Comparing the Results of Youth Training Programs in Latin America and the Caribbean

Oscar A. Mitnik
Laura Ripani
David Rosas-Shady

Office of Strategic Planning and Development Effectiveness

Labor Markets Division

DISCUSSION
PAPER Nº
IDB-DP-484

October, 2016

Comparing the Results of Youth Training Programs in Latin America and the Caribbean

Oscar A. Mitnik
Laura Ripani
David Rosas-Shady

Inter-American Development Bank

October, 2016

IDB
Inter-American Development Bank

1300 New York Ave, NW Washington, DC 20577

Oscar A. Mitnik (omitnik@iadb.org); Laura Ripani (laurari@iadb.org); David Rosas-Shady (davidro@

Comparing the Results of Youth Training Programs in Latin America and the Caribbean*

Oscar A. Mitnik† Laura Ripani‡ David Rosas-Shady§

October 26th, 2016

Abstract

The evidence on the effectiveness of youth training programs in Latin American and the Caribbean (LAC) tends to be encouraging regarding the quality of employment of beneficiaries (positive impacts are observed regarding the access to formal employment), although there is significant heterogeneity across countries and by gender of the beneficiaries. It is not clear how easily one can generalize from the results of an impact evaluation in a particular country. We address the underlying heterogeneity in the characteristics of the beneficiaries of youth training programs in LAC by relying on the individual-level data used in the experimental impact evaluations of three of these programs. We show that we can identify, and characterize, the individuals satisfying a common support condition, i.e. those who are similar across programs. We use non-experimental multiple treatment estimators to eliminate differences across programs, which work better for men than for women. For men satisfying common support (i.e. comparable), who have worse initial conditions than non-comparable individuals, the positive treatment effects on formality disappear for some programs. The results highlight the importance of treatment effect heterogeneity, which may have implications for targeting and program design. They also make explicit the limits to the external validity of each of the experiments, and how difficult is the interpretation of the results from meta-analysis studies. By stressing for which types of individuals it *is* possible to make comparisons across programs, our study points towards a more nuanced comparison of impact evaluation studies of youth training programs.

JEL classification: J24, J38, J64, O15, O17
Keywords: impact evaluation, labor market outcomes, randomized controlled trial, youth training, Colombia, Dominican Republic, Peru

* We want to thank Juan Mejalenko for superb research assistance. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

† Inter-American Development Bank, omitnik@iadb.org.

‡ Inter-American Development Bank, laurari@iadb.org.

§ Inter-American Development Bank, davidro@iadb.org.

1 Introduction

Governments in developing countries spend substantially in work-force skills development programs.[1] Despite considerable evidence on the results of these programs it is difficult to form definite conclusions on their effectiveness, given the ample heterogeneity in results. This is particularly true of programs oriented to develop the skills of the unemployed or vulnerable workers. The evidence from the literature points towards training programs for those groups as being more effective for adults and women, than for young people, with heterogeneity between regions. Training programs to improve the skills of unemployed or vulnerable groups tend to give better results in terms of employment and income for the adult population than for youth, and within adults for women than for men (Dar and Tzannatos, 1999; Betcherman, Olivas and Dar, 2004; Kluve, 2010; Card, Kluve and Weber, 2010). These programs have small impacts (and in some cases even negative) in the short term but increase (become positive) over time (Card, Kluve and Weber, 2015). Regarding youth programs, most of the international literature shows little or no significant impacts of training programs for out-of-school youth.

In Latin America and the Caribbean (LAC), the results from impact evaluations tend to be more encouraging regarding impact of training programs on the quality of employment of young people, although there is significant heterogeneity across countries, and by gender of the beneficiaries. Positive impacts are observed regarding the access to formal employment (i.e. employment that gives access to social security and/or health insurance benefits), both in the short term (Gonzalez-Velosa, Ripani and Rosas-Shady, 2012) and the long term (Ibarraran, Kluve, Ripani and Rosas-Shady, 2015; Attanasio, Guarin, Medina and Meghir, 2015; Kugler, Kugler, Herrera and Saavedra, 2015; Diaz and Rosas-Shady, 2016). These results may be explained by the fact that the programs that have been

[1] For example, in the case of Latin America and the Caribbean, the average expenditures in active labor market policies in 2010 was between 0.2% and 0.3% of GDP, but as high as 0.5% of GDP in countries like Argentina, Brazil and Chile, which compares to 0.5% to 0.7% of GDP on average for OECD countries between 2004 and 2013 (OECD, 2013). Between 2002 and 2012 the World Bank supported skills-development projects in 93 countries at a cost of almost $9 billion (Twose, 2015). In Latin America and the Caribbean, the portfolio associated to labor market programs (which is broader than skills programs) for the World Bank was around US $550 million between 2009 and 2013, while for the Inter-American Development Bank was close US $900 million for the same period.

evaluated in LAC try to be demand-driven and offer financial incentives to employers and beneficiaries (González-Velosa, Ripani and Rosas-Shady, 2012; ILO, 2016; Berniell and Mata, 2016).

A recent meta-analysis study by Kluve (2016) of training programs in LAC finds high heterogeneity in the results of the training programs, similar to the results in Card et al (2010, 2015) worldwide. Additionally, a recent meta-analysis (Escudero, Kluve, López Mourelo and Pignatti, 2016) that includes a wider set of outcomes for LAC programs finds that training programs in LAC have positive impacts on formality (almost 70 percent of impact estimates reported for formality are statistically significant and positive). As in the global meta-analyses of Card et al. (2010, 2015), estimates for female program participants are significantly more likely to be positive than for men. In comparison with older workers, estimates for youth participants are significantly more likely to be positive, which highlights the importance of youth training programs. Furthermore, there are questions on how much can be generalized about the effectiveness of interventions in general through meta-analysis (Vivalt, 2016). Both meta-analysis studies and systematic reviews need to mostly ignore by design the heterogeneity in beneficiaries, program characteristics and context.[2]

In this paper we contribute to the literature that relies on meta-analysis to obtain average effects across programs by addressing explicitly the underlying heterogeneity in the characteristics of the beneficiaries of youth training programs in LAC. For this we rely on the individual-level data used in the available experimental impact evaluations of these programs. Thus, our approach is similar in spirit to that of the study by Hotz, Imbens and Mortimer (2005) and, in particular, by Flores and Mitnik (2013), who use individual-level data and rely on a *selection on observables* or *unconfoundedness* assumption to eliminate the differences in observed characteristics of individuals in different locations. Our paper contributes to the growing literature interested in understanding the external validity of experiments (see Dehejia, Pop-Eleches and Samii, 2015, for a different methodological approach to a similar question), the role of contextual effects (see Allcott, 2012) and the information they provide for policy decision-making (see Pritchett and Sandefur, 2013; Banerjee, Chassang and Snowberg, 2016).

[2] However, it is possible in meta-analysis to control for aggregate characteristics of the beneficiaries and the context of the interventions.

We use the data from three randomized controlled trials (RCTs) conducted in Colombia, the Dominican Republic and Peru, which had somewhat different findings**.** We show the degree to which the beneficiaries of these programs are comparable, and show how the treatment effects change when only the comparable individuals are used in estimation. Our analysis is based on adjusting for differences in individual-level characteristics of the participants in the programs. We rely on the empirical strategy followed by Flores and Mitnik (2013) who show that it is possible to eliminate the differences in individual characteristics for the control individuals in several labor market RCTs in the U.S. using methods from the multiple-treatment literature. We follow the same strategy for three programs simultaneously, and show that we can identify a subset of individuals that are comparable in their characteristics and labor market histories, in particular for men.

We extend the work of Flores and Mitnik (2013) by applying the same strategy to the treated individuals in each program, as well as to the controls, and interpret any difference in treatment effects between programs as those *not* explained by differences in the characteristics of the program beneficiaries. This interpretation is made possible by imposing an *overlap* or *common support* condition that guarantees the *simultaneous* comparability of the individuals across programs. However, the imposition of this condition comes at a cost. Depending on the analysis, we lose up to thirty and forty percent of the men and women, respectively, in some programs when imposing overlap. For men satisfying common support (i.e. comparable), who have worse initial conditions in terms of education and labor history than non-comparable individuals, the positive effects on formality disappear for some programs, once we estimate treatment effects. For women, as the methods do not work well in equalizing outcomes across programs for controls, we do not deem treatment effects comparisons as credible. Our results highlight the importance of treatment effect heterogeneity, which may have implications for targeting and program design. They also make explicit the limits to the external validity of each of the experiments, and how difficult is the interpretation of the results from meta-analysis studies. By stressing for which types of individuals it *is* possible to make comparisons across programs, our study points towards a more nuanced comparison of impact evaluation studies of youth training programs.

This paper is organized as follows. Section 2 describes the programs compared and summarizes the results from their evaluations. Section 3 describes the methods used, while Section 4 presents the results. Section 5 concludes.

