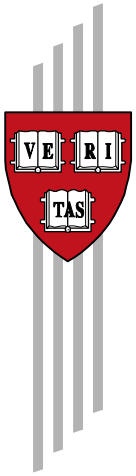
What Works for Active Labor Market Policies?

Eduardo Levy Yeyati, Martín Montané, and
Luca Sartorio

CID Faculty Working Paper No. 358

July 2019



Working Papers

Center for International Development
at Harvard University

What works for Active Labor Market Policies?

Eduardo Levy Yeyati[1]

Martín Montané

Luca Sartorio

July 2019

ABSTRACT

The past 5 years have witnessed a flurry of RCT evaluations that shed new light on the impact and cost effectiveness of Active Labor Market Policies (ALMPs) aiming to improve workers´ access to new jobs and better wages. We report the first systematic review of 102 RCT interventions comprising a total of 652 estimated impacts. We find that (i) a third of these estimates are positive and statistically significant (PPS) at conventional levels; (ii) programs are more likely to yield positive results when GDP growth is higher and unemployment lower; (iii) programs aimed at building human capital, such as vocational training, independent worker assistance and wage subsidies, show significant positive impact, and (iv) program length, monetary incentives, individualized follow up and activity targeting are all key features in determining the effectiveness of the interventions.

Keywords: vocational training, labor policies, wage subsidies, randomized controlled trials. JEL:

J21, J48, E24

[11] Eduardo Levy Yeyati is with Universidad Torcuato Di Tella; Martín Montané is with Universidad Torcuato Di Tella; Luca Sartorio is with Universidad Torcuato Di Tella and the ministry of Production and Labor of Argentina. The usual disclaimers apply.

1. Introduction

In the past 10 years, Active Labor Market Policies (ALMPs) have accounted for more than 0.5% of the GDP of OECD countries. ALMPs is a general denomination for several specific policies that could be broadly grouped into four big clusters: vocational training, assistance in the job search process, wage subsidies or public works programs, and support to micro-entrepreneurs and independent workers.

ALMPs are inherently complex interventions and their incidence depends on a broad range of variables associated with design, context and implementation. This has been implicitly assumed by various recent meta-analysis that centered not so much on the aggregated impact but rather on what specific features makes them work (Card et al., 2010; Card et al., 2017; Escudero et al., 2017). However, these meta-analyses usually group together different evaluations approaches, combining quantitative assessments, and a simplified metric that merges diverse interventions and contexts to get more power at the expense of precision.

Fortunately, the past 5 years have witnessed a flurry of experimental evaluations through Randomized Control Trials (RCT) that shed new light on the impact and cost effectiveness of ALMPs (Figure 1).

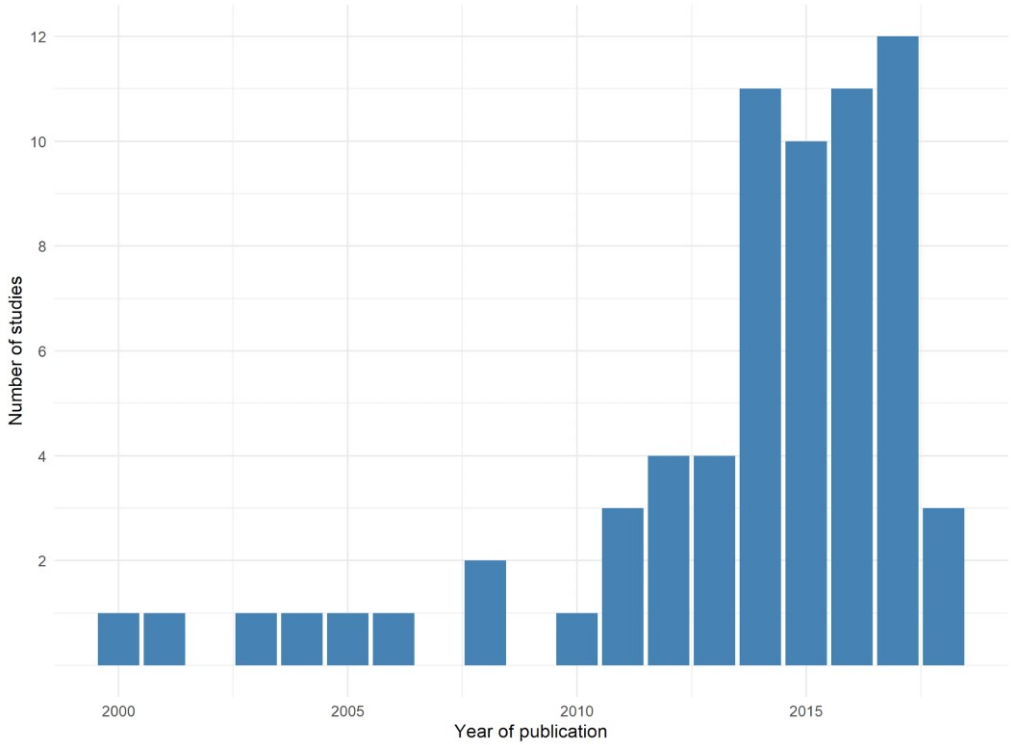


Figure 2. Distributions of studies included in our sample according to the year of publication.

In this paper, we contribute to this blooming literature in two ways:

- **We focus exclusively on programs evaluated through Randomized Control Trials (RCTs).** This choice is not without tradeoffs, as it reduces the number of

relevant evaluations, but allows us to focus in estimates with high internal validity and to refine the metrics used to compare results, making the findings from individual evaluations more naturally comparable. The high number of experimental evaluations carried out since 2014 allows us to assemble a sample large enough to consider exclusively this "gold-standard" methodology that usually represents a minority part of the most extensive reviews of the literature like Card et al. (2010) or Card et al. (2017); indeed, two thirds of our sample comes from papers published in 2014 or after. In the process, we collect data from old and recent evaluations of ALMPs effectiveness to build a workable dataset of 652 impact estimates on employment and income variables from 102 interventions around the globe, evaluated through 73 rigorous impact evaluations with experimental design, covering the four broad groups of ALMPs mentioned above. To our knowledge, this constitutes the most extensive review of the available empirical evidence on these group of policies that is entirely composed of impact evaluations based in an experimental design.[1]

- **We create a metric of new variables to capture the key (implementation, context and target) determinants of ALMP success.** More specifically, we propose a design space that captures standardized variables that characterizes (i) the specific components in which the programs can be decomposed; (ii) the implementation features and the type of public-private participation; and (iii) the economic context and the target population of the programs. This allows us to refine the analysis and identify why policies that are similar in paper can differ in their impact and cost-effectiveness.

Comparing the overall impact of the four policy clusters analyzed, we find that wage subsidies and independent worker assistance show the greatest median impact in earnings relative to the control group, with improvements of 16.7% and 16.5%, respectively. On the other hand, vocational training programs have a median impact of 7.7%, while employment services show an almost negligible impact.

