Do Youth Employment Programs Improve Labor Market Outcomes? A Quantitative Review

Jochen Kluve, Humboldt University Berlin and RWI (jochen.kluve@hu-berlin.de)

Susana Puerto, International Labour Organization (puerto-gonzalez@ilo.org)

David Robalino, World Bank (drobalino@worldbank.org)

Jose M. Romero, World Bank (jromero1@worldbank.org)

Friederike Rother, World Bank (frother@worldbank.org)

Jonathan Stöterau, Humboldt University Berlin and RWI (jonathan.stoeterau@rwi-essen.de)

Felix Weidenkaff, International Labour Organization (weidenkaff@ilo.org)

Marc Witte, University of Oxford (marc.witte@economics.ox.ac.uk)

Correspondence concerning this paper should be addressed to Prof. Dr. Jochen Kluve, RWI – Berlin Office, Invalidenstraße 112, 10115 Berlin, Phone: +49 (30) 2021598-11, Fax: +49 (30) 2021598-19, jochen.kluve@hu-berlin.de

Abstract

Bringing young people into productive work is a key labor market challenge in both developing and developed economies, and a multitude of labor market interventions have been implemented to assist vulnerable youths. To assess whether these interventions have succeeded in improving young people's labor market outcomes, this study systematically and quantitatively reviews 113 impact evaluations of youth employment programs worldwide. Of a total of 3,105 effect estimates we extract from these studies, one-third are positive significant. The unconditional average effect size across all programs is small, both for employment-related outcomes (Hedges' g = 0.05, SE = 0.02) and earnings-related outcomes (Hedges' g = 0.04, SE = 0.02). We analyze correlates of success in a meta-regression framework. We find that (i) programs are more successful in middle- and low-income countries; (ii) the intervention type is less important than design and delivery; (iii) programs integrating multiple services are more successful; (iv) profiling of beneficiaries, individualized follow-up systems and incentives for services providers matter; and (vii) impacts are of larger magnitude in the long-term. Some of these findings provide new and important insights about the design and delivery of interventions, whereas others confirm those of previous reviews. Ultimately, our findings provide practitioners with an improved evidence base about how certain design features contribute to successful youth employment programs in different contexts.

Keywords: Youth employment, Active Labor Market Policy, Impact evaluations, Systematic review, Meta-analysis

JEL codes: J21, J48, E24

Acknowledgements

We are grateful for excellent research assistance to Viviana Perego, Selsah Pasali, Misina Cato, Yamila Simonovsky, Jonas Jessen, Karishma Tiwari, Cheng Qian, Emily Yan, and Rijak Grover. We thank Azita Berar-Awad, Gianni Rosas, Niall O'Higgins, Louise Fox, David Margolis, Pierella Paci, Gloria La Cava, Cem Mete, Roberta Gatti, Omar Arias, Edmundo Murrugarra, Janice Tripney, Hugh Waddington, John Eyers, Martina Vojtkova, Michael Grimm, Anna Luisa Paffhausen, Kristin Hausotter, Gerhard Ressel, as well as participants in an IZA workshop on employment institutions, in the 2015 World Bank/IZA conference on employment and development, in the 2016 ILO/IZA conference, in a 3ie workshop in Washington, DC, in the 2016 annual meeting of Building Evidence on Education, and in the 2015 and 2017 INCLUDE conferences on youth employment, donors of the "MDTF for Job Creation" (BMZ of Germany; SECO of Switzerland; ADA of Austria; and Norad of Norway), donors of the ILO Youth Employment Programme (Swedish International Development Cooperation Agency and the Ministry of Foreign Affairs of Denmark) and to several anonymous referees at the Campbell Collaboration for helpful comments. Financial support from 3ie and the Leibniz Association is gratefully acknowledged.

1. Introduction

The global financial crisis that began in 2007 reversed the gradual declining trend in global youth unemployment rates observed between 2002 and 2007 and lead to increasing youth unemployment between 2007 and 2010. In 2017, the global youth unemployment rate settled at 13.1 percent, with nearly 70.9 million youth actively looking for jobs (ILO 2017). While the unemployment rate among youth is expected to remain relatively constant in the near future, it is still well above its pre-crisis level of 11.7 percent. Youth remain overrepresented among the unemployed and shaken by working poverty and changing patterns in the labor market. Over 160 million youth are working, yet living in poverty (ibid). To respond to the youth employment challenge, many countries have implemented active labor market programs (ALMPs) aiming to connect youth to wage- or self-employment.

Despite the importance of this challenge, little systematic evidence exists about the effectiveness of programs targeting youth employment. While several studies have synthesized research findings on the effectiveness of ALMPs for the general population (e.g., Card et al. 2010, 2017), very few reviews have focused specifically on programs and outcomes for youth. The most relevant review of labor market interventions for youth to date, Betcherman et al. (2007), has served as the basis for technical assistance and policy advice worldwide. Since then, a large amount of research has been produced and published, using experimental or quasi-experimental methods to determine the effects of new – and sometimes innovative – youth employment programs. These studies, for instance, range from experimental evaluations of prototypical skills training for high-school dropouts in the Dominican Republic (Ibarrarán et al. 2014), a wage subsidy pilot for female community college graduates in Jordan (Groh et al. 2016), an innovative

entrepreneurship training for university students in Tunisia (Premand et al. 2016) to a quasi-experimental study of a temporary-work-based approach in Germany (Ehlert et al. 2012).

Whereas there are two recent reviews that have covered parts of this new evidence, they either did not proceed to apply designated empirical methods such as meta-analysis (JPAL 2013), or they have focused on (potentially selective) subsets of the available evidence (IEG 2012). Other review studies have put their emphasis on specific types of intervention or outcomes (Tripney et al. 2013 on TVET programs, Grimm and Paffhausen 2015 on micro-entrepreneurs, both in low- and middle-income countries) or only covered specific regions (Escudero et al. 2017 and Vezza 2014 for Latin America and the Caribbean, Hardoy et al. 2018 for the Nordic Countries). Moreover, knowledge gaps unpacking how issues related to design and implementation affect the effectiveness of different programs have not been previously addressed.

This paper consolidates the evidence from all available rigorous evaluations of ALMPs to understand the relative effectiveness of youth-targeted interventions and some of the key factors that influence their performance. The analysis builds on an extensive systematic review, which relied on a comprehensive search of all available impact evaluations, yielding 113 studies fulfilling inclusion criteria of adequate thematic focus and methodological rigor to be included in a meta-analysis. These 113 studies reported results for a total of 107 different youth employment interventions within 87 separate ALMPs. The systematic search and selection process allows us to address potential issues stemming from publication bias and selective reporting. The criteria for inclusion and exclusion, search strategy, methods and statistical procedures, selection process, and quantitative results are presented comprehensively in the systematic review protocol and

technical report in Kluve et al. 2014 and 2017, respectively. This paper builds on the review's database (Kluve et al. 2018) and expands the knowledge base through new empirical evidence from multivariate meta-regression analysis.

The paper contributes to the existing literature in at least four important dimensions: first, it is based on the most thorough and comprehensive systematic reviewing effort of youth employment programs to date. We collect statistical information to construct a total of 3,105 treatment effect estimates. This sample is considerably larger than in previous studies in labor market economics, and allows us to make more generalizable inference, which is one of the primary goals of this genre of research.

Second, the findings introduce important nuances in the debate on the effectiveness of youth employment programs, in particular with regard to developing countries. In contrast to recent, rather sobering evidence about ALMPs for youth in developed countries (Card et al. 2017), we find that youth employment interventions are more successful in middle- and low-income countries compared to programs conducted in richer economies. Importantly, this result holds when accounting for differences in program and evaluation design of interventions, and is thus not simply an artefact of different types of interventions implemented in contexts that are hard to compare.

Third, while there is no evidence that certain types of programs, or combinations of programs, systematically outperform others, we find that programs that integrate multiple interventions and services are more likely to have a positive impact – in particular in low- and middle-income countries. We conjecture that the success of youth employment programs rests on their ability to respond to multiple needs and constraints facing a heterogeneous group of beneficiaries. One implication is that successful youth

employment programs will need to be able to offer a comprehensive set of interventions, from training to counselling, intermediation, and income support.

And fourth, looking at characteristics that contribute to successful youth employment programs, we find that profiling beneficiaries and individualized follow-up and monitoring systems are crucial design features. Profiling means proactively using information about individual participants to direct them to the services that best fit their constraints, while engagement through follow-up and monitoring means including a focus on features that increase the likelihood that participants finish and/or perform well in the programs. These findings support the idea that interventions responding to specific participant needs are likely to lead to better employment and earnings prospects for their beneficiaries.

In sum, relating our findings to the above-mentioned reviews, we obtain some results that are new, some that add additional detail to previous findings, and some that agree with findings so far. Novel results include our meta-analytical evidence on integrating multiple interventions and on profiling and monitoring characteristics detailed above. Our results are also more detailed by showing, for instance, that the importance of the comprehensiveness of programs only applies to low- and middle-income countries and less so for high-income countries, and by substantiating that there are general differences in youth program effectiveness by country income group, a fact for which until now the evidence had been rather tentative. Finally, our study and newly collected data confirm previous results, in particular the importance of human capital based programs and their dynamic time horizon with increasing effect sizes over the post-program time.