2 Programs compared

Traditionally, vocational training in Latin America and the Caribbean is offered by the state through National Institutes of Vocational Training.[3] For example, Colombia has the SENA (*Servicio Nacional de Aprendizaje*) and the Dominican Republic has the INFOTEP (*Instituto de Formación Técnico Profesional*). These institutions are largely responsible for the provision of technical training whose contents are determined centrally. In the early nineties, a different model emerged in Chile with the *Chile Joven* program.[4] This program was developed in 1991 and its objective was to improve the labor market insertion of vulnerable youth through the provision of short-term semi-qualified training in specific occupations demanded by the productive sector. Its three main features were: (i) the separation of the financing from the provision of training, i.e. the government selected training courses competitively, through a process where private or public training institutions could participate; (ii) the nature of the training was demand driven, i.e. responded to the specific needs of the productive sector[5]; and (iii) it offered both classroom training and on-the-job training: an initial classroom-training course was followed by an internship in a formal firm (Ibarrarán and Rosas-Shady, 2009).

[3] In Peru and Brazil, these institutions are sectorial. For example, for the manufacturing sector, Peru has the SENATI and Brazil has the SENAI.

[4] An alternative model was developed in México in 1984: the program *Probecat*. This program did not offer classroom training courses but only on-the-job training that lasted three months. *Probecat* was replicated in Honduras and in El Salvador. However, there are no robust impact evaluations of these programs. For more information of these programs see Gonzalez-Velosa et al (2012).

[5] The program required that training centers submit course proposals together with a "letter of intent" in which private firms expressed their willingness to receive interns and stated that the course contents responded to their skill requirements.

Table 1. Main characteristics of youth training programs in Colombia, the Dominican Republic, and Peru

	Programs		
	***Juventud y Empleo* (Dominican Republic)**	***Projoven* (Peru)**	***Jóvenes en Acción* (Colombia)**
Execution period	2001 to date	1996-2010	2002-2005
Classroom Technical Training	Yes (150 hours, approx. 1.5 months)	Yes (approx. 3 months)	Yes (360 hours, approx. 3 months)
Classroom "Soft" Skills Training	Yes (75 hours, approx. 1.5 months)	No	Yes (part of the 3-month course)
On-the-job training (internship)	Yes (240 hours, approx. 2 months)	Yes (approx. 3 months)	Yes (approx. 3 months)
Firms must have vacancies	Partially	No	No
Firms must hire % of the trainees	No	No	No
Eligible population	Youngsters aged 16 to 29 from lower socioeconomic strata who have not completed secondary education and who are not studying	Youngsters aged 16 to 24 from lower socioeconomic strata without higher education or university studies	Unemployed youngsters aged 18 to 25 from lower socioeconomic strata
The firms must finance the internship partially or totally	No	Yes (the firm pays a wage)	No
Average cost of the program per beneficiary (US$)	700	420	750

Source: Authors' elaboration based on Gonzalez-Velosa, Ripani and Rosas-Shady (2012).

The model introduced by the *Chile Joven* program was replicated, with the support of the Inter-American Development Bank (IDB) and the International Labour Organization (ILO), in many LAC countries: in

Venezuela (1993), Argentina (1994), Paraguay (1994), Peru (1996), the Dominican Republic (1999), Colombia (2000), Panama (2002) and Haiti (2005).[6]

In this paper, we use the information from the experimental (RCT) impact evaluations of three of these training programs: those implemented in Dominican Republic (Youth and Employment, or *Juventud y Empleo* in Spanish), Colombia (Youth in Action, or *Jóvenes en Acción* in Spanish), and Peru (*Projoven*). Even though all three programs were based on the *Chile Joven* model and were oriented towards vulnerable, young beneficiaries, they have important differences, explained in detail in the next subsections. A summary of these differences is presented in Table 1.

The three programs have differences regarding the specific groups targeted, the rules associated to participation, in particular regarding education requirements, and the length and types of services provided. Also they were implemented at a different period of time. The results of the impact evaluations of each program were also different, and are summarized in Table 2.

[6] The years in parenthesis are the years in which the design of each program started. The actual training courses started later than the design year, i.e in the Dominican Republic the courses started in 2001, while the design of the program started in 1999.

Table 2: Main results of the impact evaluations of youth training programs in Colombia, the Dominican Republic and Peru[7]

Main impacts	Programs		
	Juventud y Empleo **(Dominican Republic)**	***Projoven*** **(Peru)**	***Jóvenes en Acción*** **(Colombia)**
Authors	Ibarraran et al. (2014)	Diaz and Rosas–Shady (2016)	Attanasio, Kugler and Meghir (2011)
Program components	Training in technical and "soft" skills + internship	Training in technical skills + internship	Training in technical and "soft" skills + internship
Probability of employment	Not significant	Not significant	Women: 5.3 pp Men: not significant
Weekly hours worked	Not significant	Not significant	Women: 3 hours Men: not significant
Labor income	Monthly earnings Employees: 7%	Not significant	Women: 22% Men: not significant
Formal Employment	Employment with health coverage: Men: 17 %	Employment with health insurance: All: 3.8 pp Men: 6.8 pp	Employment with health insurance, pension or family benefits: Men: 5 pp Women: 7 pp
Evaluation date	18 to 24 months after graduation from the program	3 years after graduation from the program	13 to 15 months after the program

Source: Authors' elaboration based on Gonzalez-Velosa, Ripani and Rosas-Shady (2012).
Notes: Impacts estimates for *Projoven* correspond to those from the follow-up survey. The evaluation also estimates effects using administrative data.

[7] Impact estimates included in this table only include short term impacts for the case of the Dominican Republic and Colombia. Long-term impact estimates are available for the Dominican Republic's program (Ibarraran et al, 2015) and for Colombia (Attanasio et al, 2015; Kugler et al, 2015). The reason for not including the long-term impacts in this summary table is that this paper is analyzing short-term data for these two countries. From these two countries, long-term data from household surveys is only available for the Dominican Republic, since Colombia's long-term studies are based on administrative data that only includes people that contributes to social security (formal workers).

2.1 The *Juventud y Empleo* program in the Dominican Republic

The *Juventud y Empleo* (JyE) program targets urban youth 16 to 29 years old, unemployed or underemployed, with less than complete secondary school, who live in poor neighborhoods. The program started in 2001 and it runs at the national level. The Ministry of Labor, in collaboration with *INFOTEP*, the National Training Institution in the Dominican Republic, outsources training services from private training institutions registered and approved by *INFOTEP*, and the contracts are awarded on a competitive basis. Eligible young people register in the private training institutions, and once the training center completes its available spots, it sends the list of eligible youngsters to the Ministry of Labor. At the time of the evaluation, the program combined 75 hours of socioemotional skills, 150 hours of technical skills (in courses like administrative assistant, hair stylist, car repair-person, etc.) and 240 hours of internship in a private firm. The average per-person cost of the program is about US$700, of which US$200 are transfers to participants as stipend for transportation and meals.

The data used in this paper is from the short-term[8] follow-up of a cohort (control and treatment groups) interviewed in 2008 (baseline) and at the end of 2010 (18 to 24 months after graduation of the treatment group). This data was analyzed by Ibarraran, Ripani, Taboada, Villa and Garcia (2014) and the evaluation shows that the program has a positive impact on job formality for men of about 17 percent, and causes a seven percent increase in monthly earnings among those employed. The evaluation did not find impacts on employment rates.[9]

2.2 The *Projoven* program in Peru

Projoven was created by the Ministry of Labor in 1996, and operated until 2010.[10] Its main objective was to facilitate access into the formal labor

[8] The results of a long-term follow of the Youth and Employment program is presented in Ibarraran et al (2015). The cohort with baseline data in 2008 was followed up to 2014, six years after the treatment.

[9] Ibarraran et al. (2014) also analyze the impact of the program on teenage pregnancy, socioemotional skills and expectations. These non-labor outcomes are not analyzed in this paper.

[10] During the period 1996-2004, the program was financed with resources from the Public Treasury. During the 2005-2010 period its implementation was financed with resources received through a loan from the Inter-American Development Bank

market for urban young people with limited resources. It operated in the most important cities of Peru, offering in-classroom training for three months, later to be complemented with an internship for three additional months. The program did not include a module of socioemotional skills. *Projoven* hired private or public training institutions that were responsible for the design and the provision of training for the program's beneficiaries. Training courses were supposed to be designed in coordination with firms in which beneficiaries would later do their internships. Contrary to the programs in the Dominican Republic and Colombia, eligible youths were selected by the program (not by the training institutions). After they chose their preferred course (from a list of all the courses offered) the program sent the potential trainees to the corresponding training institutions to undergo a selection process. This selection was usually based on vocational and basic skills tests as well as on interviews (training institutions used their own evaluation criteria), and it determined which eligible candidates would become *Projoven*'s beneficiaries. This process took place until enough young people were assessed and declared suitable to cover the number of available openings for each course. Another difference of *Projoven* is that during the three-month internship stage, beneficiaries received a stipend lower than the minimum wage as well as health insurance coverage. Both the stipend and the health insurance were paid by the firms, since employment of interns had to be made through the Peruvian legal framework of vocational training agreements (a special type of contract between employers and interns). This requirement was supposed to reinforce the demand driven mechanism. Finally, the cost of the program per beneficiary was lower: US$ 420 (including operating costs and the stipend for the beneficiaries).