The reported impacts of ALMPs on employment and earnings outputs, although moderately positive on average, are subject to a great variability, possibly due to the multidimensional design space of these policies. To address this, we develop meta-analytic regressions that exploits the descriptive granularity of the proposed design space, seeking to identify policy components associated with a greater probability of success.

The main findings of this exercise can be summarized as follows:

- Wage subsidies show the greatest impact on labor earnings and employment relative to the control group, followed by independent worker

[1] Kluve et al.'s (2019) meta-analysis also benefits from this recent batch of ALMPs´ RCTs, but they restrict attention to youth-targeted programs and complement their sample with other evaluation approaches.

assistance and vocational training programs, while the incidence of employment services is almost negligible.

- However, the ALMPs show great variability in reported impacts:
  - Design and implementation matter: individualized coaching and follow up of the participants, training exclusively to a specific industry and giving monetary incentives to trainees all correlate with better outcomes in vocational trainings programs (the most frequent ALMPs in our dataset); training programs tend to be more effective for young people (we find no difference across genders or educational levels).
  - Context matters: the effectiveness of this kind of programs positively correlates with per capita GDP growth and negatively with the unemployment rate in the year of implementation.
- Although there is little evidence on the delivery costs of the programs, we do find greater volatility in independent training programs, relative to employment services or vocational training programs. On average, employment services are inexpensive, and they stress the need of further research to focus on cost effectiveness analysis.

2. Active Labor Market Policies and its design space

The effectiveness of multidimensional and complex policies, such as ALMPs, depends on how they were specifically designed, on the quality of their implementation, on the context in which they were developed and on their target population. For example, a vocational training program may differ in its cost and duration, in its curricular content, and in whether or not, and how, the private sector participates, and may address a very diverse public, from experienced software programmers in New York or Berlin to disadvantaged youth in the state of Madhya Pradesh. The number of potentially relevant factors and related aspects may quickly render the dimensionality of the comparison practically unmanageable.

On the other hand, an analysis that ignores these considerations can hardly give specific and conclusive lessons for policymakers. Following Pritchett et al. (2013), our four groups (vocational training, wage subsidies or public works programs, support to micro-entrepreneurs and independent workers or assistance in the job search process) can all be thought of as "classes" of policies that could be designed and implemented in very different ways and target diverse demographic groups, with widely varying effectiveness. A review that does not consider this variability could draw conclusions of the type "wage subsidies work" or "vocational training does not work", statements as imprecise as "the ingestion of chemical components works".[2] In order to account for the

[2] Pritchett et al. point out that the question "Does the ingestion of chemical compounds improve human health?" is under-specified, as some chemical compounds are poison and some are aspirin or penicillin and their effects will vary widely depending on the frequency of the applied dose or the particular

dimensionality of the problem in an operational way, we need to approach the evidence from the perspective of a simplified *design space*, namely, a parsimonious version of the space of all of the possible instances of a class of policy, arrived at by specifying all of the choices necessary for a project to be implemented.

Systematic reviews generally consider their policies evaluated as low-dimensional (uniform and with few relevant decisions to make in their design) and with smooth and non-contextual response surfaces. This could be the case of purely "logistical tasks" [3] such as conducting vaccination campaigns in which once we know the "optimal design" of the medical solution and all its contraindications and requirements, the effects will have a strong homogeneity and effectiveness in very dissimilar contexts and there will be almost no relevant decisions to be made in their design beyond ensuring the application of standardized protocols of proven performance. But ALMPs are generally complex policies with high-dimensional design spaces and rugged and contextual response surfaces, highly dependent on a good implementation. Any systematic review that does not exhaustively describe the design space of the policies evaluated and considers the existing variability within the same intervention class, or their interactions with the context and the target population, may have limited use from a practical policy perspective.

A design space for ALMPs

Elaborating a complete and exhaustive design space that homogeneously portrays the evaluated policies requires building a set of standardized variables that characterize at least five fundamental dimensions of the ALMP:

(i) its type,

(ii) its specific components (the content),

(iii) the way it is implemented (including the nature of public-private involvement),

(iv) the implementation cost, and

(v) the implementation context and the target population.

---

conditions of the individual in which it is applied. According to these authors, the currently conventional approach to "the evidence" is of limited value due to the inability to extrapolate the lessons of an impact evaluation of a particular policy to the analysis of another policy that has small changes in some elements of its design (lack of "construct validity") or that has different target populations or is implemented in a different context (lack of "external validity"). For this purpose, he introduces the concept of *response surface*, defined as the average gain on a target indicator of a selected population exposed to a specific program (as an element of the overall design space) compared to those of an ex-ante identical population not exposed.

[3] Following Andrews, Pritchett & Woolcock (2017) a policy intensive in logistical tasks is a policy that requires an important number of agents to implement this policy (is "intensive in transactions") but who do not need to make significant decisions (they do not require "local discretion" from their agents) beyond following established protocols based on known and proven technologies.

To characterize these dimensions, we analyzed all the information available in the academic publications and condensed the description of the criteria used for the identification of each of the variables into unified protocols that articulated the review process.

Types

As noted, our taxonomy starts from four big types of ALMP:

(i) Vocational training

(ii) Wage subsidies or public works program

(iii) Support to micro-entrepreneurs and independent workers

(iv) Assistance in the job search process

In some cases, a program may combine elements corresponding to different policy types, in which case it was classified within these four pure categories according to the main intervention. This dimension is described by a unique categorical variable that indicates the ALMP type.

Content Components

Next, we identify granularly all the specific components of the design of the specific policy. Even if the main intervention of an ALMP is training (classifying it as a "Vocational training" program), it can be complemented with characteristic elements of other types: for example, it may help services in the job-search process as a complement. Thus, the description of the program content allows us to describe more thoroughly programs that entail elements of more than one type.

We define nine possible content components that an ALMP can contain, each one specified by an individual dummy variable:

- Training in technical skills
- Training in soft skills
- Work experience, internship or on-the-job training
- Individualized mentoring or participants tracking
- Assistance services in the job-search process
- Monetary transfer or financing of transportation expenses
- Voucher system
- Asset transference
- Loan
- Duration (a continuous variable measured in months, whenever this information is available).

Implementation features

Even if the specific components of the content of the policies are similar, there may be considerable variations in their delivery, in the existence of incentive schemes for a

correct implementation, and public-private participation in their execution. This dimension tries to capture variables of the design space that characterize the incentives and the nature of the implementation of the policy, independently of its curricular and formative content.

For this purpose, we define the following nine dummy variables:

- Participation of the private sector
- Participation of private firms
- Participation of NGOs
- Participation of multilateral organizations
- Participation of the non-public sectors in the design of the policy
- Participation of the non-public sectors in the financing of the policy
- Participation the non-public sectors provider of the main content of the policy
- Field experiment or pilot
- Orientation to specific economic sectors or occupations

Implementation costs

When the information is available, we add a continuous variable that identifies the average cost per person of the intervention, in 2010 PPP dollars. It is important to stress that only 51 interventions reported this critical variable, and only 22 carried out a rigorous cost-benefit analysis by means of NPV, IRR or payback periods, highlighting an important limitation of the usual practice in the impact assessment literature.