2. Interventions of interest

Youth can face different constraints that affect access to wage- or self-employment; constraints that can be addressed through targeted interventions. For instance, they might not have the necessary skills and/or work experience, they might not have information about job opportunities and/or knowledge about the job search process, or they might be less able to access capital to start a business. This paper seeks to shed light on which type of intervention (or combination of interventions) is most successful in releasing these constraints. We cluster the youth employment interventions into four typologies: (i) training and skills development; (ii) entrepreneurship promotion; (iii) employment services; and (iv) subsidized employment interventions.

The ALMPs examined in this paper include at least one of these intervention types but, importantly, often combine multiple interventions and/or provide different interventions to individual participants. The review hence makes an important distinction between programs and interventions: a youth employment program is a single entity that may consist of one or several interventions. Interventions are components/tracks of programs, or in some contexts also identified as sub-programs within an overall larger program. For example, if a program has a training track and an employment services track, and participants take one or the other, these are considered to be two individual sub-tracks within the same program.[1] It is also possible to find a comprehensive intervention that offers, for instance, both skills training and employment services to the same participant. Program examples consisting of several interventions include the Job Corps program in the United States, the Economic Empowerment for Adolescent Girls program in Liberia, the Projoven program in Peru, and the Employment Fund in Nepal. Interventions are defined based on characteristics such as the type of intervention or the population

targeted. The identification of components within programs allows for the analysis of interactions across interventions.

We hypothesize that participation in ALMPs will ultimately improve the employment and earnings outcomes of youth. Table 1 provides a stylized results chain for each type of intervention, mapping out the causal process from intervention delivery to (potential) labor market effects. For more details on how we define each intervention type, please refer to Appendix 2. Broadly, employment service interventions provide youth with job counseling, job search assistance and/or mentoring services; entrepreneurship promotion interventions provide recipients with business advisory services and/or finance and market access; skills training interventions offer skills training for young people to improve their employability and job access; and subsidized employment interventions comprise wage subsidy and public work programs.

[Table 1 here]

3. Search, selection, and coding of primary studies

The underlying systematic review focused on studies that investigate the impact of single ALMPs on labor market outcomes of young people, and that meet well defined criteria related to program population and context (youth targeting), type of intervention, whether the study is based on counter-factual analysis, looks specifically at labor market outcomes, and several others. The publication format of the study can range from grey literature to articles in academic journals, published or posted between 1990 and 2014.

Inclusion criteria

Population and context. —The review is global in coverage and considers interventions from all countries, regardless of their level of development. Studies must have investigated ALMPs that (i) are designed for – or primarily target – young women and men aged 15 to 35; (ii) target unemployed youth or those with low levels of skills or limited work experience or who are generally disadvantaged in the labor market; and (iii) aim to promote employment and/or earnings/wage growth among the target population, rather than simply providing income support (cf. Heckman et al. 1999).

Intervention. —Eligible studies must evaluate an ALMP that provided at least one of the four categories of interventions (also shown in Table 1) – training and skills development, entrepreneurship promotion, employment services, and/or subsidized employment.

Comparison. —The systematic review includes counterfactual impact evaluation studies that measure change in at least one outcome of interest among intervention participants and relative to non-intervention participants based on a counterfactual analysis (comparing treatment and control groups). Eligible comparison groups include those that receive no intervention or are due to receive the intervention in a pipeline or waitlist study. Hence, we exclude studies that only measure impacts relative to some other intervention without a comparison group. Note that the comparison group of some studies might be exposed to interventions other than the evaluated intervention.

Outcome. —Eligible studies must report at least one measure of one of the primary outcomes of interest, namely: employment, earnings, or business performance. The review also captures outcomes that are measured conditional on other ones and excludes

studies that focus only on intermediary outcomes without measuring impacts on the above-mentioned primary outcomes.

Study design and characteristics. —The review focuses on completed experimental and quasi-experimental evaluations, and considers the following research designs and impact evaluation methods: (i) randomized experiments, (ii) methods for causal inference under unconfoundedness (classical regression methods, statistical matching, propensity score matching), and (iii) selection on unobservables (instrumental variables, regression discontinuity design, difference-in-differences). We consider studies that have been published in a peer-reviewed journal, as a working paper, mimeo, book, policy or position paper, evaluation or technical report, or as dissertation or thesis. Studies published in any language were eligible and the pre-registered search and selection process included a translation of studies where necessary. The date of publication or reporting must have been between 1990 and 2014 (inclusive).

Study search and selection methods

The search for relevant literature was based on a variety of sources to ensure that published and unpublished studies ("grey literature") relevant to the research question are included. The search process consisted of (i) a primary search – searching of a wide range of general and specialized databases, and (ii) a complementary search – hand-searching of relevant websites; searching of dissertations, theses, and grey literature databases; literature snowballing; and contacting authors and experts. The search included search terms in English, Spanish, French, German, and Portuguese. For each source, the review team tested and documented several strategies and identified one or more preferred search strategies that yielded a comprehensive and precise set of potentially relevant results.

The primary and complementary search resulted in 32,117 records based on search in more than 70 sources, including: 12 specialized databases; 11 general databases; 35 websites, such as institutional and conference websites; five dissertations, theses, and grey literature databases; and nine other reviews and meta-analyses. Eighty-six studies were selected as eligible for the analysis. After extracting data from the preliminary set of 86 included studies, the review team screened 6,782 additional records that were identified through reference lists and citation tracking of included studies. This led to the selection of 27 additional studies. Overall, our comprehensive search and selection process resulted in 113 studies considered eligible for inclusion in this review, which report results for a total of 107 different youth employment interventions within 87 separate ALMPs.

The next step in the data collection process was to extract information from each study about the evaluated ALMP, the study sample, and the corresponding effect size estimates using a coding tool and a written coding manual along with other quality assurance mechanisms. [2] The information collected included (i) study characteristics, (ii) intervention characteristics, (iii) characteristics of the subject samples of analysis, and (iv) detailed information about the impact estimates. Rather than coding a risk-of-bias score explicitly for every study, we conduct a design-based bias assessment following Duvendack et al. (2012).

4. Study and intervention characteristics

Characteristics of included studies

Table 2 provides an overview of the 113 studies included in the database. Sixty-five of the studies analyzed interventions from high-income countries and 48 studies covered

interventions from low- and middle-income countries, demonstrating the global coverage of the review and the substantial share of impact studies from developing and emerging markets (Table 2, panel (a)). We were successful in identifying unpublished studies, which is one of the core contributions of this review: only around one-third of the studies is from peer-reviewed journals, with the remainder split between technical reports from implementing organizations and working papers (panel (c)).

Almost half of the studies in the sample were published after 2010, with 21 studies published in 2014 alone (Table 2, panel (b)). Most of the latter are working papers. Table 2 also provides an overview of the types of outcomes that are evaluated by these studies (panel (g)). Within the employment outcome category, most studies estimate the effect on the individual employment probability, while 35 studies estimate the effect on hours worked (not reported separately in the table). Since only entrepreneurship promotion interventions measured business performance outcomes, there are limited observations for these outcomes across the sample.

[Table 2 here]

In contrast to other systematic reviews, we find a large share of experimental studies in the form of randomized control trials (RCTs). Most of the experimental results have been published recently (66 percent after 2010) and, hence, are not included in previous reviews. Figure 1 shows the increase in rigorous experimental evidence; furthermore, before 2011 most RCTs in the sample were conducted in developed countries, while the last five years have seen a remarkable increase in RCTs run in developing countries. Most

notably, in 2014, 12 out of 15 RCTs included in this review come from non-developed countries; five of them evaluated youth employment programs in Sub-Saharan Africa.

Of the 50 quasi-experimental studies in our sample, 18 make use of cross-sectional covariate adjustment (including matching), 12 employ a difference-in-differences (DiD) estimator, six use a panel-based covariate adjustment (including matching), and one study uses instrumental variable techniques. The remaining studies employ other methods, most often a combination of DiD and propensity score matching.

[Figure 1 here]

Regarding the program evaluation features, 39 studies provide impact estimates at multiple time points. In addition, 71 studies measure changes in outcomes of interest at more than 12 months after treatment exposure (Table 2, panel e)). These longer-term effects are estimated primarily for skills training interventions. Relatively few studies provide a sub-group analysis in addition to the overall analysis (Table 2, panel (f)), unless the study did not evaluate an intervention already targeted at a specific group (e.g. females). In particular, only half of the reports in the sample provide separate results for males and females. Only a few reports in our sample provide separate treatment effects for disadvantaged, low-income, or low-educated youth.

Characteristics of evaluated interventions

Some of the 87 youth employment programs covered in this review consist of several interventions. To provide evidence of which interventions and combinations work best, these different types are evaluated separately in the meta-analysis.

Table 3 provides an overview of the characteristics of the 107 interventions in the sample. Many of the interventions provide a combination of different components (panel (c)), with more than 50 percent of studies having skills training as the main category of intervention (panel (a)).[3] In particular, over 30 percent of interventions provide a combination of different components for participants (panel (c)). Among single-component interventions, skills training-only programs constitute a similarly large share (30 percent).

The 113 studies included in this review cover a wide range of countries from all major world regions. As in previous reviews of ALMPs (e.g., Card et al. 2010 and Betcherman et al. 2007), a large share of the interventions that are evaluated have been implemented in developed countries. Another large share of studies comes from Latin America and the Caribbean, where many countries have experimented with ALMPs – in particular, skills training – and started to evaluate their impact early in the 1990s using quasi-experimental and experimental designs. As mentioned above, we include a relatively large number of recent ALMP evaluations from Sub-Saharan Africa (15), which contrasts with the number of evaluations found in other developing and emerging regions. With regard to scale, most interventions have a national coverage. In 30 cases, the evaluations considered localized interventions implemented as pilots (Table 3, panel (f)).