The data used in this paper correspond to the follow-up household survey implemented three years after beneficiaries completed the training, analyzed by Diaz and Rosas-Shady (2016).[11] The evaluation found a positive impact of *Projoven* only on formal employment. The probability of

(IDB). In total, during almost 15 years of service, the program trained approximately 73,000 youths. In 2010, after a change of government and the end of financing from the IDB, the Ministry of Labor modified the program's design and name, becoming the program *Jóvenes a la Obra.*

[11] The evaluation evaluated impacts using a household survey, and complemented the estimations using administrative data (data from social security records for formal workers).

having a job with health insurance and the probability of having a pension increased in 3.8 and 3.3 percentage points, respectively, for the treatment group when compared to the control group. These effects are of considerable magnitude, since they represent an increase greater than 20% in the probability of having health insurance or a pension for the treatment group. Also, their magnitude is almost twice for male youths and those between 14 and 18 years of age.

2.3 The *Jóvenes en Acción* program in Colombia

Jóvenes en Acción (JA) was a labor training program for urban young people between the ages of 18 and 25 in the two lowest socioeconomic strata of the population. The program started in the seven largest cities[12] of Colombia in 2001 and ended in 2005. As JyE and *Projoven*, it offered three months of classroom training and three months of on-the-job training. The classroom training courses included a module of soft skills (named "*Proyecto de Vida*" or life project). Private training institutions chose the courses to be taught as part of the program and received applications from youngsters. The average per-person cost of the program was about US$750, which included transfers to participants as stipend.

The data used in this paper is the short-term[13] follow-up of a cohort (control and treatment groups) that were interviewed in January 2005 (baseline) and between August 2006 and October 2006 (13 to 15 months after graduation follow-up). According to Attanasio, Kugler, and Meghir (2011), the program raised earnings and employment for women. Women offered training earned 19.6 percent more and had a 0.068 higher probability of paid employment than those not offered training, mainly in formal-sector jobs.

[12] Barranquilla, Bogotá, Bucaramanga, Cali, Cartagena, Manizales, and Medellin.

[13] The results of a long-term follow of the *Jóvenes en Acción* program is presented in Attanasio et al. (2015). The cohort with baseline data in 2005 was followed using social and labor market administrative records collected between July 2008 and June 2014, three to six years after the treatment. Kugler et al (2015) analyze long-term impacts of the same program on both educational and labor market outcomes. For the latter, the authors use social security records collected between 2008 and 2013.

3 Methodology

We follow the methodology proposed by Flores and Mitnik (2013), FM hereafter, who rely on a *selection on observables* or *unconfoundedness* assumption, and use multiple-treatment non-experimental estimators to eliminate the differences in observable characteristics between individuals in different locations. The methodology proposed by FM essentially implies three steps. First, we estimate the generalized propensity score (GPS) (see Imbens, 2000; Hirano and Imbens, 2004). This GPS provides the probability of an individual *i* to be in location L=*k*, and we estimate it with a multinomial logit model (which FM show works well in this type of setting). This can be done for two (in which case the GPS becomes a standard propensity score) or more locations. In the latter case it is necessary to keep in mind that while the GPS is the probability for individual *i* who belongs to location *L* to be in location *k*, $\Pr(L_i = k|L_i = k)$, we also obtain the probability of any individual *not* in location L to be in location *k*, $\Pr(L_i = k|L_i \neq k)$. This latter probability is relevant for the second step, while the first probability is used in the third step.

Second, in several instances we impose that the individuals in each location are *simultaneously* comparable across all locations. For this we impose a *common support condition*, or which is the same, we find the individuals for which there is *overlap* in observed characteristics. We impose this condition with the procedure suggested by FM. For each individual we take his or her probability of being in each location *k* (both for those in location *k* and for those *not* in location *k*) and mark the individuals with a probability of being in the location higher or equal to $\rho = \max\{\mathrm{q}(\Pr(L_i = k|L_i = k)), q(\Pr(L_i = k|L_i \neq k)\}$, where $q(\cdot)$ denotes the q^{th} quantile over the distribution of the GPS for the particular group indicated in parenthesis. Those individuals satisfy the overlap condition in location *k.* Then we mark as satisfying the overlap condition overall those individuals that satisfy the overlap condition in *every* location *k*. This guarantees that the individuals satisfying the overlap condition are comparable in *all* locations simultaneously.[14]

[14] We use the quantile q=0.0075 (i.e. 0.75%). In addition we consider as not satisfying the overlap condition those observations for which the GPS is so low as to imply a weight (for estimators based on weighting by the inverse of the GPS) of above 3% of the sample. This eliminates a very small number of observations, four for men, and five for women.

Third, we estimate alternative estimators of the mean outcome associated to each program (location) for control and treated individuals. We call *raw* the estimator that just computes the outcome means with no adjustments of any type (we estimate them both for the full sample, and for the sample of individuals that satisfy the overlap condition). The other estimators, following FM, all use the fact that in this setting the mean outcomes associated to each location can be obtained through a *partial mean*, which is an average of a regression function over some of its regressors while holding others fixed (see FM for details). We denote as *linear regression* the estimator where we first estimate with a standard linear regression the expectation of the outcome conditional on the treatment group, location and covariates. Second, this regression function is averaged over the covariates, keeping fixed the treatment group and location, for all the individuals, not only those in the group and location. This estimator does not impose the overlap condition. We denote as *inverse probability weighting* (IPW) the estimator where a similar procedure is followed, but imposing the overlap condition and weighting each individual by the inverse of its own GPS, $\Pr(L_i = k|L_i = k)$. The regressions estimated have the following form:

$$Y_{ik}^t = \sum_t \sum_{k=1}^K \beta_k^t L_{ik}^t + \gamma X_i + \eta_c + \varepsilon_i \qquad (1)$$

where Y_{ik}^t is an outcome variable at follow-up; L_{ik}^t is a dummy equal to 1 if the individual *i* is in treatment group *t* (*t*=C,T for control and treated individuals respectively) in location *k* and 0 otherwise; X_i is a matrix of individual characteristics at baseline; η_c represents a training course fixed effect, to control for the fact that individuals or the programs made the course selections; and ε_i is an error term. The coefficients β_k^t are key because they allow, as the regressions are run without a constant, obtaining an estimation of $\mathrm{E}[Y_{ik}^t|L_i = k, t = C]$, the average (adjusted) mean of the outcome for control individuals in each location and similarly for treated individuals, $\mathrm{E}[Y_{ik}^t|L_i = k, t = T]$. The difference between the treated and control means in a location, represents the *average treatment effect* of the program in that location. And, if we are successful in controlling for the differences in observed characteristics among control individuals in different programs, then we expect that for the control groups these adjusted means should be equal, after weighting by the GPS (and imposing the overlap condition). We test this latter hypothesis by estimating a version of (1) only for control individuals.

An issue that arises in trying to estimate (1) is that the fixed effects η_c are perfectly collinear with the β_k^t coefficients, because the course fixed effects are country-specific. Thus, we actually estimate (1) in two parts. First, for *each country*, we run a regression of the outcome variables against only the course fixed effects, pooling in these regressions control and treated individuals. We take the residuals from those regressions and create a new outcome variable $\tilde{Y}_i$ that is equal to the residuals from each of these regressions plus the mean of the outcome variable for the individuals in each country. This new variable $\tilde{Y}_i$ has the same grand mean, and country-specific mean, of the original variable, but with the course fixed-effects subtracted. We then estimate a regression similar to (1):

$$\widetilde{Y_{ik}^t} = \sum_t \sum_{k=1}^{K} \beta_k^t L_{ik}^t + \gamma X_i + \varepsilon_i \qquad (2)$$

where, as before, the results are used to obtain the adjusted mean outcomes for individuals in a treatment group and location (country).

We follow FM in calculating two other partial mean estimators of the mean outcomes in a location for treated and control individuals. In the *parametric partial mean estimator*, after imposing the overlap condition, the conditional means of the outcome are obtained from a regression on a quadratic function of the GPS, not controlling for covariates:

$$\widetilde{Y_{ik}^t} = \sum_t \sum_{k=1}^{K} \beta_k^t L_{ik}^t + \sum_{p=1}^{2} \gamma_p {P_{ik}^t}^p + \sum_t \sum_{k=1}^{K} \sum_{p=1}^{2} \delta_k^t L_{ik}^t \times {P_{ik}^t}^p + \varepsilon_i \qquad (3)$$

where P_{ik}^t denotes the GPS for individual *i* that belongs to group *t* in location *k*. It is important to note that when calculating the mean outcomes of interest, after the regression has been estimated, the means are calculated over all the individuals, using the estimated probability that they belong to the particular location and group being calculated, even if they do not belong to that location and group. The final estimator is a nonparametric version of (3) that employs a local linear regression of the outcome on the GPS to estimate the regression function for each location and group. We denote this estimator as *nonparametric partial mean*.