Although the sample of ALMPs for which we have cost data is limited, we can identify some indicative patterns. Wage subsidies, support to independent workers or microentrepreneurs and vocational trainings have comparable median cost per participant, ranging from 1744 and 1518 2010 PPP U.S. dollars, with much greater variability in the second group. In turn, employment services are notably less expensive policies, with a median cost per participant of 277 2010 PPP U.S. dollars and limited variability across programs.

These important differences in median costs –and the dispersion in the case of independent workers programs– highlights the need for precise information on the unit costs of these policies as the policy maker should care not just about what works, but also about what is cost-effective.

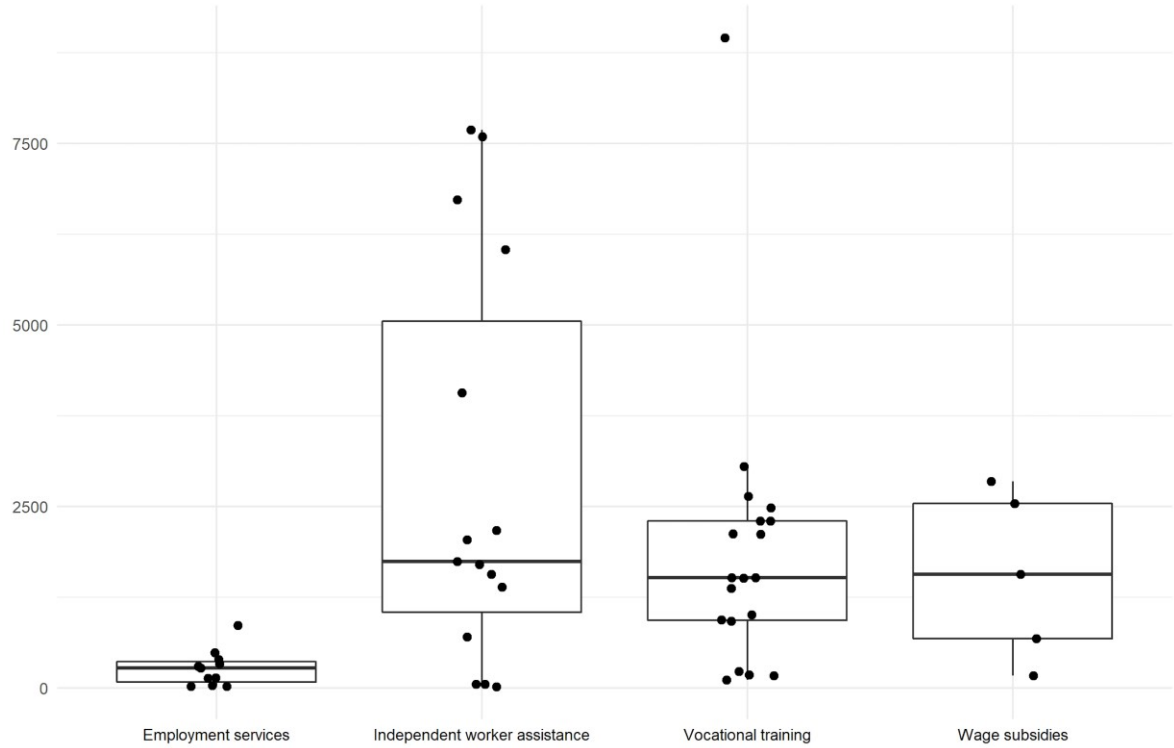


Figure 2. Boxplot of unit costs: cost per treated participant by four-way program classification. 2010 PPP US Dollars.

Context and target population

Our dataset also contains information about the context in which the policies were implemented identifying the country and the cities in which the programs were developed, including whether the intervention was conducted in a rural or urban area.

In addition, a set of variables indicate the demographic characteristics of the target population:

- Average, minimum and maximum age of the participants
- Proportion of participants with a high-school degree
- Proportion of participants with a university degree
- Average years of study of the participants
- Proportion of unemployed participants
- Gender distribution of the participants

3. Results

We reviewed 73 studies that cover a total of 102 ALMPs evaluations, collected from two sources: i) evaluations used in other meta-analysis that addressed similar interventions;

ii) Google Scholar searches[4]. All the programs in our sample were evaluated using Randomized Controlled Trials (RCTs).

Our sample has a high percentage of vocational trainings (45.1%), in line with other meta-analyses. Entrepreneurship training and employment services are almost evenly represented with 24 (23.5%) and 22 (21.6%) interventions respectively. Wage subsidies and public works programs account for the remaining 10% of interventions.

Regarding the target of the programs, only 10 interventions were specially designed for women, and 23 targeted exclusively participant 29 years old or younger. Although the share of programs targeting the youth may look small, the unweighted average of the mean age reported by all the interventions in the sample is only 27.7 years: either by design or by demand, participants tend to rather young.

The interventions took place in countries with widely different income levels, with approximately one half of them being implemented in lower-income or lower-middle-income economies. North America is an exception, with RCT evaluation being the standard way before the rest of the world adopted it.

One key feature of our analysis is the participation of the private sector in the interventions. Other meta-analysis used dummy variables to capture whether the private sector is actively involved in the intervention. We went one step further and created Boolean variables that code the type of participation: 40 interventions were at least partially designed by the private sector, 18 were financed partially or totally by the private sector, while in 68 it took part in the delivery of the services.

Finally, almost half of the interventions were field experiments, and some information on the final cost by participant is available only for 55 interventions.

Type of intervention	n	%
Skill training	46	45.1%
Entrepreneurship training/promotion	24	23.5%
Employment services	22	21.6%
Wage subsidies	7	6.9%
Public works	3	2.9%
Target groups		
Gender		
Only women	10	9.8%
All genders	92	90.2%

[4] Available on request.

Age			
	Maximum age 29 or less	23	22.5%
	Maximum age 30 or more	36	35.3%
	No age restrictions	43	42.2%
Country income			
	Low-income	27	26.5%
	Lower-middle-income	22	21.6%
	Upper-middle-income	20	19.6%
	High-income	33	32.4%
Region			
	Africa	31	30.4%
	Latin America and the Caribbean	20	19.6%
	North America	25	24.5%
	Asia	18	17.6%
	Europe	8	7.8%
Field experiment			
	Yes	48	47.1%
	No	54	52.9%
Private sector participation			
	Designing	40	39.2%
	Financing	18	17.6%
	Implementing	68	66.7%
Information about costs			
	Costs per participant	51	50%

Table 1. Summary statistics on selected variables of our sample of 102 interventions

Impact metrics

Impacts are quantified by the evaluation coefficients. To a question of the type: which is the effect of a training program on the labor income of program participants relative to comparable workers that do not participate in the program, an evaluation study would respond with a coefficient identifying the mean differential effect over the "treated" sample (participants) relative to the control sample (non-participants), as well as some measure of the precision of the coefficient (ideally, its standard error).