As regards to program targeting, we identify 16 interventions (15 percent) that serve only young women and 45 interventions (42 percent) that focus exclusively on low-income and disadvantaged youth. Forty-five percent of interventions target unemployed youth. With respect to implementation, most interventions have public and private entities delivering services, ranging from the provision of in-classroom training to internships agreements and mentoring. Non-Governmental Organizations (NGOs) appear as

implementers in about one-third of the evaluated interventions in the sample (Table 3, panel (g)).

[Table 3 here]

5. Effect size computations and meta-analysis methods

In order to summarize individual study results and to conduct a multivariate analysis we create two measures from the treatment effects reported in the original studies. The measures are the standardized mean difference (SMD) and a binary variable indicating whether the reported treatment effect is positive and statistically significant (PSS) at a five percent significance level.[4]

The SMD captures the relative magnitude of the treatment effect in a way that is unit-less and hence comparable across outcomes and studies. It is the ratio of the treatment effect (ATT, ITT, or LATE, see below) for a specific outcome relative to the standard deviation of that outcome within the evaluation sample that is used to estimate the treatment effect. The true effect size (θ) is the mean difference between the treatment (μ_t) and control groups (μ_c) as a proportion of the standard deviation of the outcome variables (σ):

$$(1)\ \theta = \frac{\mu_t - \mu_c}{\sigma}$$

The most intuitive form of estimating θ is applying Cohen's d (Cohen, 1988) defined by

$$(2)\ d = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\frac{(n_c-1)*S_c^2 + (n_t-1)*S_t^2}{n_t+n_c-2}}}$$

where $\bar{Y}_t$ is the mean outcome of the treatment group and $\bar{Y}_c$ that of the control group. The numerator of *d* captures the treatment effect and is often reported as a treatment effect parameter estimate, such as an average treatment effect on treated (ATT), intention-to-treat effect (ITT), or local average treatment effect (LATE), rather than as differences in means; thus, we use D to denote a treatment effect estimate. The denominator of *d* is the pooled standard deviation from standard deviations of the treatment and control groups and is equivalent to

$$(3)\ S_p = \sqrt{\frac{(n_c-1)*S_c^2+(n_t-1)*S_t^2}{n_t+n_c-2}}$$

where n_c and n_t are the sample sizes of the control and treatment groups, respectively, and S_c and S_t are the sample standard deviations of the control and treatment groups, respectively. While *d* is an intuitive estimator for θ, it has been shown that *d* has a bias, as it overestimates the absolute value of θ in small samples (Hedges, 1981). For this reason, we use a small sample size adjusted estimator referred to as Hedges' g, which is discussed in Appendix 4.

A challenge we encountered in the data extraction is the limited information available to compute *g*. Standard deviations for the treatment, control, and full sample groups were often missing, even after contacting the original study's authors in attempts to acquire this information. In such cases, the standard deviation of the outcome variable is approximated using the formula from Borenstein et al. (2009b)

$$(4)\ S_p = SE * \sqrt{\frac{n_c*n_t}{n_c+n_t}}$$

where SE is the Standard Error of a means test (e.g. standard error of the regression coefficient estimate).

The median number of treatment effect estimates per study is 12, with some reports providing more than 100 estimates. We implement the following procedure to arrive at summary effect sizes for each study and avoid permitting outcomes of one individual beneficiary influence the aggregate effect size measure twice: first, we identify a single effect size for each same independent group of study participants and per outcome to remove redundancy.[5] In cases without clear justification for dropping some effect sizes over others, we apply combine effect sizes from the same independent population as suggested by Borenstein et al. (2009a). The procedure is explained in detail in Appendix 4.

With one effect size per intervention, we can create aggregate effect sizes for different groupings of interventions as well as an aggregate effect size for the whole sample. Given the breadth of interventions included in our sample it is likely that not all interventions have an identical effect size but, rather, that each intervention's true effect size (θ_i) deviates from the true aggregate effect size for the overall intervention group it belongs to. Furthermore, each observed effect size, estimated by Hedges' g, contains a sampling error, and consequently g_i will either be less than or greater than θ_i. This can be expressed as

$$(5)\quad g_i = \mu + \zeta_i + \varepsilon_i = \theta_i + \varepsilon_i,$$

where μ is the true aggregate effect size for the group as a whole, ζ_i is the deviation of the true effect size of intervention i from the group's aggregate effect, and ε_i is the sampling error. In order to estimate the true aggregate effect size for the group as a whole (μ), equation (5) is estimated using a random-effects regression. Moreover, we use a weighted random effects regression with weights defined as each study's inverse variance to obtain the most accurate estimate of μ.

In the case of PSS, we provide a weighted average of the effect sizes such that each study within the group carries equal weight for the aggregates. As with Hedges' g, the PSS average is based on independent groups created by methods described in Appendix 4.

In this paper, we add to the results from the technical report of the systematic review (Kluve et al. 2017) and extend the quantitative analysis with multivariate meta-regression models. While equation (5) allows us to analyze difference between groups of interventions, multivariate regression models enable us to explore correlations between effect sizes and correlates of interest in the subsequent sections while controlling for potential confounding factors (e.g. differences in study design). We employ three different estimation procedures. In the Appendix 4, we discuss the strength and weaknesses of each approach as well as sources for potential difference in results.

First, we estimate a random effects inverse-variance weighted random effect regressions on Hedges' g as in equation (5). Second, as a robustness check to this estimation procedure, we estimate weighted least square regressions with clustered standard errors on Hedges' g:

$$(6)\quad Y_{ij} = X_{ij}\delta + \varepsilon_{ij},$$

where Y_{ij} is effect size *i* extracted from study *j* and X_{ij} are the relevant covariate values for the sample (or sub-sample) used in estimating Y_{ij}. To control for publication bias, we follow the procedure suggested by Stanley and Doucouliagos (2012). The authors argue that including the squared standard error (i.e. the variance) of the effect size estimate in the weighted least squares model accounts for the potential effect of publication bias, and the resulting coefficient estimate would provide an indication of the magnitude (and significance) of the effect (for a detailed discussion of the effect of small-sample and

publication bias on our results, see Appendix 5). We weight the regressions by the inverse of the number of effect size observations contributed by each intervention and cluster standard errors at the intervention level. Third, for analysis of the bivariate effect size indicator $I_{pss,i}$, we estimate the following probit model via Maximum Likelihood:

$$(7)\quad \text{Prob}(I_{pss,i} = 1|X_i) = \Phi(X_i\delta) ,$$

where $\Phi(.)$ is the Cumulative Distribution Function (CDF) for the standard normal distribution, X_i is a vector of the covariates of interest, and δ is the vector of parameter being estimated. The covariates include intervention characteristics, outcome characteristics, and study characteristics. The results from this model are reported as marginal effects.

For all regressions, we use the fully disaggregated effect sizes to retain the variation with respect to covariates.

6. Results

Aggregate effect sizes

After reviewing the 113 primary studies, we identified and coded 3,567 treatment effects. We have the necessary statistical information to construct the binary PSS indicator variable for a total of 3,105 effect sizes and to compute standardized mean differences in 2,258 cases. Our number of effect sizes is substantially higher than in other systematic reviews: Card et al. (2017) use 857 estimates for 526 different program-type/participant subgroup combinations; Tripney et al. (2013) calculate 92 effect sizes; Grimm and Paffhausen (2015) are able to code 116 effect sizes. Our substantially larger sample of effect sizes is, to some extent, the result of intensive efforts to acquire missing

information from authors: solely relying on primary studies would have provided the required information to compute Hedges' g for only 13 percent of reported treatment effects. Using the methods described in the previous section and Appendix 4, we create the aggregate PSS for the pooled sample as well for several subgroups of interventions and outcomes separately.[6] These resulting average effect sizes are reported in Table 4.

Overall, the results from the random effects model show that many youth employment interventions have a positive and statistically significant effect, but that this does not apply to all sub-groups of interventions. Slightly more than one-third of the impact estimates are positive and statistically significant. Across the various dimensions captured in Table 4 we observe that the percentage of positive statistically significant estimates typically lies in this range, with a few exceptions that are above forty percent (subsidized employment interventions and some particular intervention design features). Results from estimating the aggregate random effects model (equation (5)) are consistent with this, with an overall estimate for Hedges' g of 0.04 and a 95 percent confidence interval ranging from 0.02 to 0.06. The fact that the unconditional mean effect size across all studies is as small does not necessarily imply that youth employment programs are not effective, but that effect size estimates vary significantly across intervention types and/or program features.

Looking at the differences across intervention categories, the unconditional analysis provides initial evidence on whether specific types of programs are more likely to succeed: we find that programs in which the main intervention focus is employment services or subsidized employment (along with interventions whose single focus is not specified) have an effect size that is statistically indistinguishable from zero. We estimate Hedges' g for these three types of interventions quite precisely, without extreme or

outlying values relative to the estimates of other sub-groups, suggesting that our null results are "real" and that the reason for the generally insignificant effect sizes are indeed the small intervention impacts. The reason may be twofold: first, on the design-side, this may be related to within-program / within-study heterogeneity. That is, programs are not well targeted or they are not offering the right combination of interventions to address a diverse set of beneficiaries facing multiple constraints. Second, even if programs are well designed, they may be affected by implementation challenges – in particular whether participants are actively participating and completing all components; and whether program staff are facing the right incentives to implement the program as designed.