For all the estimators, to take into account the statistical uncertainty associated to the multi-step procedures, we obtain confidence intervals by bootstrapping all the steps involved in estimation, including the course fixed effects and GPS estimation, when appropriate.

As mentioned above, at first we run all the estimators described only using the control individuals, to test the hypothesis that the estimated mean outcomes in each location are *jointly* equal. Once we obtain with a

particular estimator the mean outcome in each location, we simply calculate a Wald test that the means are jointly equal. As these tests of equality of means can be affected by the size of the standard errors, we follow FM and calculate the *root mean squared distance* (*rmsd*) as a measure of distance between each of the estimated means. Letting $\hat{\theta}_k$ denote the estimated value of $\theta_k = \mathrm{E}[Y_{ik}^t | L_i = k, t = C]$ and $\hat{\mu} = k^{-1} \sum_{k=1}^{K} \hat{\theta}_k$, the *rmsd* is defined as $rmsd = \frac{1}{|\hat{\mu}|}\sqrt{k^{-1} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\mu}\right)^2}$. As discussed in FM, the lower the value of the *rmsd*, the closer are the control individuals in each location to each other in terms of outcome means. The advantage of the *rmsd* is that it is not dependent on standard errors of the coefficients, but as is shown in FM, even under complete randomization of individuals to locations, its value is not expected to be zero, due to sampling variability. As in FM, we obtain a *benchmark* for this sampling variability by running simulations where we randomly assign individuals to locations (keeping location proportions fixed), and calculate the *rmsd* associated to these simulations.

4 Results

In this section, we present the results from following the methodology explained in Section 3 in comparing the *Jóvenes en Acción* program in Colombia (CO hereafter), the *Juventud y Empleo* program in Dominican Republic (DR hereafter) and the *Projoven* program in Peru (PE hereafter). Table 3 shows the means and standard deviations for the outcomes and for the covariates used in the GPS estimation and the estimators based on covariates, separately for men and women. The table also includes covariates not used in the GPS estimation, for information purposes. These covariates are not part of the GPS specification because either they are not available in the version of the dataset we have for Colombia, or because they cause the GPS to perfectly predict a location. The table pools control and treated individuals, because within each country the differences in means between those groups are not statistically significant (as expected given random assignment within each country).

Table 3. Descriptive statistics CO, DR, PE - treated & control individuals

Variable	CO		DR		PE	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
			Men			
Follow-up outcomes						
Employment	0.83	0.38	0.77	0.42	0.79	0.40
Formal employment	0.38	0.49	0.26	0.44	0.20	0.40
Baseline characteristics included in GPS						
Age	21.0	2.0	20.8	3.1	19.0	2.3
Married	0.10	0.31	0.08	0.28	0.03	0.16
Years of Schooling	10.2	1.7	7.4	3.2	10.5	1.4
Employment 1 qtr. before program	0.34	0.47	0.19	0.39	0.26	0.44
Income/MW 1 qtr. before program	0.30	0.46	0.26	0.79	0.37	1.63
Baseline characteristics not included in GPS						
High school complete	0.75	0.43	0.04	0.19	0.81	0.40
Employment at baseline	0.58	0.49	0.06	0.24	0.59	0.49
Employment 2 qtrs. before program	-	-	0.21	0.41	0.27	0.44
Formal employment 1 qtr. before program	0.10	0.30	0.05	0.22	0.02	0.15
Formal employment 2 qtrs. before program	-	-	0.06	0.24	0.02	0.15
Income/MW 2 qtrs. before program	-	-	0.29	0.80	0.42	1.77
Number of observations	671		509		219	
			Women			
Follow-up outcomes						
Employment	0.67	0.47	0.53	0.50	0.56	0.50
Formal Employment	0.23	0.42	0.13	0.34	0.14	0.35
Baseline characteristics included in GPS						
Age	21.3	2.0	21.9	3.5	19.2	2.3
Married	0.27	0.44	0.33	0.47	0.12	0.33
Years of Schooling	10.1	1.7	7.5	3.2	10.7	1.1
Employment 1 qtr. before program	0.28	0.45	0.12	0.33	0.20	0.40
Income/MW 1 qtr. before program	0.19	0.36	0.10	0.38	0.16	0.85
Baseline characteristics not included in GPS						
High school complete	0.73	0.45	0.04	0.21	0.88	0.33
Employment at baseline	0.47	0.50	0.02	0.15	0.39	0.49
Employment 2 qtrs. before program	-	-	0.14	0.35	0.22	0.42
Formal Employment 1 qtr. before program	0.06	0.23	0.03	0.16	0.01	0.09
Formal Employment 2 qtrs. before program	-	-	0.03	0.17	0.01	0.10
Income/MW 2 qtrs. before program	-	-	0.11	0.35	0.21	1.10
Number of observations	860		825		327	

Some patterns that emerge from Table 3 are clear consequences of the differences in the targeted groups by each program. Individuals in PE are younger,[15] less likely to be married, and with higher level of schooling and high school completion. DR individuals show much lower years of schooling, compared to CO and PE, and much lower high school completion rates, compared to PE. These differences are expected, since

[15] The upper limit in terms of age for PE was the lowest of the three studies programs (16-24 for PE, 16-29 for DR, 18-25 for CO).

the program in DR targets youth with, at most, incomplete high school education (high-school drop-outs), while both CO and PE allowed beneficiaries to have completed high school (although without higher education attendance). Employment rates at baseline and before the programs, as well as employment income as function of minimum wages before the program, are lower in DR than in CO and PE. However, formality rates prior to the programs are higher in DR than in PE. According to the eligibility rules, for the three programs the focus was on urban unemployed youth. The high rates of informality in LAC countries make it hard to know if a person is unemployed, since he or she can work informally and not show up in social security records as employed. Thus, it is not surprising that the baseline surveys show that many individuals were working at baseline.

We estimate the GPS with a multinomial logit model using the variables indicated in Table 3, separately for men and women, and using only control individuals at first. Unfortunately, the variables included in the GPS are the only ones available in all three datasets that work in estimation, which makes the models very limited. We use the estimated GPS to identify those individuals satisfying the overlap condition. The bottom panel of Table 4 presents the number of control individuals that are outside the overlap region, and their percentages by country and gender. Almost 30% of men in DR and around 20% in CO and 23% in PE do not satisfy the condition. For women the shares are higher indicating that the characteristics of the women in all three programs are more different across countries. Almost 40% of women in PE, 26% in DR and 23% in CO do not satisfy the overlap condition. Overall, almost 24% of men and 27% of women appear as not comparable. The top panels of the table make this clear. The first three columns show the average for each variable by country and gender. The next three columns show the average for the individuals satisfying the overlap condition, while the next three columns, under the heading IPW (inverse probability weighting), weight the individuals satisfying the overlap condition by the inverse of the GPS of each individual, to balance covariates. It is an indication of how well the GPS works in eliminating differences between the individuals. The next two columns show the p-value of a joint test of equality of means between the three countries in the raw and IPW case, while the last two columns calculate the *rmsd* as a measure of distance between the averages of the variables in the three countries, not dependent on standard errors.

Table 4. Balancing of baseline characteristics for control individuals CO, DR, PE