From our dataset, we can collect 652 coefficients obtained from 102 interventions, or approximately six coefficients per intervention. If we group these coefficients according to the type of program and the outcome category, we have that (i) there is slightly more coefficients related to employment outcomes (55%) relative to earnings and, more important, (ii) vocational training account for 54% of all coefficients. Since some type of interventions, such as wage subsidies, are underrepresented in our sample, we address

this bias by working with several meta-regressions with different subsamples. Figure 3 illustrates this composition.

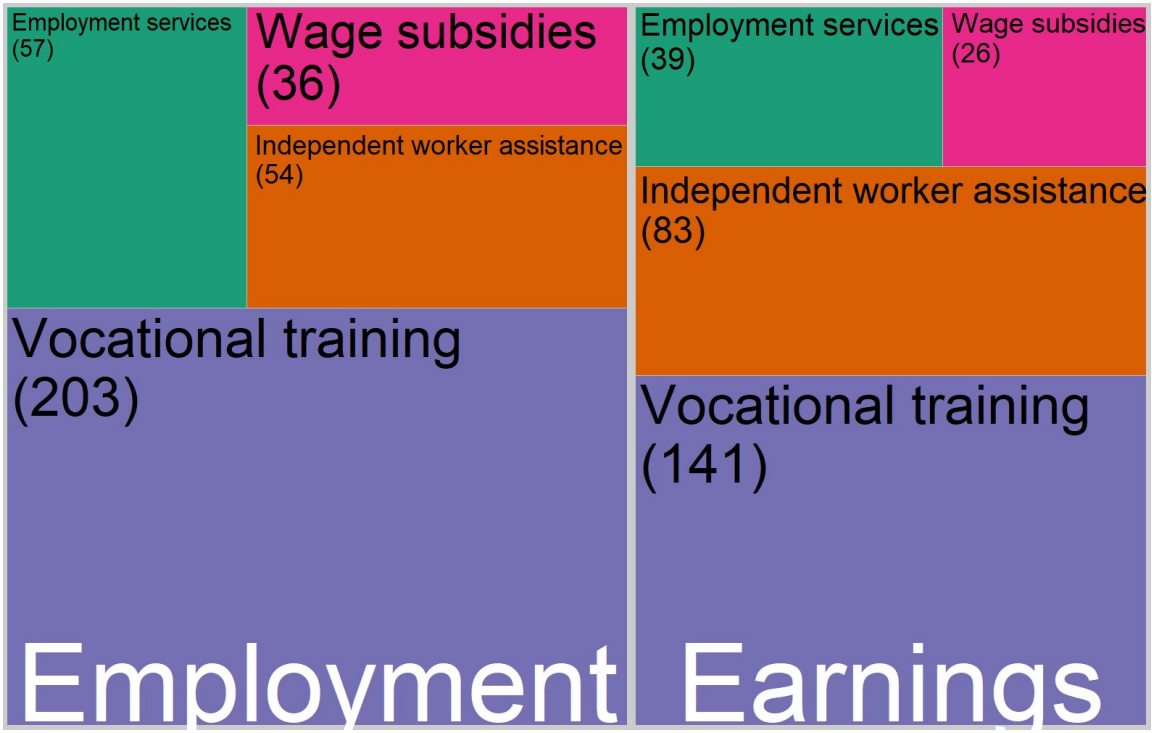


Figure 3. Distribution of the 652 reported coefficients in the 73 selected articles. Left hand side shows the composition of employment-related outcomes. Right hand side shows the composition of the earnings-related outcomes

If we focus on the median impact on earnings, wage subsidies and independent worker assistance show the greatest impact relative to the control group, with improvements of 16.7% and 16.5%, respectively. Vocational training programs have a median impact of 7.7%, while employment services show an almost negligible impact. The median impact on employment outcomes exhibits a similar pattern, with wage subsidies being the type of program which reports the highest impact on this outcome category, while independent worker assistance and vocational training showing a median impact of 11% and 6.7%, respectively. Interestingly, employment services interventions have a median impact of 2.6%, consistent with short-lived and inexpensive interventions that do not attempt to help build human capital, but rather to improve the propensity to find employment.

Importantly, there is a substantial variability in reported impacts on earnings and employment outcomes. This is especially true for the type of interventions in which we have more than 10 cases, such as employment services, independent worker support or assistance and vocational training (Figure 4). We think that this is evidence of the multidimensional nature of the design space in which the programs we are analyzing are deployed. This variability suggests that the impact should be adjusted according to the

different components of each program, the context in which it is implemented and the target population it was designed for.

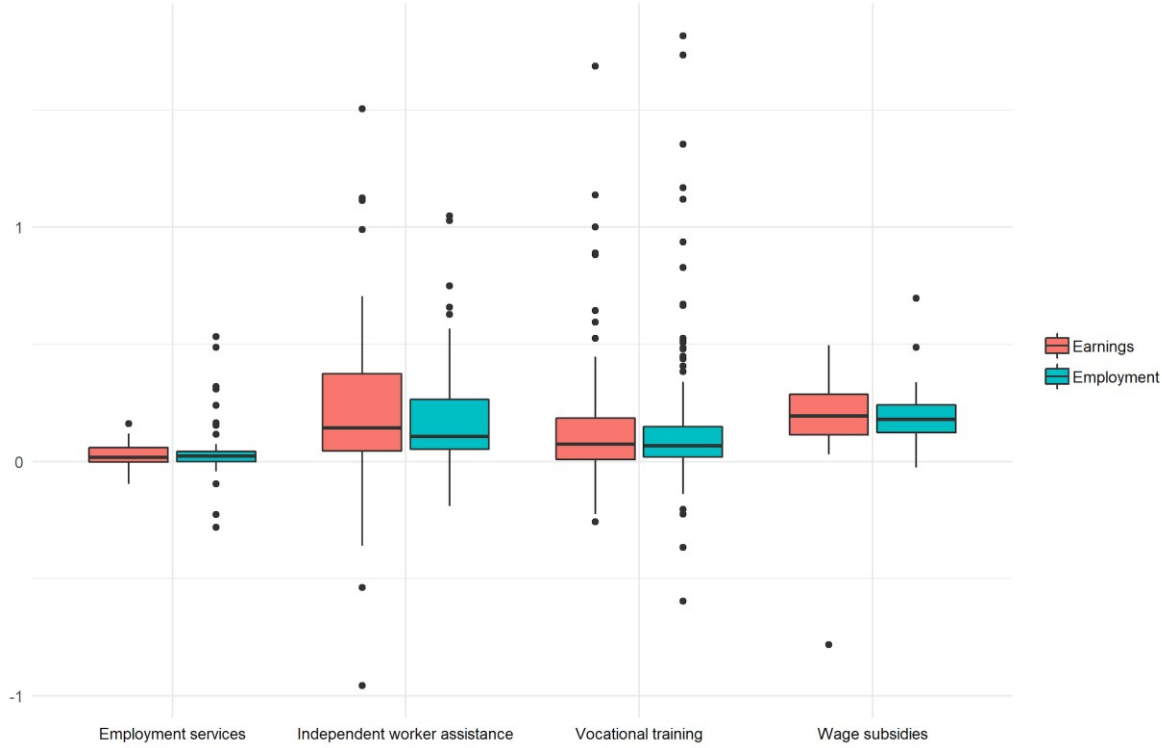


Figure 4. Boxplot of the 652 coefficients according to the estimated effect relative to the control group. Group by type of program and outcome category. Boxes represent the 50% central coefficients reported. The horizontal lines show the median value. The vertical lines show the last coefficient that falls into the +/- 1.5*IQR limit. Points are observations that lay above or below the +/- 1.5*IQR limit.