Further, there is a substantial contrast between entrepreneurship promotion interventions and skills training interventions with respect to the unconditional aggregate effect sizes. Whereas both types of interventions have similar proportions of positive and statistically significant impact estimates (0.37), the aggregate effect size for entrepreneurship programs is more than twice the aggregate effect size for skills programs (0.12 compared to 0.05), while having a much lower level of precision (standard error of 0.08 compared to 0.02). Nonetheless it seems – from the perspective of unconditional estimates – that entrepreneurship promotion programs have greater treatment effect magnitudes. At the same time, it should be noted that entrepreneurship programs and skills training programs are often implemented in different contexts. There is large variation for skills training programs (partly due to the large number of skills programs in our sample) in terms of target populations, scale, implementers, and location, among other factors. In contrast, the entrepreneurship programs in our sample are more homogeneous, as they tend to be implemented at a smaller scale, targeting poor and disadvantaged populations, and more frequently taking place in lower-income countries.

Hence, the difference in effect size magnitudes should be interpreted with caution and requires testing with conditional analysis, which we do in the multivariate regressions below.

In terms of outcomes, the aggregate estimates indicate comparable findings for earnings and employment results, with a similar proportion of positive and significant impact estimates (0.32 and 0.36) and comparable aggregate effect sizes (0.04 and 0.05). Breaking down these categories into more specific outcomes (where we have a sufficiently large number of observations), we find that the impacts of youth employment programs tend to be the largest on the probability of employment.[7] Youth employment program effects on other employment outcomes, which include labor force participation, unemployment duration, quality of employment,[8] and hours worked, yield smaller values and drive the aggregate effect size for employment outcomes down. For the two most common types of earning outcomes, income and hourly wages, we observe very similar impacts (0.04 and 0.03).

Both the PSS results and aggregate effect sizes show that several elements of the program design have a strong impact on the outcomes of programs, specifically i) including a focus on features that increase the likelihood that participants finish and/or perform better in the programs (engagement), ii) proactively using information about individual participants to direct them to the services that best fit their constraints (profiling), and iii) providing service providers with incentives based on results. The contrast is starkest with respect to the proportion of the estimates in evaluations that were positive and statistically significant. Programs with engagement mechanisms have over twice the proportion of positive and statistically significant estimates (0.41 compared to 0.17) than those without, while interventions that profile participants or incentivize

service providers are each 11 percentage points more likely to report a significant positive impact. However, as with the other unconditional estimates discussed in this section, without conditional (multivariate) analysis the weight placed on these differences is limited, and they cannot be generalized, since often profiling and engagement mechanisms correlate with certain program types. For example, subsidized employment programs by their very nature include financial incentives to continue participation, since there is a direct link between payment and attendance.

[Table 4 here]

Finally, the aggregate estimates provide initial evidence that youth employment programs in certain macroeconomic contexts and targeting certain populations have better evaluation results in general: programs that target low-income populations have larger effects compared to programs without income targeting, and programs in high-income countries have smaller effects compared to those in low- and middle-income countries.

Multivariate meta-regression results

Our main objective is to analyze the effects of covariates on employment and earnings outcomes. We exclude business outcomes from the multivariate meta-analysis since the number of effect sizes is very small, pertains only to the sample of entrepreneurship interventions, and is often measured at enterprise rather than the individual level. As outcome measures, we use the probability of a PSS effect, as well as Hedges' g, which is the SMD corrected for small-sample bias.

In terms of covariates, we focus on four dimensions: the type of program; the country context and income level; individual characteristics of the participants; and general information about program design. We start with the most parsimonious specification, including only the main type of interventions as covariates. As a first step, we add variables to control for evaluation design and publication type. Some meta-analyses aim to control for the quality of evaluations by coding a risk-of-bias score for each effect size based on how well the study addresses potential confounders (e.g. attrition). We favor a design-based approach as suggested by Duvendack et al. (2012) and include dummies for the respective method to control for unobservable sources of confounding in quasi-experimental studies. Acknowledging the caveat that the risk of bias can vary within certain categories of study design, we nonetheless believe this provides a more objective and transparent assessment to which extent empirical approaches correlate with effect size estimates without enforcing an (ex-ante) hierarchy of methods. In this regard, we slightly deviate from standard practice of most systematic reviews. We further include an indicator variable whether the treatment effect estimate represents an ITT estimator rather than the Average Treatment Effect on the Treated (ATET). The rationale is that the two types of estimators can have largely different implications in terms of effect size magnitude, external validity and policy making. Finally, we control for sample size to account for the possibility that large-sample studies are more likely to find statistically significant impacts.

In a second step, we include further explanatory variables that are likely to correlate with the effect size estimate according to our stylized results chain of Table 1. Due to missing information about intervention characteristics in the original studies, the inclusion of further covariates reduces the number of studies in the sample. Hence, the

set of specifications reported aim to strike a balance between informative content and sample size. We provide further specifications in Appendix 6 which show that our qualitative results hold across specifications.

The results for our full sample of effect size estimates are presented in Table 5. Columns one to three of Table 5 show different specifications of the weighted least squares regressions on Hedges' g in equation (6); columns four to six show the same specification using a (inverse-variance weighted) random effects regressions on Hedges' g; and columns seven to nine display the probit regressions on whether the outcome is positive and statistically significant of equation (7). Tables 6 and 7 contain results for estimating the same nine specifications separately in the sample of high- and low-income countries.

Results from full-sample regressions

The regression tables (Table 5) report the three different specifications mentioned above, each comprising different sets of covariates. Moving from specification (1) to (3), the power gains from the additional explanatory variables come at the cost of decreasing number of observations. We add covariates in four blocks: evaluation/study features, program design features, outcome characteristics, evaluation sample, and the implementing organization.

We start by assessing differences in effect sizes across intervention types, with skills training programs as base category against which coefficient estimates should be interpreted. Overall, we do not observe a clear pattern indicating that specific program types or combinations of interventions systematically outperform others in the estimation results for the pooled sample. Even though coefficient estimates are negative in some specifications for certain intervention types (implying that skills training are the most

effective), these patterns are generally not consistent across specifications and regression models. This result suggests that some of the differences in unconditional aggregate effect sizes found in the prior section may not hold when accounting for other features related to the intervention or evaluation design. Following our stylized results chain, this result is not surprising: we expect that the effect of any given intervention on labor market outcomes will depend on the interaction of specific design features with beneficiaries' characteristics and country context. Hence, such design features may explain some of the variation within intervention types and we discuss these in turn.

First, the multivariate regression results support the stylized fact that combining several interventions in one program ("additional services") increases the likelihood of success of a given intervention type. Once controlling for relevant covariates, offering services that complement the main intervention can increase the magnitude of the effect roughly between 0.07 and 0.09 standard deviations and the probability of success by 19 percentage points. This finding is particularly pronounced in the low- and middle-income country sample, as shown in the next sub-section. Most population groups are likely to face multiple constraints affecting their likelihood of getting a job, the types of jobs they get, and the associated earnings. One exemplary program addressing multiple constraints is the Economic Empowerment of Adolescent Girls (EPAG) program in Liberia, which shows strong evaluation results, combines six months of classroom-based training followed by employment services through six months of follow-up support in entering wage employment or starting a business. Also having a positive impact, the Teenage Parent Demonstration in the U.S. aims to address multiple constraints to youth employment. Being mandatory for teenage mothers receiving welfare, it provides a wide array of services that were employment oriented including enrolment in alternative

education programs, participation in job training, job search guidance, and employment. It is not possible, however, to identify the one specific multi-component combination that always works; the types of interventions that are needed seem to be specific to the individual and the country context. Even within the Teenage Parent Demonstration, implementation in some sites included additional specific services like transportation stipends that were not broadly applied.

[Table 5 here]

Second, we document that programs which profile beneficiaries are more likely to succeed and have larger effect sizes. Profiling, for example, allows program managers to better understand and respond to the needs and constraints facing different groups of beneficiaries. A program proactively taking information from participants to enable them to succeed, such as the Programa de Capacitación Jóvenes con Futuro (JCF) in Colombia or the Galpao Program in Brazil, uses information about participant aptitudes to place them in the type of training where candidates may be most likely to succeed. However, profiling does not necessarily imply that each individual needs to have a differentiated treatment. Instead, the ability to group beneficiaries into broad categories – from those who require only minimal support to those who are hard to serve – seems to be critical for the performance of the program. The Adolescent Girls Employment Initiative (AGEI) in Nepal applied an innovative approach by incorporating a results-based system whereby training providers received different bonus payments for successfully placing participants from specific vulnerable populations in “gainful” employment.

Third, the continuous follow-up and engagement of beneficiaries is important for program performance. Following up on beneficiaries is not only necessary to assess whether a given intervention is delivering the expected results, but also to obtain timely feedback in terms of whether adjustments to the intervention are required, both in the composition and intensity of different services. In general, this requires having in place adequate monitoring and evaluation systems. For example, the Women's Income Generation Support (WINGS) in Uganda, which focuses on entrepreneurship promotion activities, requires that staff maintain close supervision of business activities for the first few business cycles and provide advice on meeting market challenges and implementing sound business practices. While continuous follow-ups and monitoring also address the problem of beneficiary drop-out, providing incentives for beneficiaries to stay in the program can also be used for this purpose. The Satya/Pratham program in India, which provides young women with specific skills training, requires beneficiaries to deposit Rs 50 per month for continuing in the program. This means that participants have to be ready to commit a total of Rs 300 for the entire duration of the training program with a promise that upon program completion, they would be repaid Rs 350.