Variable	Variable Means									P-value Equality of Means Test		Root Mean Square Distance	
	Raw			Raw w/Ovlp			IPW						
	CO	DR	PE	CO	DR	PE	CO	DR	PE	Raw	IPW	Raw	IPW
							Men						
Baseline characteristics included in GPS													
Age	21.1	20.8	18.8	20.8	20.0	18.9	20.4	20.0	20.3	0.00	0.02	0.05	0.01
Married	0.12	0.08	0.03	0.02	0.06	0.01	0.02	0.07	0.01	0.00	0.01	0.51	0.78
Years of Schooling	10.1	7.2	10.4	10.3	8.8	10.3	9.5	9.2	9.6	0.00	0.05	0.16	0.02
Employment 1 qtr. before program	0.30	0.22	0.23	0.26	0.22	0.14	0.24	0.24	0.17	0.00	0.39	0.15	0.14
Income/MW 1 qtr. before program	0.28	0.31	0.42	0.26	0.31	0.15	0.24	0.29	0.17	0.46	0.09	0.18	0.21
Baseline characteristics not included in GPS													
High school complete	0.74	0.04	0.77	0.74	0.05	0.73	0.61	0.09	0.57	0.00	0.00	0.65	0.56
Employment at baseline	0.54	0.07	0.58	0.49	0.04	0.55	0.48	0.05	0.60	0.00	0.00	0.59	0.63
Employment 2 qtrs. before program	-	0.24	0.23	-	0.25	0.18	-	0.26	0.21	0.00	0.34	0.02	0.11
Formal employment 1 qtr. before program	0.10	0.06	0.03	0.08	0.05	0.02	0.07	0.05	0.03	0.00	0.04	0.44	0.38
Formal employment 2 qtrs. before program	-	0.07	0.04	-	0.07	0.02	-	0.07	0.03	0.00	0.03	0.32	0.46
Income/MW 2 qtrs. before program	-	0.35	0.49	-	0.35	0.19	-	0.34	0.18	0.02	0.00	0.16	0.32
							Women						
Baseline characteristics included in GPS													
Age	21.4	21.8	19.4	21.4	21.2	20.1	21.1	21.0	20.9	0.00	0.57	0.05	0.00
Married	0.27	0.33	0.14	0.28	0.30	0.19	0.28	0.28	0.25	0.00	0.77	0.31	0.06
Years of Schooling	10.0	7.4	10.7	10.2	9.0	10.8	9.8	9.3	10.3	0.00	0.00	0.15	0.04
Employment 1 qtr. before program	0.27	0.13	0.21	0.13	0.12	0.06	0.10	0.18	0.10	0.00	0.00	0.29	0.28
Income/MW 1 qtr. before program	0.19	0.10	0.16	0.10	0.10	0.10	0.08	0.14	0.11	0.00	0.02	0.23	0.21
Baseline characteristics not included in GPS													
High school complete	0.71	0.04	0.90	0.70	0.05	0.88	0.63	0.09	0.71	0.00	0.00	0.67	0.57
Employment at baseline	0.46	0.03	0.39	0.35	0.02	0.34	0.33	0.02	0.34	0.00	0.00	0.65	0.64
Employment 2 qtrs. before program	-	0.14	0.20	-	0.14	0.10	-	0.19	0.12	0.02	0.00	0.17	0.21
Formal employment 1 qtr. before program	0.05	0.03	0.00	0.03	0.03	0.00	0.02	0.04	0.00	0.00	0.00	0.71	0.83
Formal employment 2 qtrs. before program	-	0.03	0.00	-	0.03	0.00	-	0.04	0.00	0.00	0.00	0.82	1.00
Income/MW 2 qtrs. before program	-	0.11	0.19	-	0.11	0.17	-	0.15	0.14	0.25	0.94	0.27	0.02

Overlap	Men			Women		
	CO	DR	PE	CO	DR	PE
Total number of control observations	671	509	219	860	825	327
Number observations outside overlap region	132	152	51	196	213	130
Percentage observations outside overlap region	19.7%	29.9%	23.3%	22.8%	25.8%	39.8%

Note: Raw refers to the data in its original form, Raw w/Ovlp refers to the data after observations not satisfying the overlap condition have been dropped, IPW refers to data weighted by the inverse of the GPS associated to each observation

Table 4 shows how imposing the overlap condition is more successful in eliminating differences for some variables than others. Of the five variables included in the GPS estimation, the balancing in age and years of schooling is much improved for both males and females (even in the case where the equality of means tests is rejected, the *rmsd*, which measures the distances between the means, decreases significantly for both variables). For percent married the results are more mixed, with success in reducing differences only for women. For employment and income in the quarter prior to the program there are no significant reductions in differences for either group.

For the variables *not* included in the GPS estimation balancing does not improve much after imposing overlap, except for high school completion which sees some improvement, but with a still very large difference for DR. Still these variables provide useful information, even just in the comparison of the columns raw and raw with overlap, because the difference between the two sets of columns indicate the characteristics of those individuals that are *not* comparable across countries. For example, individuals with more labor experience, higher likelihood of formal employment and higher income one quarter before the program are outside the overlap region in both CO and PE, and similarly two quarters before the program for PE. This makes sense as the characteristics of the individuals in DR indicate that they start from a more disadvantaged position, compared to those in CO and PE, both in terms of education and labor experience. Thus, any comparison between the three countries that relies on the overlap condition will have to apply to relatively more disadvantaged individuals.

Table 5. Outcome means at follow-up for control individuals CO, DR, PE

Estimator	Employment at follow-up					Formal employment at follow-up				
	Outcome means			P-value	RMSD	Outcome means			P-value	RMSD
	CO	DR	PE			CO	DR	PE		
					Men					
Raw mean (full sample)	0.83	0.79	0.77	0.06	0.03	0.34	0.24	0.18	0.00	0.27
	[0.81, 0.86]	[0.76, 0.82]	[0.72, 0.82]		[0.01, 0.06]	[0.31, 0.37]	[0.21, 0.27]	[0.13, 0.22]		[0.19, 0.37]
Raw mean (imposing overlap)	0.82	0.78	0.76	0.16	0.03	0.32	0.25	0.14	0.00	0.31
	[0.79, 0.85]	[0.74, 0.82]	[0.70, 0.82]		[0.01, 0.07]	[0.29, 0.36]	[0.21, 0.29]	[0.10, 0.20]		[0.20, 0.43]
Linear regression (full sample)	0.79	0.83	0.81	0.00	0.02	0.24	0.27	0.26	0.01	0.04
	[0.78, 0.82]	[0.80, 0.83]	[0.81, 0.85]		[0.01, 0.02]	[0.25, 0.29]	[0.27, 0.32]	[0.26, 0.31]		[0.02, 0.07]
Parametric partial mean (imposing overlap)	0.78	0.82	0.81	0.07	0.02	0.22	0.24	0.24	0.71	0.03
	[0.75, 0.81]	[0.78, 0.86]	[0.79, 0.86]		[0.01, 0.05]	[0.23, 0.28]	[0.24, 0.32]	[0.22, 0.29]		[0.02, 0.12]
Nonparametric partial mean (imposing overlap)	0.77	0.82	0.81	0.54	0.03	0.23	0.23	0.24	0.97	0.03
	[0.73, 0.83]	[0.77, 0.89]	[0.78, 0.86]		[0.01, 0.07]	[0.22, 0.30]	[0.22, 0.33]	[0.20, 0.32]		[0.02, 0.17]
IPW regression (imposing overlap)	0.78	0.82	0.83	0.05	0.03	0.23	0.27	0.24	0.10	0.07
	[0.76, 0.81]	[0.79, 0.84]	[0.79, 0.87]		[0.01, 0.04]	[0.23, 0.28]	[0.26, 0.32]	[0.21, 0.30]		[0.02, 0.14]
Benchmark	0.81	0.81	0.81	0.66	0.02	0.28	0.28	0.28	0.73	0.06
	[0.79, 0.83]	[0.79, 0.83]	[0.77, 0.85]		[0.00, 0.04]	[0.26, 0.30]	[0.25, 0.31]	[0.23, 0.32]		[0.02, 0.12]
					Women					
Raw mean (full sample)	0.64	0.52	0.55	0.00	0.09	0.20	0.13	0.13	0.00	0.19
	[0.61, 0.66]	[0.49, 0.55]	[0.51, 0.60]		[0.06, 0.12]	[0.17, 0.22]	[0.11, 0.15]	[0.10, 0.17]		[0.11, 0.30]
Raw mean (imposing overlap)	0.63	0.52	0.57	0.00	0.07	0.18	0.14	0.12	0.08	0.17
	[0.60, 0.66]	[0.49, 0.56]	[0.51, 0.62]		[0.05, 0.12]	[0.16, 0.22]	[0.12, 0.17]	[0.08, 0.16]		[0.09, 0.33]
Linear regression (full sample)	0.54	0.59	0.57	0.00	0.03	0.12	0.19	0.12	0.00	0.24
	[0.55, 0.59]	[0.56, 0.60]	[0.55, 0.60]		[0.00, 0.02]	[0.14, 0.17]	[0.16, 0.19]	[0.13, 0.16]		[0.06, 0.13]
Parametric partial mean (imposing overlap)	0.54	0.59	0.56	0.04	0.04	0.11	0.18	0.13	0.00	0.22
	[0.54, 0.59]	[0.53, 0.60]	[0.53, 0.63]		[0.01, 0.06]	[0.13, 0.17]	[0.15, 0.20]	[0.12, 0.20]		[0.04, 0.17]
Nonparametric partial mean (imposing overlap)	0.54	0.58	0.38	0.52	0.18	0.11	0.18	0.13	0.67	0.21
	[0.48, 0.65]	[0.46, 0.67]	[0.45, 0.63]		[0.01, 0.24]	[0.07, 0.19]	[0.13, 0.21]	[0.11, 0.22]		[0.04, 0.51]
IPW regression (imposing overlap)	0.54	0.59	0.54	0.04	0.04	0.11	0.18	0.10	0.00	0.27
	[0.54, 0.59]	[0.54, 0.60]	[0.51, 0.64]		[0.01, 0.07]	[0.13, 0.16]	[0.16, 0.20]	[0.10, 0.19]		[0.05, 0.22]
Benchmark	0.57	0.57	0.57	0.44	0.03	0.16	0.16	0.16	0.12	0.07
	[0.55, 0.60]	[0.55, 0.60]	[0.53, 0.61]		[0.01, 0.05]	[0.15, 0.18]	[0.15, 0.18]	[0.13, 0.19]		[0.02, 0.14]

Notes: Numbers between brackets are confidence intervals at the 90% confidence level, based on 1,000 bootstrap replications for estimators other than the benchmark, and the 5th and 95th percentile of the simulations for the benchmark estimator. P-value refers to the p-value associated to a Wald test of equality of outcome means between the three programs. RMSD refers to the root mean square distance of the outcome means for each program.