One key question in this type of interventions is whether they have a lasting effect on the participants. Impact evaluation tend to focus in the short term, especially when they rely on survey data instead of administrative records. Fortunately, our dataset of studies includes reports that estimate the effect of programs even after three or four years. The reported coefficients are most dispersed in the first years, including some negative outcomes, and become gradually less volatile after the second year. More interestingly, the coefficients do not seem to lose their statistical significance over time, as illustrated by the 28 coefficients reported after 4 years of program completion (see Figure 5), or in the coefficient reported in Table 3.

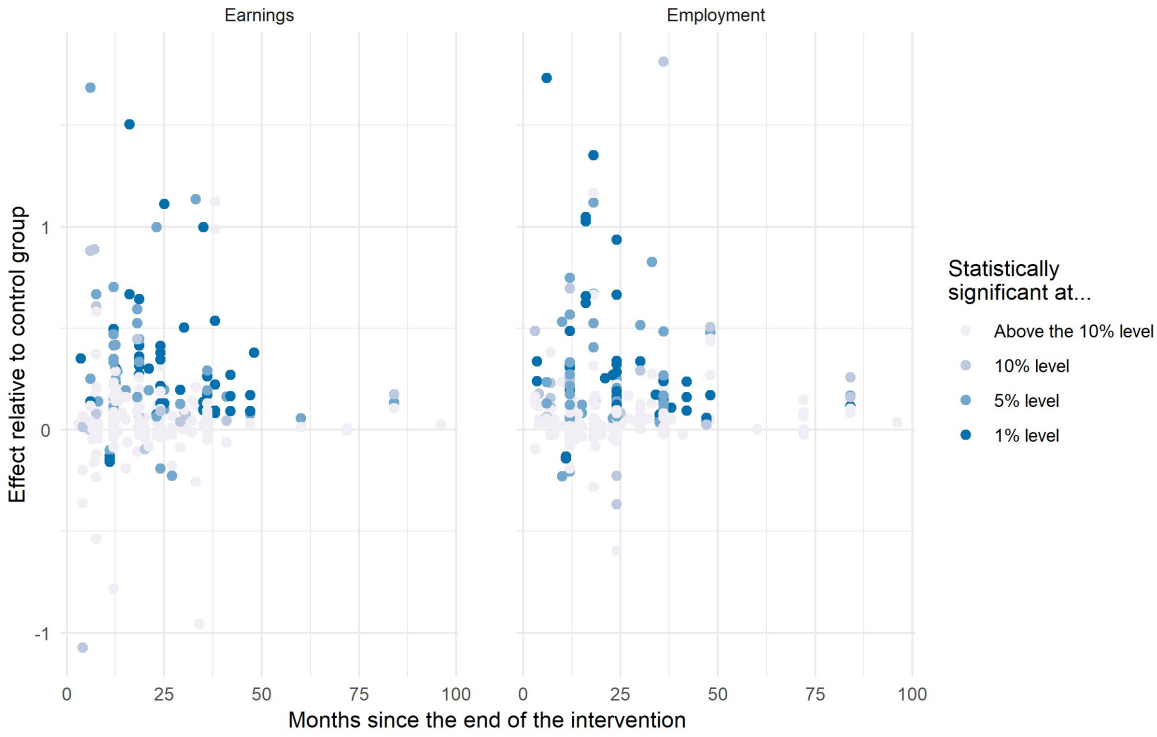


Figure 5. Coefficient reported on earnings and employment status of the treatment group relative to the mean of the control group by months after the end of the program.

Meta-analysis

In this paper we follow Card et al (2017) and Escudero et al (2017) approach to meta-analysis in the ALMP context.

Let assume that the estimated impact of an active labor market program on the outcomes of the participants (b) has an approximately normal distribution with mean (β) and a precision (P) that depends both in the sample size and the design features of the study[5]. Under this usual assumption, an estimate of one of our studies can have the following representation: $b = \beta + P^{-1/2}z$, (1)

where z is a realization of an approximately normal distribution that, if the sample size is big enough, will be close to N (0,1). $P^{-1/2}z$ can be interpreted as the sampling error captured by b.

In turn, we can assume that the limiting program effect, which is the parameter that the impact evaluation studies are trying to estimate, can be decomposed into two terms:

$\beta = X\alpha + \varepsilon$ (2),

Where the multiplication of α, a vector of coefficients, and X, that captures the observed sources of heterogeneity (our design space), measures the "explained" part of the effect

[5] We use Card et al (2017) notation.

on the treated, and ε, an unexplained error term that captures other particularities of the programs that could not be observed or are not included in our taxonomy.

Equations (1) and (2) can be combined, yielding the following model of the observed effects: $b = X\alpha + \mu$ (3),

where X contains the observed characteristics of the programs reflected in our taxonomy, α is the set of parameters that measures the average contribution of the characteristics and μ is the observed error term containing both the unobserved effects and the sampling error.

Unfortunately, some studies do not report the data required to estimate this equation, namely, the estimated average treatment effect and its standard error. Although this problem is less acute in our sample than in past studies, 33% of the point estimates report the *p* value (or a traditional threshold) with no information on their standard errors. To address this problem, we estimate an equation based on a Boolean outcome dependent variable known in the literature as a Positive and Statistically Significant (PSS). The PSS takes the value 1 if the estimated effect is positive and significant below a 10 (or 5) percent threshold and is 0 otherwise. As Table 2 shows, almost no estimate has the wrong sign, so there is no need to estimate the ordered probit model used in Card et al (2017).

Negative and significant at the 5% level	Not significant at the 5% level	Positive and significant at the 5% level
23	419	210

Table 2. Distribution of the coefficients of the impact on the employment status or their earnings according to its sign and statistical significance.

The use of this strategy relies in one broad assumption: that the effect size model yields a vector coefficient close to the binary significance model. Card et al (2017) find that this is the case in the subsample where they have both the effect and the significance value.

Table 3 shows the results for the full sample, ignoring now program components and design variables, except for the inclusion of a control for private sector participation. We start from this specification because, even though the interventions are homogeneous in the sense that all of them are active labor market policies or interventions, the components of the programs are usually program-specific, and they can mask the differential impact of a program type.