Fourth, the evidence regarding incentives for providers continues to support what we see in the aggregate group statistics (Table 4), but is less persuasive regarding its influence on the proportion of results that are positive and statistically significant. In principle, programs that pay service providers based on results and performance are more likely to have positive impacts. There are different examples of how contracts and payment systems can be structured in this way. The 2008 Employment Package, a Turkish subsidized employment program, combined a payroll tax subsidy to employers for newly hired employees with cuts in their social security payments of employers. In Contrat

Jeune en Entreprise in France, a subsidized employment program with a negligible impact based on its evaluation results, firms are entitled to claim a subsidy whenever they hire an eligible young worker on an open-ended contract. Unfortunately, the information about contracting and payment system in our sample of programs is quite sparse. We are thus only able to distinguish between programs that provide some type of incentive system to providers explicitly mentioned in the program materials and programs that do not have such an incentive.

Fifth, another important finding for the design of interventions is that training programs that focus on soft or non-cognitive skills may not be the silver bullet that many expected them to be. Several studies have emphasized the role that non-cognitive skills have in determining labor market outcomes. For example, programs such as the Jovenes in Latin America and the Caribbean became well known, in part, because of their focus on non-cognitive skills training. There are also examples of soft-skills playing a prominent role in programs with large positive labor market outcomes. For instance, in The Employment Fund (Nepal) girls were taught soft skills during a 40-hour course based on a context-tailored curriculum that identified the most relevant soft skills, including negotiating skills, dealing with discrimination, worker's rights education, sexual and reproductive health, business development, and financial management skills (among others). Our results, however, suggest that other things being equal, programs that include training in socio-emotional, behavioral, and non-technical skills do not necessarily do better than other programs. On the contrary, under some specifications, once we control for key design features, these programs seem to have been less likely to achieve positive outcomes.

The JOBSTART program in the U.S. offers an illustration of a training program with soft skills components and limited impacts. JOBSTART applied an intensive exposure model through which it provided school dropouts with training on work-readiness, life, and communication skills. Another example of an intervention with soft skills training and limited impact is Entra 21, implemented in several Latin American countries by different organizations. In that case, soft skills were broadly defined from writing résumés and job search support to reproductive health (Alzua 2007).

These empirical results from soft skills may be due to the broad set of non-technical skills components that have been applied in our sample, making it difficult to determine empirically which soft skill components are most likely to have positive impacts, and for which target group. Although much of the literature (Cunningham and Villaseñor 2016, Heckman and Kautz 2012) suggests soft skills are important for successful labor market outcomes, the empirical analysis in this literature on individuals' outcomes tend to focus on one type of soft skills, rather than the breadth found in our sample. Duckworth et al. (2007) find that "grit" is key to achieving high-term goals. Almlund et al. (2011) find that the degree to which specific types of skills are important depends on the scenario; i.e. personality traits are more important for lower complexity jobs, higher complexity jobs demand better cognitive skills. In our sample, there was no standard set of skills covered. While the most common skills were communication, work place conduct, team work, work readiness, job search skills, we could identify over 30 different types of skills.[9] Moreover, there was significant variation in the design and implementation of soft skills components found in our sample, such as in modality, duration, and intensity. Training modality in our sample spanned classrooms, workshops, on the job, one-on-one instruction by counselors or mentors, through arts and sports, and through community

service. Duration and intensity went from programs like Jordan Now, who offered an intensive 45-hour course over nine days, to Entra 21's ADEC program in Argentina which covered 76 hours of life skills during its two-year duration.

Following our analysis of program design features, we turn to the relevance of having specific implementing organizations deliver a youth employment program. Results in the pooled sample suggest that programs implemented by the private sector alone, as opposed to joint public-private implementation or sole implementation by the government, lead to moderately larger gains (as shown in the random effect regressions and probit regressions). Our main interpretation of this result is that programs managed by the private sector may be more likely to have built-in incentives to respond to the needs of employers and job seekers. In the next section, we further qualify this finding when disaggregating the sample in high- versus middle- and low-income countries.

Turning to the impact on different participant groups, the multivariate regressions do not show differences in effect sizes based on participants' age or gender. We find some evidence that programs focusing on vulnerable youth – either low-income workers or youth at risk – report larger effect sizes, consistent with the evidence in the aggregates discussed in Table 4. Generally, one would expect that programs dealing with targeted groups should be better able to tailor their interventions according to specific needs. Our analysis, however, suggests that programs that are able to flexibly respond to each participant's individual barriers are more likely to succeed than those who target specific groups with an ex-ante fixed approach to each beneficiary.

In terms of the type of evaluation features, the pooled estimation results show some interesting patterns. First, in line with other meta-analyses (e.g. Card et al. 2017), studies based on randomized experiments generally show smaller effect size estimates than

quasi-experimental studies, regardless of the method of adjusting for potential confounders used in the latter. Second, studies published in peer-reviewed journals also report lower effect size estimates. Third, a larger sample size is significantly correlated with smaller effect size magnitudes, but with a significantly higher likelihood of finding a positive statistical significant effect.[10] Fourth, in line with our expectations, effect size estimates based on an ITT estimator tend to be smaller in magnitude than effect sizes based on ATET estimators, and are less likely to be positive significant. Finally, and most importantly, our meta-regression results clearly show that the duration between program exit and outcome measurement correlates significantly with effect size magnitude: program evaluations that estimate impacts over one year after the intervention are more likely to identify significant results. This finding is in line with the detailed timing patterns carved out in Card et al. (2010, 2017). Our finding indicates that, in most cases, ALMPs do not have immediate effects either on employment rates or on earnings. Consequently, long-term evaluation studies are important to gauge true program success.[11]

Finally, the pooled meta-regression results clearly show that programs implemented in low- and middle-income countries are more likely to succeed and have larger effect sizes than programs implemented in high-income countries. The average effect size is 0.08 to 0.15 standard deviations lower for programs conducted in high-income countries. One explanation is that differences in performance reflect differences in the severity of the constraints facing beneficiaries to access jobs and improve their earnings. High-income countries could be dealing, on average, with population groups that are harder to serve or for whom there are fewer job opportunities relative to those facing beneficiaries in low- and middle-income countries. Moreover, the larger impacts can be associated with low- and middle-income countries having a low starting point, i.e. representing a “catch-up”

phenomenon or conditional growth. We also conjecture that programs in the latter set of countries tend to be newer and might have benefited from better designs and technical innovations.

Results by country income level

Stratifying the sample by country income group addresses a key dimension of impact heterogeneity of youth employment interventions. The types of interventions in the groups of countries are markedly different. For instance, in high-income countries employment services and subsidized employment interventions are more frequent. Middle- to low-income countries, on the other hand, tend to focus more on entrepreneurship programs. The empirical results, estimating the same outcomes and model specifications as in the pooled analysis for both groups separately are found in Table 6 and Table 7. For brevity, we focus the discussion on those results for which we observe differences between the two country groups or results deviating from the pooled results.

As in the pooled analysis, differences between program types are not very strong overall, and few clear patterns emerge that hold across specifications. In high-income countries, for instance, we find some evidence that wage subsidies tend to be less successful than other interventions, which is consistent with other reviews of this type of program (see Almeida 2012). In middle- and low-income countries, studies of employment service programs often report small and/or non-significant effect sizes. However, only few employment service programs remain in the low- and medium income country sample, meaning that our results could be driven by a small number of programs.

Our findings in terms of program design and implementation are less ambiguous and generally support the results of the pooled analysis. Following the five indicators of program design (as in the pooled analysis above), we observe that: first, programs that add extra services to the main intervention tend to do better, regardless of the country income level. Second, the positive correlation of participant profiling in the pooled sample appears to be driven by low- and middle-income country programs (increasing both the proportion of positive evaluation estimates and the effect size magnitude). Third, participant engagement mechanisms are associated with more successful programs, regardless of country income. Fourth, and somewhat surprisingly, incentive mechanisms for service providers are correlated positively with the likelihood of obtaining positive and significant program effect sizes in high-income countries, but are negatively correlated in low- and middle-income settings. Finally, in line with the pooled results, we do not find evidence that programs delivering soft skills report more positive effect size estimates in either country group.

With regards to the role of the implementing organizations, we observe marked differences across country groups. For high-income countries, the evidence suggests that programs implemented by the private sector (including NGOs) or public sector alone perform worse than those jointly implemented in a public-private partnership. The results for low- and middle-income countries reverse this pattern and show there is an advantage in outcomes for those programs that were solely implemented by the public or private sector. Both report on average around 0.11 standard deviations larger effect sizes that programs implemented jointly by governments and the private sector. This finding suggests that, while public-private collaborations can be beneficial, they require a strong

institution set-up which may not always be the case for programs implemented in low- and middle-income countries.

[Table 6 here]

[Table 7 here]

7. Conclusion

Labor market prospects for youth are a cause of concern for policymakers worldwide. As a consequence, many programs have been implemented to bring young people into the labor market, connect them to jobs, increase their earnings, and/or help them set up and grow a business. However, the majority of these programs have not been properly evaluated and therefore there is, to date, limited information available about the types of interventions that work and the reasons why.