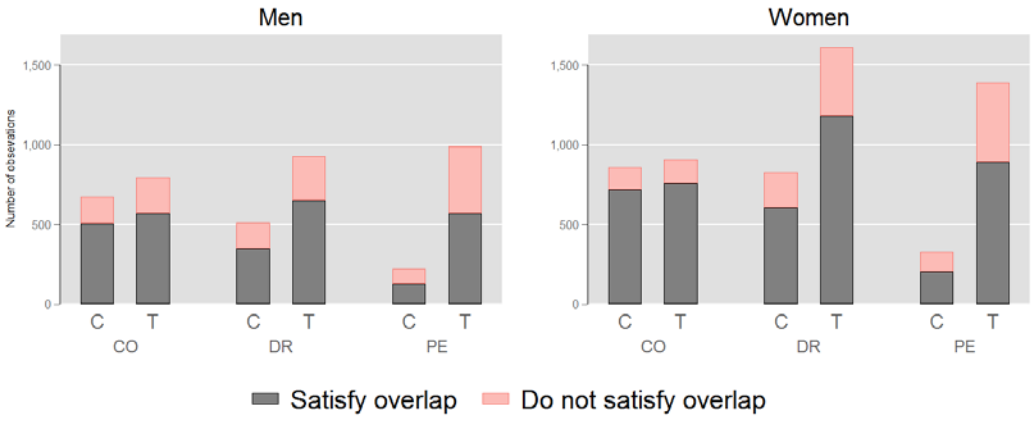
Table 5 presents the estimators discussed in Section 3, using only the outcomes for the control individuals, as a way of checking whether it is possible to eliminate differences in *outcomes* between the countries, for those individuals that are not beneficiaries of the programs. The top panel shows the results for men while the bottom panel shows the results for women. The table presents employment at follow-up and formal employment at follow-up on the left and right side of the table, respectively. The rows show the estimators discussed in Section 3, including the benchmark simulation, and below each, between brackets, its bootstrapped confidence interval at the 90% confidence level (5^{th} and 95^{th} percentiles for the benchmark). For each outcome, the first three columns of Table 5 show the mean outcome in each country, in the fourth column the p-value associated to the Wald test of *joint* equality of these means, and in the fifth column, the *rmsd* associated to the same means. The best way to evaluate the value of a particular *rmsd* is to compare it to the corresponding *benchmark* value, which is a measure of the expected *rmsd* under random assignment of location.

The table shows that for employment at follow-up for men, the differences in raw means are statistically significant at the 10% level and the *rmsd* has a confidence interval that overlaps the benchmark *rmsd* confidence intervals. While the estimators show *rmsd* values that are an improvement upon the raw *rmsd*, their confidence intervals overlap that of the raw estimator and all the p-values imply in general significant mean differences, suggesting not much improvement upon an initial situation with relatively small differences. For formal employment at follow-up for men, the raw differences between the countries are quite large, with a value of the *rmsd* of 0.27 (compared to a benchmark value of 0.06). In this case, the linear regression is able to reduce the differences between the mean outcomes, as the *rmsd* goes down to 0.04, but with a p-value for the test of equality of means that still is significant at around the 1% level. The partial mean estimators, imposing overlap, improve the *rmsd* further, both compared to the raw and to the linear regression case, and the difference of means appears as not statistically significant. The IPW estimator, however, is not as successful in reducing the *rmsd* and the p-value is exactly at the 10% level. Overall, the results for men suggest that the non-experimental estimators are good enough to eliminate differences across countries. For women, Table 5 shows that while for employment at follow-

up the estimators are able to reduce differences in means; this is not the case for formal employment at follow-up. Indeed, the differences become larger instead of smaller, which signals that comparisons for women may be problematic.

As explained in Section 3, the estimation of treatment effects in each country follow (2) and (3), obtaining mean outcomes for both treatment and control individuals, and calculating treatment effects by the difference in mean outcomes between treated and control individuals within each country. For the estimation of treatment effects we estimate the GPS again, now using both control and treated individuals. Based on this GPS, Figure 1 shows for control (C) and treated (T) men and women the number of observations that satisfy overlap, in each country. The figure makes clear the very small number of observations that satisfy overlap in Peru for controls, and the high proportion of treated individuals dropped from this program. This has consequences for the statistical power of the estimated treatment effects, as is discussed below.

Figure 1
Observations Satisfying Overlap Condition
Youth Training Program in CO, DR & PE



Here is very interesting to understand both the characteristics of the individuals who satisfy the overlap condition, as those are the individuals for which the treatment effects are comparable across countries, and the characteristics of those *not* satisfying the overlap condition. The latter group is of interest because they were part of the target population of the programs, highlighting the differences in focus of each program, even though their design was quite similar, and because, as we discuss below,

the treatment effects seem to be different between the individuals inside and outside the overlap region. Thus, it becomes a very useful exercise to characterize them.

Figure 2 presents for selected characteristics the averages for the individuals inside and outside the overlap region (pooling treated and controls), by country and for men (top panel) and women (bottom panel). The figure makes clear that men who are in the overlap region are less likely to be married, have similar years of schooling as those outside the overlap region, except for DR, where those in the overlap region have more than double the years of schooling compared to those outside. Regarding labor history, the individuals inside the overlap region are less likely to be employed at baseline for CO and have much lower employment and formal employment rates and income, in the quarter prior to the program in CO and PE, with smaller or no differences for DR. For women the patterns are very similar, except for the percent married, which is higher in the overlap region for both CO and PE. The figure makes clear the same message suggested by Table 4: the individuals inside the overlap region are more disadvantaged and have worse initial conditions than those outside the region, which is important to keep in mind when analyzing treatment effects.

As explained above, average outcomes are obtained for each group (treated, controls) and program, running regressions (2) and (3). The average differences inside each country are the estimated treatment effects, with confidence intervals obtained by bootstrapping. Figure 3 shows the treatment effects for the employment at follow-up outcome. It depicts three estimators: Linear regression uses the full sample prior to imposing overlap, while parametric partial mean and IPW regression are estimated after imposing overlap. The circles depict the point estimates, while the vertical lines depict the 90% level confidence intervals. For men (top panel), the treatment effects appear as not statistically significant with the three estimators, and the point estimates relatively stable (i.e. imposing overlap does not affect them much). For women (bottom panel) the results become much smaller and not significant for CO, while they remain not significant and stable for DR and not significant, but higher, for PE. This suggests that the effects for CO are higher for those individuals outside the overlap region, i.e. with better initial conditions, as shown in Figure 2.

Figure 2
Average Selected Baseline Characteristics
Treated & Contol Individuals Inside & Outside Overlap