First, we can tell that the classification into different types of programs captures part of the effectiveness variation. Job search assistance services tend to be less effective than the rest of the programs, while vocational trainings and support to micro-entrepreneurs and independent workers are less effective than wage subsidies. This difference is like

the variation shown in the median costs of the programs when we had data about them (see Figure 1).

Second, target population show that on average the programs have been less effective for people aged 24 or older, while it seems to be no difference among genders or educational level. Context variables hint that the GDP growth in the year of the implementation of the program correlates with better outcomes, while the income level of the country and the unemployment rate are statistical insignificant at any of the conventional levels.

	Coefficient (t stat)	Confidence interval (95%)	
		Lower bound	Upper bound
Type of program			
Vocational trainings	0.2 (0.07)	0.05	0.34
Supports to micro-entrepreneurs and independent workers	0.2 (0.09)	0.03	0.38
Wage subsidy	0.58 (0.13)	0.32	0.84
Population			
Gender (omitted = pooled)			
Women	0.01 (0.06)	-0.1	0.13
Men	-0.06 (0.06)	-0.17	0.05
Age (omitted = pooled)			
Aged 24 or lower	-0.041 (0.09)	-0.21	0.13
Aged above 24	-0.19 (0.07)	-0.33	-0.05
Education (omitted = pooled or no information)			
Incomplete high school or lower	-0.26 (0.07)	-0.4	-0.13
Complete high school or higher	-0.13 (0.08)	-0.30	0.03
Context			
Lower or lower-middle income country	0.12 (0.08)	-0.04	0.28
GDP growth in the year of implementation	0.02 (0.01)	0	0.04
Unemployment in the year of implementation	-0.01 (0.01)	-0.02	0

Other controls			
Impact measured more than a year after program completion	-0.01 (0.07)	-0.14	0.12
Field experiment	-0.02 (0.06)	-0.15	0.11
Non-public sector participation	0 (0.07)	-0.13	0.13
Number of observations	652		
R squared	0.16		

Table 3. Coefficients, t-stat and confidence interval (at the 95% level) of a linear probability model with a Positive and Significant Sign (PSS) as a dependent variable.

In our sample of evaluated interventions, it is critical to show different models since some of the coefficients are dependent on a specific type of program and, on average, have a wide confidence interval. Table 4 shows coefficients and its significance at conventional level of eight different models. They are the combination of two cut offs for the PSS binary variable (5% and 10%) and four subsamples. Several insights arise from this result.

1. We can see that variables that attempt to capture the context in which the programs were implemented tend to be significant across every subsample. This is especially true among supports to micro-entrepreneurs and independent workers and vocational training programs. GDP per capita growth in the year of the experiment is significant at the 90% level or more in three out of four subsamples, and it is always statistically significant whenever we exclude employment services programs. The unemployment rate at the year of the experiment coefficient is always negative, although only significantly different from zero in the specifications with all programs except employment services and vocational training.

2. We find little evidence of a positive impact of the nonpublic sector in any of the type of participations we considered. Nonpublic sector financing is positive and statistically different from zero at the 90% percent level whenever we include wage subsidies in the sample. Once we exclude wage subsidies, nonpublic financing seems to have no effect.

3. We find some evidence of this kind of policy being more effective in finding formal employment opportunities for the participants rather than boosting earnings or other type of employment, in line with findings reported in other meta-analysis. However, we think that this could be partly due to the null hypothesis testing when working with administrative data, which is always larger than survey data.

4. We find that the individualized coaching and follow up of the participants, the explicit activity targeting and monetary transfers for the participants while actively participating in the programs are all associated with a higher chance of finding an effective response either in earning or employment outcomes. The explicit activity targeting, and the monetary transfers are statistically significant for the vocational training subsample.

	Pooled programs		All programs except employment services		Independent and vocational training		Vocational training	
	PSS (5%)	PSS (10%)	PSS (5%)	PSS (10%)	PSS (5%)	PSS (10%)	PSS (5%)	PSS (10%)
Intercept	0.1	0.21	0.12	0.22	0.09	0.13	0.12	0.1
Category of outcome (omitted = employment)								
Formal employment	0.12	0.17**	0.08	0.14	-0.06	0.08	0.03	0.18*
Earnings	0.03	0.02	0.03	0.03	0.01	0.03	0.07	0.08
Impact measured more than a year after program completion	-0.03	-0.09	0.01	-0.06	0.05	-0.01	0.1	0.07
Nonpublic sector participation								
Designs	-0.11	-0.1	-0.17	-0.16	0.01	-0.02	-0.03	-0.16
Implements	0.08	0.07	0.07	0.07	0.02	0.02	-0.09	-0.04
Finances	0.26**	0.29**	0.3**	0.33***	0.06	0.16	-0.06	0.14
Population properties								
Gender (omitted = pooled gender)								
Women	0	0.04	0	0.04	-0.04	0.02	-0.05	0.03
Men	-0.1*	-0.09	-0.09	-0.07	-0.11	-0.1	-0.1	-0.11
Age (omitted = pooled)								
Aged 24 or lower	-0.08	-0.1	-0.16**	-0.18***	-0.07	-0.08	0.04	-0.05
Aged above 24	-0.19**	-0.27***	-0.2***	-0.27***	-0.13*	-0.21***	-0.05	-0.12

Education (omitted = pooled or not available)								
Highschool dropout or lower	-0.12	-0.05	-0.07	0.01	-0.06	0.02	-0.05	0.03
Complete high school or higher	-0.27***	-0.33***	-0.1	-0.15	0.15	0.07	0.25	0.16
Context								
Lower or lower-middle income country	0.08	0.04	0	-0.04	0.01	-0.06	0.14	0.02
GDP growth in the year of implementation	0.02	0.02	0.02	0.02*	0.04***	0.04***	0.03***	0.04***
Unemployment in the year of implementation	-0.01	0	-0.02**	-0.02**	-0.02	-0.01	-0.03**	-0.02
Other controls								
Field experiment	-0.02	-0.01	0.08	0.06	0.05	0.07	0.12	0.12
Training aims to a specific industry	0.05	0.12	-0.1	-0.01	-0.12	-0.03	0.19	0.32**
Soft skills module	0.07	0.03	0.02	-0.02	-0.1	-0.12	-0.06	-0.07
Individualized mentoring or follow up	0.05	0.08	0.16*	0.19**	0.19**	0.25***	0.08	0.24**
Monetary incentive for the participants	0.04	0.07	0.07	0.09	0.11	0.15	0.19*	0.26***
Program length (omitted = not available)								
A year or less than a year	0.13	0.05	0.2*	0.11	0.16*	0.05	0.07	-0.14
More than a year	0.21**	0.18*	0.3***	0.26**	0.25**	0.2*	0.34***	0.22**

Table 4. Coefficients and p value of the null hypothesis testing of a linear probability model with a Positive and Significant Sign (PSS) as a dependent variable. Each column shows the p value used as cut off for the PSS binary variable. * significant at the 10% level, ** significant at the 5% level and *** significant at the 1% level

Discussion

As we pointed out, without a clearly defined design space that comprehensively characterizes the characteristics of the design, implementation, context and target population of the evaluated policy, the systematic accumulation of learning among the multiple experiences evaluated is clearly limited, inhibiting the finding of identifiable empirical regularities in the determinants of the success of a policy. Generalizing the adoption of a taxonomy validated by the evaluating community would allow the consolidation of information between the different empirical evaluations in a granular way and in universally comparable variables. In the absence of systematization and coordination efforts, the impact evaluation will be limited to the documentation of unconnected experiences and not to the construction of a true collective learning experience.