This paper aims to improve our understanding of the effectiveness of youth employment programs, focusing on skills training, entrepreneurship promotion, employment services, and subsidized employment. To this end, we identify all relevant empirical studies with rigorous evaluations produced over the last ten years. We create a database with 113 studies and code information about 3,402 treatment effects on employment and earnings, beneficiaries, and program design and implementation (Kluve et al. 2017). Our multivariate meta-analysis draws on three different empirical approaches to ensure robust results: (i) random effects meta-regressions on the small-sample corrected standardized mean difference (Hedges' g); (ii) Weighted least square

regressions on Hedges' g; and (ii) probit regressions on whether the original evaluation estimate is reported positive and statistically significant.

The results of the meta-analysis show that, on average, evaluations of youth employment programs report statistically significant positive effects. However, the unconditional average magnitude of the effect size is small (the overall estimate for Hedges' g is 0.04 with a 95 percent confidence interval ranging from 0.02 to 0.06). Furthermore, just above one third of the programs in our database display statistically significant positive effects. The interpretation of this result is not that youth employment programs do not work. Instead, differences in performance seem to be related to design and implementation factors, as well as the characteristics of the country and population of beneficiaries. Our meta-regressions suggest several important factors that correlate with program success.

First, there is no evidence that certain types of programs, or combinations of programs, systematically outperform others. Rather, we find evidence that programs that integrate multiple interventions and services are more likely to have a positive impact, in particular in low- and middle-income countries. Hence, while there is no specific combination of services that always works, programs that add complementary services to the main intervention, regardless of what those are, tend to do better. The interpretation is that the success of youth employment programs rests on their ability to respond to multiple needs and constraints facing a heterogeneous group of beneficiaries. In other words, the efficient portfolio of services is specific to the population of beneficiaries. Programs that target multiple categories of beneficiaries are likely to need multiple portfolios of services. One implication is that successful youth employment programs will need to be able to offer a

comprehensive set of interventions, from training to counselling, intermediation, and income support.

Second, youth employment interventions are more successful in middle- and low-income countries. One interpretation could be that these programs are more recent and might have benefited from innovations in design and implementation. However, our result holds when accounting for differences in program and evaluation design of interventions implemented in low-income countries. An alternative conjecture may be that labor market constraints in middle- and low-income countries are more easily released. In addition, the programs' investments in such contexts may more strongly affect highly disadvantaged populations, e.g. low-skilled and low-income youth.

Third, we find evidence about the importance of profiling and individualized follow-up and monitoring systems in determining program performance. Consistent with the finding above, programs that profile beneficiaries are also able to better respond to their needs. But profiling does not necessarily imply having services tailored to each individual. Instead, it often involves being able to group beneficiaries in broad categories, from those requiring minimal support to the most disadvantaged or hard-to-reach. Efficient follow-up systems and incentives to keep youth in the program are consequently also critical for success. This often implies having in place robust monitoring and evaluation systems that allow a positive feedback loop to improve program performance.

Fourth, evidence about the importance of incentive systems for services providers is positive, yet not uniform, pertaining only to programs implemented in high-income countries. At least conceptually, programs that pay providers based on performance are more likely to achieve their objectives. It remains unclear, however, what the best types of contracting and payments systems are and how these need to be adjusted depending on

the context. We are unable to capture these differences in design and therefore only code whether a given program offers some type of incentive at all.

Fifth, the evidence suggests that involving non-public actors in the implementation of youth employment programs leads to moderately higher gains compared to pure public sector implementation. The added benefit of such collaboration appears much stronger in high-income settings suggesting that complementarities among government, private sector and civil society implementers can effectively factor in the needs of employers and jobseekers and translate into better labor market outcomes of youth. In lower income settings, sole implementation by non-public actors reports larger effect sizes than joint implementation. In those contexts, the success of multi-sectoral collaboration may be hampered by insufficient strategic alignment and institutional development.

Last, but not least, we show that evaluation design matters: most importantly, our meta-regression models show that the timing of outcome measurement is clearly correlated with reported effect size magnitude and statistical significance. This result shows the importance of evaluating programs in the medium- and long-term to gauge their success. At the same time, there are no systematic differences in reported effect sizes related to the age or gender of beneficiaries for which the effect size is estimated. Only programs that focus on vulnerable populations more often report large effect sizes.

Relating our findings to other reviews, several of these results are new, some add additional detail to previous findings, and some confirm findings so far. Novel results include our meta-analytical evidence on integrating multiple interventions and on profiling and monitoring characteristics. Our results are more detailed, for instance, when carving out important differences by country income group. Finally, our newly collected

data confirm, for instance, the importance of human capital based programs and their positive effect dynamics over the post-program time horizon.

The findings of this systematic review also bring to light the importance of including information on program costs in impact evaluations. Our findings provide insights into mechanisms and designs features that may improve youth employment program performance, but the general unavailability of standardized information on program costs limits our ability to make absolute statements about the efficient allocation of resources available to improve outcomes. The sporadic presentation of standardized program costs measures alongside impact evaluation results may be the largest gap in our knowledge of what works and why to improve labor market outcomes of youth.

REFERENCES

Almeida, R., Behrman, J., & Robalino, D. (2012). The right skills for the job? Rethinking training policies for workers. Human Development Perspectives. Washington, DC: World Bank.

Almlund, M., Duckworth, A., Heckman, J. J., & Kautz, T. (2011). Personality Psychology and Economics. In E. Hanushek, ed., *Handbook of the Economics of Education*, Volume 4, 1–181. Amsterdam: North Holland.

Betcherman, G., Godfrey, M., Puerto, S., Rother, S. F., & Stavreska, F. A. (2007). A review of interventions to support young workers: Findings of the youth employment inventory. Social Protection Discussion Paper Series, No. 715. Washington, DC: World Bank.

Borenstein, M., Cooper, H., Hedges, L.V., & Valentine, J.C. (2009a). Effect sizes for continuous data. In H. Cooper, V. Hedges, & J. C. Valentine (Eds.). *The handbook of research synthesis and meta-analysis*, 221–235. New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009b). *Introduction to Meta-Analysis.* Hoboken, NJ: John Wiley & Sons, Ltd.

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*. doi: 10.1093/jeea/jvx028.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum.

Cunningham, W., Sanchez-Puerta, M.L., & Wuermli, A. (2010). Active Labor Market Programs for Youth: A Framework to Guide Youth Employment Interventions. Washington, DC: World Bank.

Cunningham, W.V., & Villaseñor, P. (2016). Employer Voices, Employer Demands, and Implications for Public Skills Development Policy Connecting the Labor and Education Sectors. *The World Bank Research*, Volume 31, pp. 102 – 134.

Duckworth, A., Peterson, C., Matthews, M.D., & Kelly, D.R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), pp. 1087-1101.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634.

Ehlert, C., Kluve, J., & Schaffner, S. (2012). Temporary work as an active labor market policy: Evaluating an innovative program for disadvantaged youths. *Economics Bulletin* 32, 2, 1765– 1773.

Escudero, V., Kluve, J., López Mourelo, E., & Pignatti, C. (2017). Active Labour Market Programmes in Latin America and the Caribbean: Evidence from a Meta-Analysis. Ruhr Economic Papers #715. Essen, Germany: RWI.

Grimm, M., & Paffhausen, A.L. (2014). Do interventions targeted at micro-entrepreneurs and small and medium-sized firms create jobs? A systematic review of the evidence for low and middle income countries. *Labour Economics* 32 (2015) 67–85.

Groh, M., Krishnan, N., McKenzie, D., & Vishwanath, T. (2016), Do Wage Subsidies Provide A Stepping-Stone To Employment For Recent College Graduates? Evidence From A Randomized Experiment In Jordan, *The Review of Economics and Statistics* 98(3), 488–502.

Grosh, M., del Ninno, C., Tesliuc, E., & Ouerghi, A. (2008). For protection and promotion: The design and implementation of effective safety nets. Washington, DC: World Bank.

Hardoy, I., Røed, K., Zhang, T., & von Simson, K. (2018). Initiatives to Combat the Labour Market Exclusion of Youth in Northern Europe: A Meta-analysis, In M.A. Malo & A. Moreno Minguez (Eds.), *European Youth Labour Markets. Problems and Policies, 235–253.* Springer Nature.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451-46.

Heckman, J.J., LaLonde, R.J., & Smith, J.A. (1999). The economics and econometrics of active labor market programs, in Ashenfelter, O. and D. Card (Eds.), *Handbook of Labor Economics 3*. Amsterdam: Elsevier.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators, *Journal of Educational Statistics* 6(2), 106-128.

Ibarrarán, P., Ripani, L., Taboada, B., Villa, J.M., & García, B. (2014). Life Skills, Employability and Training for Disadvantaged Youth: Evidence from a Randomized Evaluation Design. IZA *Journal of Labor & Development* 3, 10.

IEG (2012). *World Bank and IFC support for youth employment programs,* Washington, DC, World Bank.

ILO (2017). *Global Employment Trends for Youth 2017. Paths to a better working future.* Geneva, International Labour Office.

J-PAL (2013). *J-PAL Youth Initiative Review Paper*. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab.