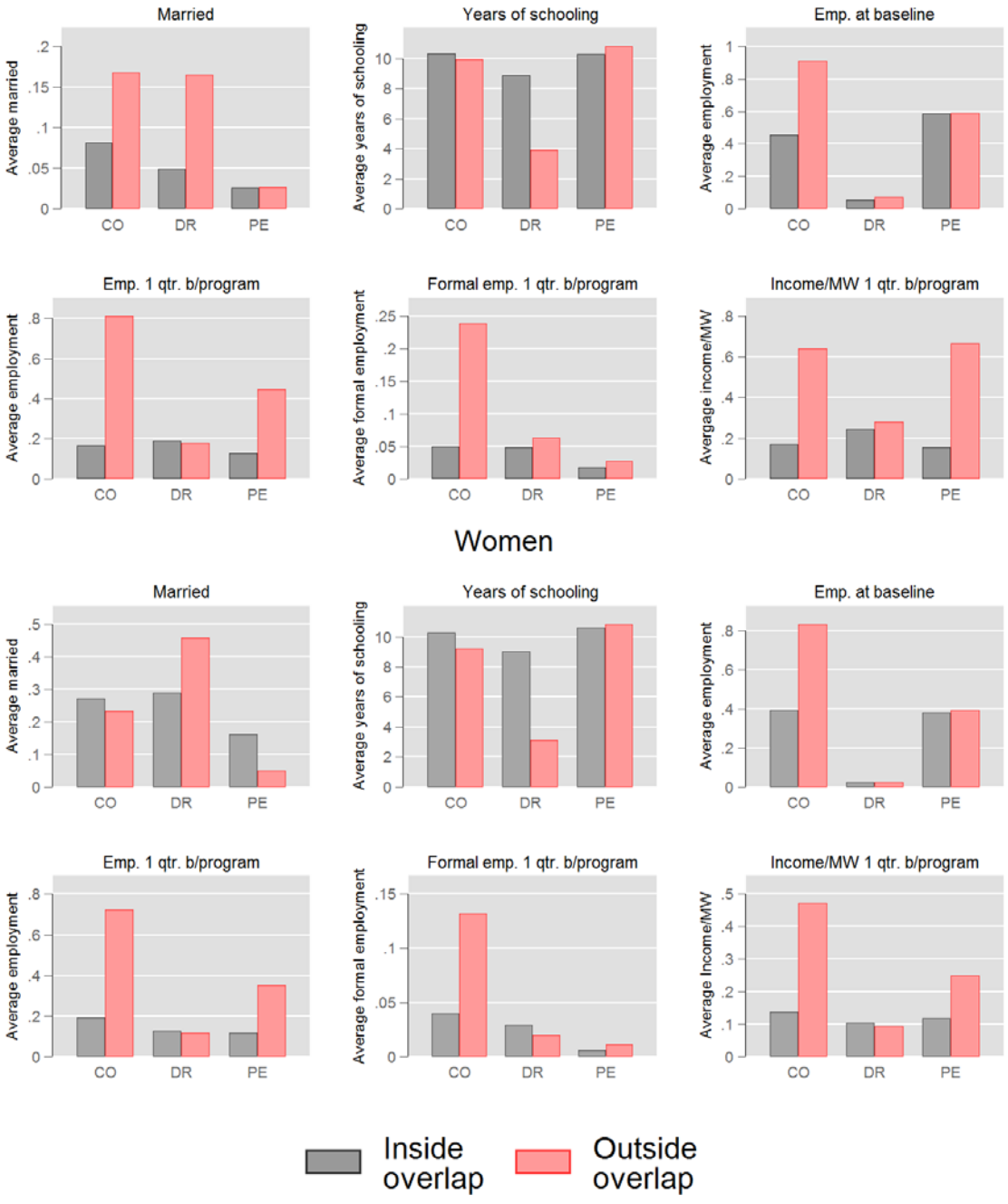


Figure 3
Treatment Effects: Employment at Follow-up

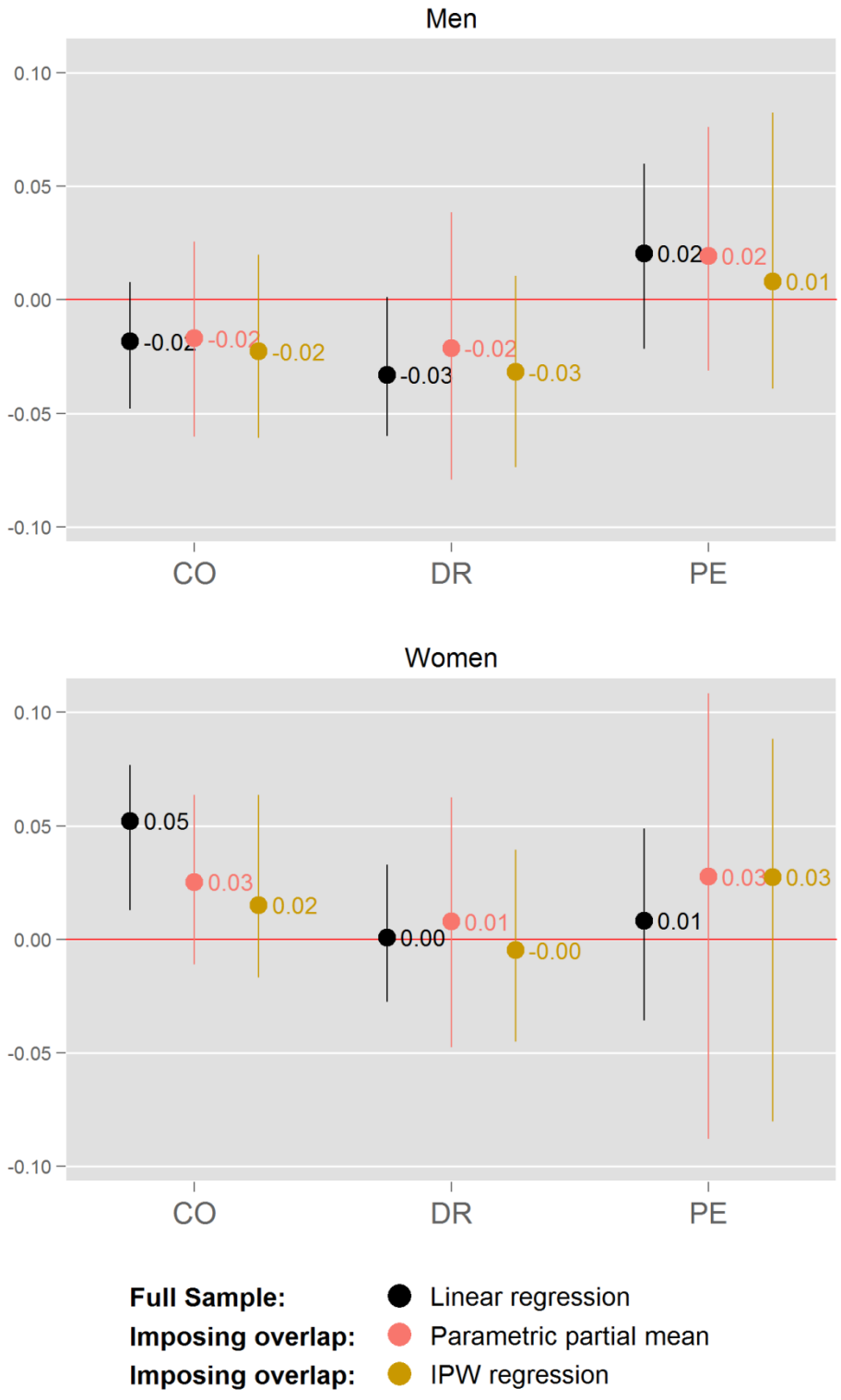


Figure 4
Treatment Effects: Formal Employment at Follow-up

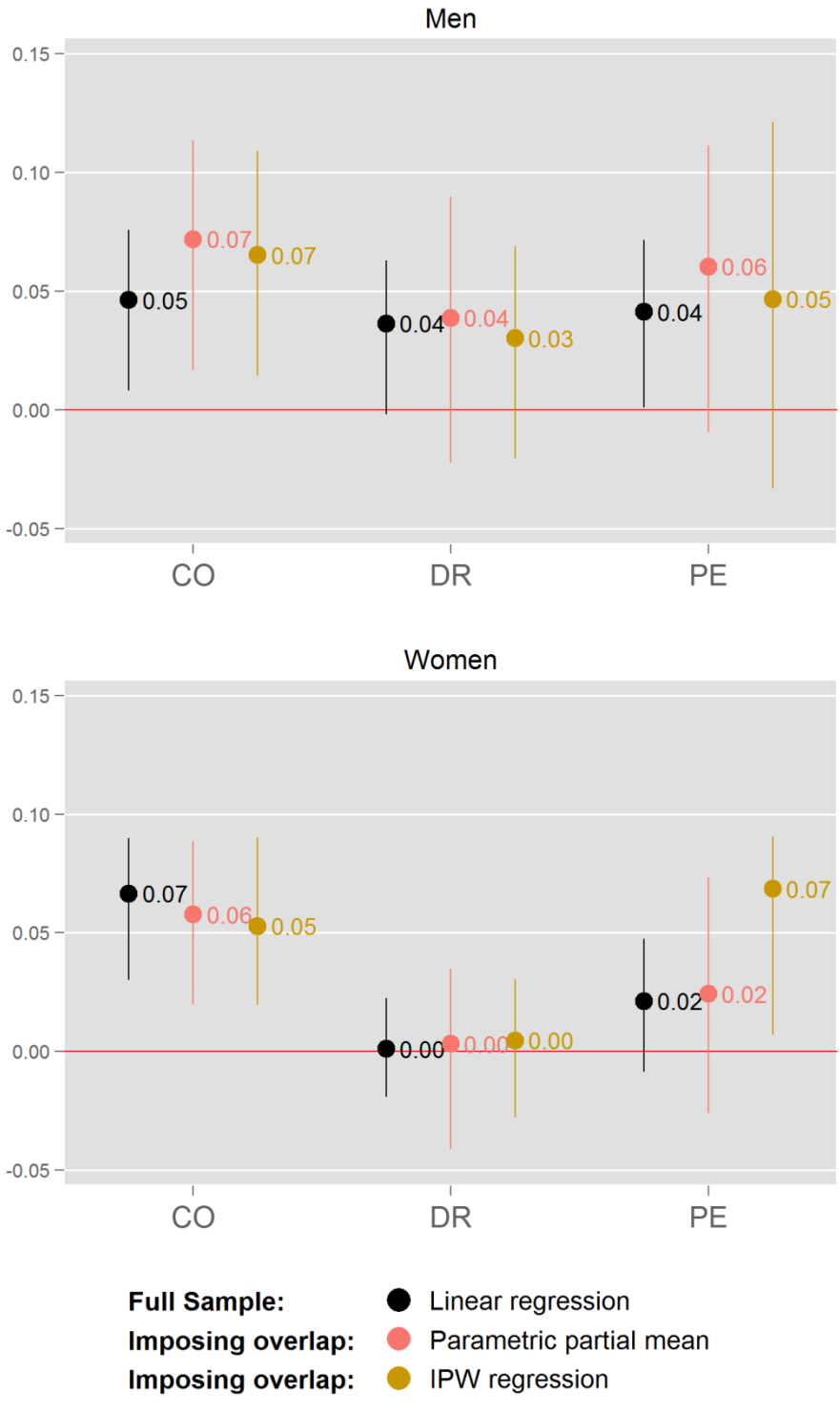


Figure 4 is similar to Figure 3, and presents the results for the outcome formal employment at follow up. This is the outcome where the original evaluations found effects for men and in CO also for women. For men the effects become stronger for CO when imposing overlap, and remain stable in DR and PE. However, in DR and PE the treatment effects become statistically not significant. It is important to consider that for both DR, but particularly for PE, statistical power is reduced when imposing overlap, given the number of observations outside the overlap region (see Figure 1).[16] For women, the results need to be analyzed more carefully given the difficulties encountered in eliminating the differences in this outcome for control individuals (see Table 5). The results suggest some small reduction in effects for CO from the linear estimator to the other estimators, and a large increase in effects for PE, using the IPW estimator. Given the results in Table 5, though, we deem the results for women as problematic, and probably not credible.

Overall, the results highlight that treatment effects within each program can be somewhat heterogeneous, and that they are quite heterogeneous across programs. It is not clear from the analysis is this treatment effect heterogeneity arises from differences in the focus of the programs on different types of individuals[17], or if it arises from different contextual effects (i.e. the macroeconomic conditions) affecting different types of individuals in different ways.

[16] In the case of PE, the treatment effect using the linear regression estimator (0.041) is much lower than the treatment effect obtained in the original evaluation (0.068). This difference is explained by several factors. First, the imposition of course fixed effects by the procedure explained in (3) is not equivalent to imposing fixed effects in a linear regression, and thus affects the results. In addition, the number of covariates used in the original evaluation is much larger than the ones used in this study.

[17] The three programs analyzed target urban unemployed youngsters from lower socioeconomic strata. However, two big differences in the eligibility criteria come from age (DR targets youth aged 16 to 29, PE 16 to 24 and CO 18 to 25), and education, where DR does not accept youth that have completed high school, while CO and PE do accept young people with a high school degree.

5 Conclusions

Several youth training programs in LAC have been rigorously evaluated using RCTs. In the case of Peru's *Projoven*, the Dominican Republic's *Juventud y* Empleo, and Colombia's *Jovenes en Accion*, the results (both short run and long run analyses) show that the programs had, in general, an impact on quality of employment (formality and earnings) and these impacts are sustained over time.

This study explicitly addresses the underlying heterogeneity in the characteristics of the beneficiaries of these three youth training programs in LAC by relying on the individual-level data used in their experimental impact evaluations. We show that by using non-experimental multiple treatment estimators we can control for the underlying heterogeneity in the characteristics of the beneficiaries and identify, and characterize, the individuals satisfying a common support condition, i.e. those who are similar across programs.

Our results show that for some of the programs analyzed, the positive treatment effects on formality disappear once only the comparable individuals are analyzed, making clear the sources of treatment effects heterogeneity and the extent of the limits to the external validity of the experiments used to evaluate the training programs. This does not mean that the programs do not have effects (they do, as it is shown in the individual impact evaluations of the three programs), but rather that comparing similar individuals (those that are part of the common-support group), some of the impacts are smaller, and some are no longer statistically significant. This suggests that some of the effects identified by the original evaluations come from the "non-comparable" groups. It is difficult to know exactly from this study, but it could be that contextual factors (i.e. the economy) have differential effects and/or that the specificities of the eligibility criteria for each country make the programs work better for those that are not comparable with the other countries.

This paper highlights the importance of thinking about treatment effect heterogeneity and its implications for targeting and program design. It also makes clear how difficult is the interpretation of the results from program comparisons that do not control for heterogeneity in the individual characteristics of the beneficiaries, as is the case in meta-analysis studies.