The proposed design space can serve as a preliminary version of a validated protocol for a systematized description of the different characteristics of an ALMP. Ideally, each academic publication could specify explicitly and in a tabulated form each of the dimensions of the design space of the policies they evaluate. Thus, not only its description would be improved, but its comparability and consolidation would be facilitated, enabling truly informative systematic reviews and an aggregate analysis that provides powerful and granular insights for the policymaker. Through the accomplishment of empirical meta-analysis, this systematized description would allow to isolate specific elements of the design of each kind of policies, identifying empirical regularities that are truly useful for the policymaker.

However, this taxonomy is only a parsimonious version of the many relevant decisions in the design and implementation of an ALMP. The design space suggested in the present work is a first approximation that does not contemplate some crucial aspects of the policies evaluated since these are not systematically described in the compiled studies. Indeed, the limited information that was available in the academic publications corresponding to each evaluation, usually leaves aside several aspects of interest not documented. For example, the synthetic and not systematized descriptions that the academic publications make of the policies they evaluate emphasize with certain rigor the content of an ALMP (the type of training, its duration, its modality, the disciplines or productive sectors involved, etc.). Nevertheless, they say very little about the procedures of its implementation like if they had a competitive selection process for training providers, if there was an incentive system that linked their remuneration to their performance or if there conducted a monitoring process and evaluation of their activity.

We believe that the boom of experimental evaluations has the potential to provide informative evidence for the design of effective policies. However, improving the effective impact of this empirical research will require a coordination effort that facilitates the systematic collection of granular and valuable information.

Encouraging the generation of agreed protocols that require the publication of certain information in a systematized and tabulated registry that characterizes a design space will facilitate and expand the collection of valuable information and enable an aggregate analysis to identify best practices in the design and implementation of ALMPs.

5. Final remarks

Our meta-analysis assessed the effectiveness of more than a hundred Active Labor Market Policies (ALMPs) that were rigorously evaluated through Randomized Control Trials (RCTs), for a total of 652 estimates of program impacts on employment or labor income outcomes obtained from 102 interventions discussed in 73 unique articles.

Given the many dimensions that can influence the effectiveness of this kind of programs, we built a design space to capture the implementation details, the components of the programs, the target population and the context in which the programs were deployed.

We find that impacts on employment and earnings outcomes are moderately positive on average. The median impact on the participants' employment outcomes of the program relative to the control group ranges from approximately 11% for wage subsidies and independent workers support, and 2% for the employment services. Vocational training lays in the center of this range, reporting a median impact of 6.7%

The median impact on the participants' earnings outcomes are higher for wage subsidies and independent workers support, reaching almost a boost of 17% in the earnings outcomes of the reported coefficients, and for vocational trainings, which have a median relative impact of 7.7%. On the other hand, employment services report null effects on earnings.

Although only 51 interventions reported the costs of providing the program per participant, we found that wage subsidies, vocational trainings, and independent workers support do not differ significatively in the median cost per participant, which is approximately 1500 and 1700 USD (2010 PPP). Employment services are inexpensive policies whose median cost is of only 277 USD. However, two caveats should be stressed: (i) there are only 5 observations of the costs of wage subsidies programs and (ii) the dispersion in the costs of independent workers support is high relative to the rest of the programs.

When we focus on the quantitative meta-analysis our main finding is that context matters: GDP per capita growth is positively corelated and the unemployment rate is negatively related to the probability of a program reporting positive and statistically significant (PSS) coefficients. In this regard, we find evidence close to Escudero et al (2017). In the pooled specification we also find some, although not robust, evidence that programs partly or fully financed by the nonpublic sector tend to be more effective. More interestingly, the individual following of the participants is correlated with better

outcomes once we drop the employment services out of the sample (in line with Kluve et al 2019)

In the vocational training subsample, in which our database is denser, we find more interesting results. First, context keeps playing an important role. Second, longer programs tend to be more effective, in line with other meta-analysis (Card et al, 2017).

Third, monetary incentives for the participants to cover for the opportunity costs of taking the program, or maybe just nudging for the participation becomes a statistically significant coefficient (in line with Kluve et al, 2019). Finally, activity-oriented vocational programs are also associated with a greater probability of success.

At the cost of limiting the number of eligible policy evaluations, our analysis has the advantage of comparing studies that are based on a homogenous approach and a powerful causal identification, which lends themselves more easily to the construction of the coefficients used as impact measure.

This effort is a startup of a continuous effort to extract systematic lessons from policy experience, and as such will be updated in the future as the dataset is enriched with new evaluations and descriptive variables which hopefully will fill the gap left by existing evidence, most notably on the cost of the programs, essential for a more reliable full cost-benefit meta-analysis (given the high variance detected in the few cases we could obtain details on the cost per participant).

Bibliography

Andrews, M., Pritchett, L., & Woolcock, M. (2017). Building state capability: Evidence, analysis, action. Oxford University Press.

Card, D., Kluve, J., & Weber, A. (2010). Active labor market policy evaluations: A meta-analysis. The economic journal, 120(548), F452-F477.

Card, D., Kluve, J., & Weber, A. (2017). What works? A meta-analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, *16*(3), 894-931.

Cho, Y., & Honorati, M. (2013). *Entrepreneurship programs in developing countries: A meta regression analysis*. The World Bank.

Escudero, V., Kluve, J., López Mourelo, E., & Pignatti, C. (2018). Active labour market programmes in Latin America and the Caribbean: Evidence from a meta-analysis. The Journal of Development Studies, 1-18.

Grimm, M., & Paffhausen, A. L. (2015). Do interventions targeted at microentrepreneurs and small and medium-sized firms create jobs? A systematic review of the evidence for low- and middle-income countries. *Labour Economics*, *32*, 67-85.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In *Handbook of labor economics* (Vol. 3, pp. 1865-2097). Elsevier.

Kluve, J., Rani, U (2016). A review of the effectiveness of Active Labour Market Programmes with a focus on Latin America and the Caribbean. Geneva: ILO.

Kluve, J., Puerto, S., Robalino, D., Romero, J. M., Rother, F., Stöterau, J., ... & Witte, M. (2019). Do youth employment programs improve labor market outcomes? A quantitative review. *World Development*, *114*, 237-253.

McKenzie, D. (2017). How effective are active labor market policies in developing countries? a critical review of recent evidence. *The World Bank Research Observer*, *32*(2), 127-154.