Kluve, J., Puerto, S., Robalino, D., Romero, J.M., Rother, F., Stöterau, J., Weidenkaff, F.

& Witte, M. (2014). Protocol: Interventions to improve labor market outcomes of youth: a systematic review of training, entrepreneurship promotion, employment services, mentoring, and subsidized employment interventions. Oslo, Norway: Campbell Collaboration.

Kluve, J., Puerto, S., Robalino, D., Romero, J.M., Rother, F., Stöterau, J., Weidenkaff, F. & Witte, M. (2017). Interventions to improve labor market outcomes of youth: a systematic review of training, entrepreneurship promotion, employment services, and subsidized employment interventions. Oslo, Norway: Campbell Collaboration.

[dataset] Kluve, J., Puerto, S., Robalino, D., Romero, J.M., Rother, F., Stöterau, J., Weidenkaff, F. & Witte, M. (2018), Youth Employment Program Quantitative Review Database, Research Data Center Ruhr at the RWI, http://en.rwi-essen.de/forschung-und-beratung/fdz-ruhr/.

Premand, P., Brodmann, S., Almeida, R., Grun, R., & Barouni, M. (2016), Entrepreneurship Education and Entry into Self-Employment Among University Graduates, *World Development* 77, 311-327.

Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*, Routledge, New York, NY.

Tripney, J., Hombrados, J., Newman, M., Hovish, K., Brown, C., Steinka-Fry, K., & Wilkey, E. (2013). Technical and vocational education and training (TVET) interventions to improve the employability and employment of young people in low- and middle-income countries: A systematic review. Campbell Systematic Reviews 9.

Vezza, E. (2014). Policy scan and meta-analysis: Youth and employment policies in Latin America, CEDLAS Documento de Trabajo No. 156 (La Plata, Universidad Nacional de la Plata).

World Bank. (2015). The role of skills training for youth employment in Nepal: an impact evaluation of the employment fund. Adolescent Girls Initiative Results Series. Washington, DC: World Bank Group.

Tables

Table 1 Types of interventions and constraints

Type of constraint faced by youth	General Type of intervention used to address	Rationale	Services under this type of intervention	Risks	Illustrative examples
Information gap (lack of adequate information about job opportunities and lack of information about skills of young applicants by employers), limited access to networks, obstacles to applying for jobs (e.g. high transport costs)	Employment and Intermediation Services	Creating mechanisms that make information exchange between (for) employers and workers less costly.	Information Systems/ Counselling, based on accurate labor market information Mentoring Training, Job search assistance, Support services	Displacement of employment (no new jobs created).	Programa Inserjovem (Portugal), Jordan New Opportunities For Women (NOW)
Limited Access to Credit; Lack of financial capital, Limited Social Networks, Limited know-how in setting up a business, bookkeeping, and similar skills, Value chain exclusion or disconnect	Youth Entrepreneur Promotion Programs	Directly supplying young entrepreneurs with access to the specific inputs needed for a business to succeed.	Microfinance, business skills training, assessments by experienced professionals, facilitating access to value chains, mentorships that teach management and other know-how (marketing, business registration)	Moral-hazard, low-potential projects, and expensive assessments of business profitability.	Women's Income Generation Supports (WINGS, Uganda), The Prince's Trust (UK)
Inadequate supply of skills – technical, cognitive, and non-cognitive, Low Skill Level, No or little work experience, Skills Mismatch (youth are not trained for the jobs requested by employers), Missing "soft" non-cognitive skills, lack of basic skills (numeracy/literacy)	Skills training for young people	Training workers with the technical, vocational, non-cognitive skills that makes them desirable to firms	Different types of training: technical and vocational skills, business skills, literacy and/or numeracy, behavioral and non-cognitive skills that are implemented both in classrooms or on the job (OJT)	Governance of program.	Job Corps (US), Chile Joven
Little or no work experience among youth Minimum Wages and mandatory benefits (e.g. social security contribution) Workers whose productivity is not high enough to outweigh the costs of employment of youth with little or no work experience	Subsidized Employment	Lowering hiring and labor costs of employing workers to allow them to gain work experiences which makes them more productive and propels them into their career path.	Direct payments to employers, tax deductions to employers, direct payments to workers, public works.	Deadweight loss (creation of more jobs than required).	JUMP wage subsidies (JWS, Germany), Youth Hires (Canada)

Table 2 Study characteristics

	n	%		n	%
a) Country income level			**e) Timing of Evaluation Follow-up**		
High-income	65	58%	Follow-up <= 1 year	58	51%
Low and middle-income	48	42%	Follow-up > 1 year	71	63%
			f) Sub-group analysis in addition to the overall analysis		
b) Year of Publication					
1991-2000	14	12%	Gender Disaggregated	56	50%
2001-2005	20	18%	Low-income participants	4	4%
2006-2010	27	24%	Education level of participants	13	12%
2011-2014	52	46%			
			g) Outcome Category		
c) Type of Publication			Employment	98	87%
Peer-Reviewed Journal	41	36%	Earnings	91	81%
Working Paper	28	25%	Business Performance	10	9%
Evaluation / Technical Report	30	27%			
Other (Book / Dissertation)	14	12%	**h) Main Intervention Type**		
			Skills Training	74	65%
d) Evaluation Design			Entrepreneurship Promotion	12	11%
Experimental	53	47%	Employment Services	11	10%
Natural Experiment	12	11%	Subsidized Employment	17	15%
Quasi-experimental	50	44%	Unspecified	9	8%

Note: Number of studies=113. Reports may not be exclusive across the different typologies in this table. E.g. one study may estimate multiple outcomes or look into more than one intervention type.

Table 3 Intervention characteristics

	n	%		n	%
a) Main category			**d) Country income level**		
Skills Training	55	51%	High-income country	60	56%
Entrepreneurship Promotion	15	14%	Low and middle-income country	47	44%
Employment Services	10	9%			
Subsidized Employment	21	20%	**e) Intervention Region**		
Unspecified	6	6%	OECD	56	52%
			Sub-Sahara Africa	15	14%
b) Has Component			Europe and Central Asia	4	4%
Skills Training	68	64%	Latin America and Caribbean	22	21%
Entrepreneurship Promotion	18	17%	Middle East and North Africa	6	6%
Employment Services	40	37%	South Asia	4	4%
Subsidized Employment	25	23%	**f) Scale of Intervention**		
c) Combinations			National	59	55%
Skills Training Only	31	29%	Regional	21	20%
Entrepreneurship Promotion Only	14	13%	Local or pilot	30	28%
Employment Services Only	9	8%			
Subsidized Employment Only	12	11%	**g) Intervention features**		
Skills Training and Entrepreneurship Promotion	2	2%	**Target group:**		
Skills Training and Employment Services	24	22%	Women only	16	15%
Skills Training and Subsidized Employment	8	7%	Unemployed at intervention start	48	45%
Entrepreneurship Promotion and Employment Services	1	1%	Low-Income/Disadvantaged Youth	45	42%
Employment Services and Subsidized Employment	3	3%	**Implemented with participation of:**		
Skills Training and Employment Services and more	3	3%	Government	75	70%
			Private Sector	63	59%
			NGO/Non-profit	37	35%
			Multilateral organization	11	10%

Note: N=107 interventions.

Table 4 Aggregate effect sizes

	Percent Positive and Statistically Significant	Aggregate* Hedges' g	95% Confidence Interval (G)	
			Lower Bound	Upper Bound
Main category of program				
Skills Training	0.37	0.05***	0.02	0.07
Entrepreneurship Prom.	0.37	0.12***	0.04	0.19
Employment Services	0.17	0.00	-0.03	0.03
Subsidized Employment	0.41	0.02	-0.01	0.05
Unspecified	0.24	0.04	-0.03	0.10
Outcomes				
Earnings Outcomes (combined)	0.32	0.04***	0.02	0.05
Employment Outcomes (combined)	0.36	0.05***	0.03	0.06
Specific outcomes:				
Employment Probability	0.39	0.06***	0.04	0.08
Number of hours/days worked	0.26	0.03**	0.00	0.06
Income	0.35	0.04***	0.02	0.06
Salary/Wage	0.35	0.03***	0.02	0.05
Design features				
Offered extra services:				
Yes	0.37	0.05***	0.02	0.09
No	0.33	0.04***	0.02	0.05
Profiled participants:				
Yes	0.39	0.06***	0.02	0.10
No	0.28	0.04***	0.01	0.06
Participant engagment mechanism:				
Yes	0.41	0.05***	0.03	0.06
No	0.17	0.03*	0.00	0.06
Incentives to service providers:				
Yes	0.43	0.06***	0.03	0.09
No	0.32	0.05***	0.01	0.08
Type of participant				
Male	0.35	0.04***	0.02	0.06
Female	0.30	0.08***	0.04	0.11
Low income/disadvantaged:				
Yes	0.35	0.06***	0.03	0.10
No	0.34	0.03***	0.01	0.05
Country				
High Income Countries	0.34	0.02**	0.00	0.04
Low and middle income (combined)	0.37	0.09***	0.06	0.12
Low income	0.34	0.15***	0.10	0.21
Middle income	0.38	0.06***	0.03	0.10
Sub-saharan Africa	0.31	0.14***	0.09	0.19
Latin American and Caribbean	0.50	0.10***	0.05	0.15
Total	**0.35**	**0.04*** **	**0.02**	**0.06**

