controlled trial in Zambia. BMC Public Health 20(349):1–15

Baird S, Ozler B (2016) Sustained effects on economic empowerment of interventions for adolescent girls: existing evidence and knowledge gaps. CDG Background Paper

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. Am Econ J Appl Econ 12(1):210–259

Bandiera O, Burguess R, Das N, Gulesci S, Rasul I, Sulaiman M (2017) Labor markets and poverty in village economies. Quart J Econ 132(2):811–870

Banerjee A, Duflo E, Goldberg N, Karlan D, Osei R, Pariente W, Shapiro J, Thuysbaert B, Udry C (2015) A multifaceted program causes lasting progress for the very poor: evidence from six countries. Science 348(6236):1260799

Bargain O, Boutin D, Champeux H (2018) Women's political participation and intrahousehold empowerment:. IZA Discussion Paper Series 11534

Beaman L, Duflo E, Pande R, Topalova P (2012) Female leadership raises aspirations and educational attainment for girls: a policy experiment in India. Science 335(6068):582–586

Beath A, Christia F, Enikolopov R (2013) Empowering women through development aid: evidence from a field experiment in Afghanistan. American Political Science Review 107(3):540–557

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57(1):289–300

Bhalotra S, Kambhampati U, Rawlings S, Siddique Z (2020) Intimate partner violence. The influence of job opportunities for men and women. World Bank Policy Research Working Paper 9118

Blair G, Imai K, Zhou Y (2015) Design and analysis of the randomized response technique. J Am Stat Assoc 110(511):1304–1319

Bott S, Guedes A, Ruiz-Celis A, Mendoza J (2019) Intimate partner violence in the Americas: a systematic review and reanalysis of national prevalence estimates. Pan Am J Public Health 43:1–12

Brooks W, Donovan K, Johnson T (2018) Mentors or teachers? Microenterprise training in Kenya. Am Econ J Appl Econ 10(4):196–221

Buehren N, Goldstein M, Gulesci S, Sulaiman M, Yam V (2017) Evaluation of an adolescent development program in Tanzania. World Bank Policy Research Working Paper 7961

Buller A, Peterman A, Ranganathan M, Bleile A, Hidrobo M, Heise L (2018) A mixed-method review of cash transfers and intimate partner violence in low-and middle-income countries. World Bank Research Observer 33(2):218–258

Bursztyn L, Gonzales A, Yanagizawa-Drott D (2020) Misperceived social norms: women working outside home in Saudi Arabia. American Economic Review 110(10):2997–3029

Buvinic M, Furst-Nichols R (2016) Promoting women's economic empowerment: what works? World Bank Research Observer 31(1):59–101

Buvinic M, O'Donnell M (2019) Gender matters in economic empowerment interventions: a research review. World Bank Research Observer 34(2):309–346

Campos F, Frese M, Goldstein M, Lacavone L, Johnson H, McKenzie D, Mensmann M (2017) Teaching personal initiative beats traditional training in boosting small businesses in West Africa. Science 357(6357):1287–1290

Chattopadhyay R, Duflo E (2004) Women as policy makers: evidence from a randomized policy experiment in India. Econometrica 72(5):1409–1443

CNC (2017) Informe de resultados y recomendaciones de la evaluacion de Transformate Tu Mujer. Producto 4. Centro Nacional de Consultoria

Cohen J (1988) Statistical power analysis for the behavioral sciences (Second ed.). Lawrence Erlbaum Associates

Colson K, Rudolph K, Zimmerman S, Goin D, Stuart E, van der Laan M, Ahern J (2016) Optimizing matching and analysis combinations for estimating causal effects. Nature Scientific Reports 6:(23222)

Croke K, Goldstein M, Holla A (2017) Can job training decrease women's self-defeating biases? Experimental evidence from Nigeria. World Bank Policy Research Working Paper 8141