Pritchett, L., Samji, S., & Hammer, J. S. (2013). It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space. *Center for Global Development Working Paper*, (322).

APPENDIX I – Lessons from past meta-analyses

Along with the publication of new impact evaluations of ALMPs came a series of meta-analysis that tried to infer what works in this kind of interventions, controlling for important variations in design, context and implementation.

Heckman et al. (1999), a seminal within this group, focused in job training, public works, wage subsidies and job search assistance services programs in the United States.[6] They find that programs in the U.S. tended to have a modest but overall persistent positive impact on the labor income of the participants, especially so in the case of adult women. Importantly, they concluded that, under certain scenarios, these interventions seemed to be 'remarkably cost effective'. They also noted that the programs appeared to be effective in raising the earnings of disadvantaged adult males, but ineffective on delivering the same outcomes when it came to disadvantaged youths. They attributed this heterogeneity, in part, to skill differences across groups, suggesting that ALMPs performed comparatively better with skilled participants.

Since this seminal contribution several meta-analyses were published (See Table A1 below). We identify three patterns that are present in the existing literature:

1. There has been an important effort not only to estimate the effectiveness of the programs, but also to identify features in the design, context and implementation of the interventions that are associated with positive outcomes. Thus, some meta-analyses focus exclusively or mostly in Latin American and Caribbean countries (Escudero et al, 2017; Kluve 2016) or in low income countries (Cho & Honorati, 2013; Griff y Pauffhausen, 2015; McKenzie, 2017), while Kluve et al. (2019) look exclusively into programs that target young workers. Even when the universe of interventions is not restricted, the data is usually subset to analysis the heterogeneity in the programs' effectiveness (Card et al, 2017; Kluve 2016).

2. While all the meta-analyses discuss the quality of the data and the methodology of the impact evaluations they work with, RCT impact evaluations usually represent a minor share of all the evaluations included in the analyses, explaining as little as 30% in some cases (Card et al, 2017). While the debate of whether to mix experimental impact evaluations with other types of evaluation design is present and discussed at some length in the texts, and even controlled for in the regressions, the main conclusions about what works is largely drawn from non RCTs impact evaluation.

[6] As the authors note, the country bias was not due to a preference for ALMPs in the US, but rather to a preference for having these policies evaluated, relative to other developed countries.

3. There is some basic agreement in the main findings of these meta-analyses. When comparing the effectiveness of the programs between public and nonpublic implementations, authors conclude that ALMPs that are not implemented by the public sector are associated with a better performance (Cho & Honorati, 2013; McKenzie, 2017; Kluve et al, 2019). Both Kluve (2016) and Card et al. (2017) conclude that public sector employment programs have negligible impact. Also, there is agreement about the magnitude of the impact over different time horizons: effects tend to be larger in the medium and long run[7] (Kluve, 2016; Card et al, 2017; Kluve et al, 2019). Another common finding is that programs targeting long-run unemployment have larger impacts. There is mixed evidence with respect to the sign of the relationship between growth and performance; in particular, in contrast with our findings, one meta-analysis (Kluve, 2016) shows that programs are *more* effective in periods of slow growth and higher unemployment, although the finding does not hold when the tests restricts itself to a Latin America and Caribbean sample.

[7] We do not find this effect in our model and data. This difference could arise from using only RCT impact evaluations, in which the problem of mixing short run estimates when the program has ended with some observations when the program was still at work is reduced.

Paper	Interventions under study	Focus on	Number of studies	Number of programs evaluation	Number of estimates	Impact evaluation methodology	Main findings
Cho and Honorati (2013)	Interventions that aim at promoting potential or current entrepreneurs. It includes Active Labor Market Policies (ALMPs) designed to enhanced technical, vocational or financial skills for self-employment.	Developing countries	37	Not available	1116	Experimental or quasi experimental	They don't find statistically significant difference among types of programs, although when interacting training with counseling the magnitude of the effects tends to be higher. Programs impacts estimated for youth and the urban population use to be positive and significant. NGOs are associated with better performance.
Grimm and Pauffhausen (2015)	Access to finance, entrepreneurship training, business development services, wage subsidies, and improvements to the business environment.	Low- and middle-income countries	53	Not available	116	Experimental or quasi experimental	Finance interventions had lower employment effects than training ones. Interventions targeting small enterprises are more successful than those that target micro-enterprises. Combined interventions did not systematically lead to larger effects, but the combination of finance and training work better compared to when they are isolated.
Kluve and Rani (2016)	Job search assistance, labor market training, private sector employment incentives and public sector employment	No restrictions	207 - LAC sample: 44	526	857 - LAC sample: 152	Experimental or quasi experimental	Effect sizes tend to increase from short to medium-run and is slightly negative between medium and long-run. Public sector employment has negligible or negative impacts. Programs targeting only young and older participants have smaller impacts, relative to those of mixed age groups. Long-term unemployed targeted programs have larger impacts. In periods of slow-growth and high unemployment there are larger impacts; but in LAC GDP growth is positively correlated with effectiveness.

Escudero et al (2018)	Active labor market policies (ALMPs) – i.e. training programs, public works, employment subsidies, self-employment and microenterprise creation programs, and labor market intermediation services	Latin America and the Caribbean	51	53	296	Experimental or quasi-experimental	Interventions with short duration are less likely to produce positive impact compared to longer ones. GDP growth is positively correlated with program effectiveness. Females and youth are more likely to benefit from the programs. Training programs are more successful, mostly impacting over employment formality.
McKenzie (2017)	Labor market policies that have provided vocational training, wage subsidies, job search assistance, and assistance moving	Developing countries	24	22		--	Traditional ALMPs have had at most modest impacts on employment and earnings in most cases. Training is more effective when given by private providers. Subsidies may be useful for temporary employment creation.
Card et al (2017)	Classroom or on-the-job training; job search assistance, monitoring, or sanctions for failing to search; subsidized private sector employment; subsidized public sector employment.	Latin American and the Caribbean	207	526	857	Experimental or quasi-experimental (30% of RCTs)	Average impacts are "more positive" 2-3 years after competition of the program. Programs that emphasize in human capital accumulation have larger average gains. There are larger impacts for females and participants who enter from long-run unemployment. ALMPs are more likely to show positive impacts in a recession. Public sector employment has negligible impacts.
Kluve et al (2019)	Youth-targeted active labor market interventions: training and skills development, entrepreneurship promotion, employment services, and subsidized employment interventions	No restrictions	113	87	3105	Experimental or quasi-experimental (66% of RCTs)	There is no evidence that some programs outperform others but those which integrate multiple services are more successful. Programs in middle- and low-income countries and are more successful. The intervention type is less important than design and delivery. Profiling of beneficiaries, individualized follow-up systems and incentives for services providers (only in high-income countries) matter. In lower income settings, implementation by non-public actors reports larger effect sizes than joint ones. Impacts are of larger magnitude in the long-term.

Table A1. A summary of previous meta-analyses of ALMPs