References

Allcott, H. 2012. “Site Selection Bias in Program Evaluation.” NBER Working Paper 18373. September.

Attanasio O., A. Guarin, C. Medina and C. Meghir. 2015. “Long term impacts of vouchers for vocational training: experimental evidence for Colombia.” NBER, Working paper 21390, July.

Attanasio, O., A. Kugler and C. Meghir. 2011. “Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial”. American Economic Journal: Applied Economics. 3 (3): 188-220.

Banerjee, A., Chassang, S., Snowberg, E. 2016. “Decision Theoretic Approaches to Experiment Design and External Validity”. NBER Working Paper 22167. April.

Betcherman, G., K. Olivas, and A. Dar. 2004. “Impacts of Active Labor Market Programs: New Evidence from Evaluations with Particular Attention to Developing and Transition Countries.” Discussion Paper on Social Protection No. 0402. Washington, D.C.: World Bank.

Berniell, L. and D. De la Mata. 2016. "Starting on the right track: Experimental evidence from a large-scale apprenticeship program", CAF Working Paper Series (Mimeo)

Card, D., J. Kluve and A. Weber. 2010. “Active Labour Market Policy Evaluations: A Meta-Analysis.” Economic Journal, Royal Economic Society, Vol. 120(548):F452-F477, November.

Card, D., Kluve, J. and Weber, A., 2015. What works? A meta analysis of recent active labor market program evaluations (No. w21431). National Bureau of Economic Research.

Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. “Dealing with limited overlap in estimation of average treatment effects.” *Biometrika* 96, 187–199.

Dar, A., and P. Z. Tzannatos. 1999. “Active Labor Market Programs: A Review of the Evidence from Evaluations.” Discussion Paper on Social Protection No. 9901. Washington, D.C.: World Bank.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2015. “From Local to Global: External Validity in a Fertility Natural Experiment.” NBER Working Paper 21459. July.

Diaz J. and D. Rosas-Shady. 2016. “Impact Evaluation of the Job Youth Training Program Projoven”. IDB Working Paper Series, Nº IDB-WP-693. IDB.

Escudero, V., Kluve, J., López Mourelo, E. and C. Pignatti. 2016. The effectiveness of Active Labor Market Programs in Latin America and the Caribbean: Evidence from a Meta Analysis, mimeo, Berlin / Geneva.

Flores, C.A., Mitnik, O.A., 2013. “Comparing Treatments across Labor Markets: An Assessment of Nonexperimental Multiple-Treatment Strategies.” *Review of Economics and Statistics* 95, 1691–1707.

Gonzalez Velosa, C., L. Ripani and D. Rosas-Shady. 2012. How can job opportunities for young people in Latin America be improved?, Technical Report 345, Inter-American Development Bank, Washington, DC, United States.

Hirano, K., Imbens, G.W., 2004. “The Propensity Score with Continuous Treatments”, in: Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, Wiley Series in Probability and Statistics. John Wiley and Sons, Hoboken, NJ, pp. 73–84.

Hotz, V.J., Imbens, G.W., Mortimer, J.H., 2005. “Predicting the efficacy of future training programs using past experiences at other locations.” *Journal of Econometrics* 125, 241–270.

Ibarraran, P., Ripani, L. and Taboada, V.: 2014, Life skills, employability and training for disadvantaged youth: Evidence from a randomized evaluation design, IZA Journal of Labor and Development 3(1), 1-24.

Ibarraran, P., J. Kluve, L. Ripani and D. Rosas-Shady. 2015. Experimental evidence on the long term impacts of a youth training program, IZA Discussion Paper 9136, Institute for the Study of Labor, Bonn, Germany.

Ibarrarán, P. and D. Rosas-Shady. 2009. “Evaluating the Impact of Job Training Programmes in Latin America: Evidence from IDB Funded Operations”. Journal of Development Effectiveness. 1 (2): 195-216.

Imbens, G.W., 2000. “The Role of the Propensity Score in Estimating Dose-Response Functions.” *Biometrika* 87, 706–710.

Kluve, J. 2010. “The Effectiveness of European Active Labor Market Programs.” Labour Economics, Elsevier, 17(6):904-918, December.

Kluve, J. 2016. “A review of the effectiveness of Active Labor Market Programmes with a focus on Latin America and the Caribbean.” International Labour Office, Research Department working paper # 9. Geneva.

Kugler A., M. Kugler, O. Herrera and L. Saavedra. 2015. “Long term direct and spillover effects of job training: experimental evidence from Colombia.” NBER, Working paper 21607, October.

OECD. 2013. “Employment and Labour Markets: Key Tables from OECD.” http://www.oecd-ilibrary.org/employment/employment-and-labour-markets-key-tables-from-oecd_20752342.

Pritchett, L., and J. Sandefur. 2013. “Context Matters for Size: Why External Validity Claims and Development Practice Don’t Mix”. Working Paper 336. Center for Global Development. August.

Twose, N. 2015. World Bank Report.