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96(1):187–99
Cullen C (2020) Methods matters. Underreporting of intimate partner violence in Nigeria and Rwanda. World Bank Policy Research Working Paper 9274
de Mel S, McKenzie D, Woodruff C (2008) Returns to capital in microenterprises: evidence from a field experiment. Quart J Econ 123(4):1329–1372
de Mel S, McKenzie D, Woodruff C (2012) One-time transfers of cash or capital have long-lasting effects of microenterprises in Sri Lanka. Science 335(6071):962–966
Deininger K, Jin S, Nagarajan H, Singh S (2020) Political reservation and female labor force participation in rural India. World Bank Policy Research Working Paper 9350
Dhar D, Jain T, Jayachandran S (2020) Reshaping adolescents' gender attitudes: evidence from a school-based experiment in India. Mimeo
Diamond A, Sekhon J (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95(3):932–945
Doepke M, Tertilt M (2011) Does female empowerment promote economic development? World Bank Policy Research Working Paper 5714
Duflo E (2012) Women empowerment and economic development. Journal of Economic Literature 50(4):1051–1079
Dupas P, Robinson J (2013) Savings constraints and microenterprise development: evidence from a field experiment in Kenya. Am Econ J Appl Econ 5(1):163–192
ECLAC (2020) Gender equality observatory for Latin America and the Caribbean. https://oig.cepal.org/
Edmonds E, Feigenberg B, Leight J (2021) Advancing the agency of adolescent girls. Technical report, Darmouth College, University of Illinois, and IFPRI
Elsayed A, Roushdy R (2017) Empowering women under social constraints: evidence from a field intervention in rural Egypt. IZA Discussion Paper Series 11240
Fafchamps M, McKenzie D, Quinn S, Woodruff C (2014) Microenterprise growth and the flypaper effect: evidence from a randomized experiment in Ghana. J Dev Econ 106:211–226
Fonseca A (2019) Informe de empoderamiento economico de las mujeres en Colombia. Situacion actual y recomendaciones de politica, Equidad de la Mujer, Gobierno de Colombia
Fox L, Romero C (2017) In the mind, the household, or the market? Concepts and measurement of women's economic empowerment. World Bank Policy Research Working Paper 8079
Ghani E, Kerr W, O'Connell S (2014) Political reservations and women's entrepreneurship in India. J Dev Econ 108:138–153
Haushofer J, Ringdal C, Shapiro J, Wang X (2019) Income changes and intimate partner violence: evidence from unconditional cash transfers in Kenya. NBER Working Paper 25627
Heath R, Jayachandran S (2018) The causes and consequences of increased female education and labor force participation in developing countries. In: Averett S, Argys L, Hoffman S (eds) Oxford handbook of women and the economy. Oxford University Press
Ho D, Imai K, King G, Stuart E (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 15(3):199–236
IDEA, PNUD, and ONU MUJERES (2019) Colombia: La hora de la paridad. IDEA and PNUD and ONU MUJERES
Imai K, King G, Stuart E (2008) Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society. Series A 172 (Issue 2):481–502
Imbens G (2015) Matching methods in practice: three examples. J Hum Resour 50(2):373–419
Imbens G, Rubin D (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press
Jayachandran S (2015) The roots of gender inequality in developing countries. Annual Review of Economics 7:63–88
Jayachandran S (2020) Social norms as a barrier to women's employment in developing countries. Northwestern University
Keele L (2010) An overview of rbounds: an r package for Rosenbaum bounds sensitivity analysis with matched data. Mimeo
Klasen S (2019) What explains uneven female labor force participation levels and trends in developing countries? World Bank Research Observer 34(2):161–197
Lafortune J, Riutort J, Tessada J (2018) Role models or individual consulting? The impact of personalizing micro-entrepreneurship training. Am Econ J Appl Econ 10(4):222–245

McKenzie D (2017) How effective are active labor market policies in developing countries? A critical review of recent evidence. World Bank Research Observer 32(2):127–154
McKenzie D (2021) Small business training to improve management practices in developing countries: re-assessing the evidence for "training doesn't work". Oxf Rev Econ Policy 2021(2):276–301
McKenzie D, Stillman S, Gibson J (2010) How important is selection? Experimental vs. non-experimental measures of the income gains from migration. Journal of the European Economic Association 8(4):913–945
Milazzo A, Goldstein M (2019) Governance and women's economic and political participation: power inequalities, formal constraints and norms. World Bank Research Observer 34(1):34–64
Montalvao J, Frese M, Goldstein M, Kilic T (2017) Soft skills for hard constraints: evidence from high-achieving female farmers. World Bank Policy Research Working Paper 8095
Mujeres ONU (2018) El progreso de las mujeres en Colombia. Transformar la economia para realizar los derechos, ONU MUJERES
ONU MUJERES (2020) Boletin estadistico: empoderamiento economico de las mujeres en Colombia. ONU MUJERES
Oster E (2019) Unobservable selection and coefficient stability: theory and evidence. Journal of Business and Economic Statistics 37(2):187–204
Prennushi G, Gupta A (2014) Women's empowerment and socio-economic outcomes. Impacts of the Andra Pradesh rural poverty reduction program. World Bank Policy Research Working Paper 6841
Rosenbaum P (2002) Observational Studies. Springer
Rosenbaum P, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33–38
Roy T, Acharya R, Roy A (2016) Statistical survey design and evaluating impact. Cambridge University Press
Sanchez M (2018) Empoderamiento educativo de las mujeres. Situacion actual y lineas de politica, Equidad de la Mujer, Gobierno de Colombia
Schreiner M (2014) The process of poverty-scoring analysis. Available at http://www.simplepovertyscorecard.com/
Sekhon J, Mebane W (1998) Genetic optimization using derivatives: theory and application to nonlinear models. Polit Anal 7:189–203
Sen A (1999) Development as freedom. Alfred Knopf
Sieverding M, Elbadawy A (2016) Empowering adolescent girls in socially conservative settings: impacts and lessons learned from the Ishraq program in rural upper Egypt. Stud Fam Plann 47(2):129–144
SMEG (2020) Transformate Tu Mujer. Secretaria de las Mujeres y la Equidad de Genero de la Gobernacion del Atlantico. https://www.atlantico.gov.co/index.php/secretarias/mujeres-y-equidad
Smith J, Todd P (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics 125(Issues 1–2):305–353
Upadhyay U, Gipson J, Ciaraldi E, Withers M, Fraser A, Prata N, Lewis S, Huchko M (2014) Women's empowerment and fertility: a review of the literature. Soc Sci Med 115:111–120
Urminsky O, Hansen C, Chernozhukov V (2019) Using double-lasso regression for principled variable selection. University of Chicago
Vyas S, Watts C (2009) How does economic empowerment affect women's risk of intimate partner violence in low and middle income countries? A systematic review of published evidence. J Int Dev 21(5):577–602
Woodruff C, McKenzie D (2012) Improving the productivity and earnings of women owned and/or managed enterprises in developing countries: What works? Available at http://www.womeneconroadmap.org/
World Bank (2012) Gender equality and development. World Development Report
World Economic Forum (2019) Global Gender Gap Report 2020. World Economic Forum

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.