*Notes: All aggregated measures estimated are based on independent groups. Aggregate Hedges' g represents estimate of μ from random effects inverse variance weighted regression (see equation 5 above). SMDs are winsorized at the 1% and 99% level, and cleaned of outliers above 0.75 or with a standard error above 0.75. For variable definitions see Appendix. *, ** and *** denote statistical significance at 10 percent, 5 percent and 1 percent level of significance respectively.*

Table 5 Meta-regression results: full sample

All countries pooled	Weighted Least Squares Hedges' g regressions			Random Effects SMD regressions			PSS probit regressions		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Main intervention category									
(base=skills training)									
Entr. Prom.	0.03	0.059	0.067*	0.036**	0.002	0.104***	0.028	0.024	0.079
	[0.59]	[1.33]	[1.70]	[2.09]	[0.08]	[4.14]	[0.25]	[0.21]	[1.00]
Empl. Serv.	-0.059**	0.015	0.056	-0.026**	-0.001	0.089***	-0.138	-0.028	0.178
	[2.38]	[0.48]	[1.04]	[2.57]	[0.10]	[3.17]	[1.45]	[0.29]	[1.18]
Subs. Empl.	-0.036	-0.084**	0.03	-0.031***	-0.035**	0.027	0.035	-0.235***	-0.105
	[1.27]	[2.14]	[0.88]	[3.56]	[2.04]	[1.20]	[0.42]	[2.62]	[1.06]
Unspecified	0.001	-0.006	-0.061	-0.041***	-0.044***	-0.038*	-0.093	0.057	0.007
	[0.02]	[0.09]	[0.80]	[3.06]	[2.79]	[1.81]	[1.09]	[0.37]	[0.03]
Evaluation features									
Log Evaluation Sample Size		-0.023**	-0.025**		-0.008**	-0.012***		0.049*	0.050**
		[2.40]	[2.56]		[2.20]	[2.95]		[1.87]	[2.06]
Publication Peer-Reviewed		-0.005	-0.022		-0.01	-0.006		0.032	0.014
		[0.19]	[0.84]		[0.96]	[0.61]		[0.43]	[0.17]
Intention-to-treat estimator		-0.026	-0.009		-0.021*	-0.008		-0.087	-0.115*
		[0.65]	[0.19]		[1.74]	[0.69]		[1.18]	[1.94]
Variance	0.092	-0.908*	-0.649						
	[0.17]	[1.86]	[1.15]						
Research Design *(base=RCT)*									
RDD		-0.007			0.009				
		[0.09]			[0.09]				
DiD		0.122**	0.144**		0.016	0.133***		-0.007	
		[2.03]	[2.28]		[0.84]	[3.97]		[0.05]	
Cross-sectional C.A.		0.094	0.106**		0.003	0.096***		0.176	0.181
		[1.48]	[2.11]		[0.15]	[3.65]		[1.57]	[1.28]
Panel C.A.		0.036	0.091		-0.028	0.091**		0.183	0.397*
		[0.60]	[1.29]		[1.27]	[2.45]		[1.12]	[1.92]
Combined/other		-0.019	0.003		0.011	0.069***		-0.19	-0.006
		[0.41]	[0.08]		[0.68]	[4.64]		[1.32]	[0.06]
High income country		-0.110***	-0.146***		-0.080***	-0.156***		-0.256***	-0.469***
		[3.47]	[4.39]		[6.69]	[10.58]		[3.04]	[4.20]
Program design features									
Additional services		0.017	0.089***		0.002	0.105***		-0.025	0.125
		[0.59]	[3.39]		[0.15]	[6.88]		[0.29]	[1.31]
Participant profiling		0.048	0.065**		0.030**	0.041***		0.104	0.113
		[1.60]	[2.44]		[2.58]	[2.75]		[1.51]	[1.35]
Participant engagement mechanism		0.094***	0.079***		0.064***	0.088***		0.276***	0.356***
		[2.90]	[3.72]		[5.15]	[7.32]		[2.95]	[3.47]
Incentives for service providers		0.060*	0.03		0.031***	0.060***		0.096	-0.028
		[1.82]	[0.96]		[2.79]	[4.62]		[1.09]	[0.33]
Program has soft skills training			0.018			-0.003			-0.141
			[0.40]			[0.19]			[1.26]
Outcome characteristics									
Employment outcome			0.008			-0.014**			-0.018
(base=earnings outcome)			[0.43]			[2.07]			[0.39]
Estimated unadj. diff. in means			-0.12			-0.006			0.021
			[1.66]			[0.44]			[0.14]
Measured over one year			0.066***			0.051***			0.270***
after exit from program			[3.36]			[6.12]			[4.01]
Target/evaluation group						0.043**			
Low income / disadvantaged			0.021			[2.53]			0.172*
			[0.68]			-0.004			[1.86]
Male			-0.041*			[0.37]			-0.08
(base=male and female combined)			[1.70]			-0.01			[1.52]
Female			0.001			[1.18]			-0.069
(base=male and female combined)			[0.07]			0.007			[1.24]
Younger Participants			0.006			[0.83]			0.007
			[0.18]						[0.07]
Type of implementer									
(base=jointly public&private)									
Government only			-0.034			-0.049**			-0.168
			[0.75]			[2.04]			[1.08]
Private sector only			0.044			0.039**			0.112
			[1.21]			[2.25]			[0.93]
Constant	0.071***	0.189**	0.136	0.049***	0.115***	0.002			
	[3.79]	[2.50]	[1.62]	[12.47]	[3.71]	[0.06]			
R2	0.02	0.21	0.34	.	.	.			
N	2,000	1,365	985	1,999	1,365	985	2,932	1,437	1,015
Number of interventions:	97	57	38	97	57	38	105	57	38
Number of studies:	96	69	52	96	69	52	104	66	50

Notes: Regressions weighted by inverse of number of observations coming from each intervention and errors clustered at the intervention level. Marginal effects evaluated at the variable means reported for probit regressions. Research designs refer to: RCT = Randomized Controlled Trial; RDD = Regression Discontinuity Design; DID = (Regression-adjusted) Difference-in-Differences; Cross-Section C-A = Covariate-adjustment (matching- or regression-based) with cross-sectional data; Panel C-A = Covariate-adjustment (matching- or regression-based) with panel data; Combined/other = Combination of methods or other methods. For variable definitions see appendix.
* $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 6 Meta-regression results: high-income country sample

High-income countries	Weighted Least Squares Hedges' g regressions			Random Effects SMD regressions			PSS probit regressions		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Main intervention category									
(base=skills training)									
Entr. Prom.	-0.073***			-0.034			-0.327**		
	[2.98]			[0.83]			[2.18]		
Empl. Serv.	-0.044	0.048	0.494***	-0.013	-0.034	0.415***	-0.137	-0.152	1.778***
	[1.53]	[0.83]	[4.41]	[1.15]	[1.59]	[3.47]	[1.10]	[0.99]	[12.82]
Subs. Empl.	-0.033	-0.026	-0.028	-0.023**	-0.014	-0.05	0.026	-0.337**	-0.809***
	[0.93]	[0.52]	[0.34]	[2.37]	[0.60]	[1.03]	[0.25]	[2.44]	[6.46]
Unspecified	-0.011	-0.031	-0.171	-0.026*	-0.034*	-0.090***	-0.092	0.001	-0.854***
	[0.28]	[0.52]	[1.67]	[1.84]	[1.83]	[3.25]	[0.98]	[0.01]	[14.06]
Evaluation features									
Log Evaluation Sample Size		-0.025*	-0.02		-0.007	-0.015***		0.066**	0.027
		[2.00]	[1.23]		[1.35]	[2.85]		[2.37]	[1.39]
Publication Peer-Reviewed		-0.033	-0.044**		-0.024	-0.023		-0.247**	-0.01
		[0.70]	[2.41]		[1.43]	[1.45]		[2.40]	[0.39]
Intention-to-treat estimator		-0.006	0.164		-0.026	0.014		-0.470***	-0.178***
		[0.08]	[1.23]		[1.41]	[0.93]		[3.67]	[5.25]
Variance	2.938**	1.010***	1.182						
	[2.58]	[2.92]	[1.66]						
Research Design *(base=RCT)*									
RDD		0.058			0.055				
		[0.56]			[0.47]				
DiD		0.077	0.377**		-0.025	0.217***		-0.185*	
		[1.13]	[2.79]		[1.08]	[4.60]		[1.73]	
Cross-sectional C.A.		0.031	0.327**		-0.035	0.340***		-0.342**	0.920***
		[0.32]	[2.27]		[1.21]	[4.76]		[2.30]	[10.11]
Panel C.A.		0.012	0.313**		-0.069**	0.123		-0.419**	0.799***
		[0.14]	[2.29]		[2.35]	[1.32]		[2.26]	[5.05]
Combined/other		-0.068	0.249*		-0.024	0.104***		-0.705***	-0.051
		[0.89]	[1.98]		[0.87]	[3.19]		[4.15]	[0.98]
Program design features									
Additional services		0.059	0.088		-0.006	-0.035		0.062	0.871***
		[1.11]	[1.10]		[0.34]	[0.99]		[0.43]	[13.22]
Participant profiling		-0.007	0.029		0.048*	0.153***		-0.092	-0.840***
		[0.20]	[0.51]		[1.72]	[2.71]		[1.11]	[10.56]
Participant engagement mechanism		0.117*	0.082		-0.024	0		0.494***	0
		[1.86]	[0.81]		[1.38]	[0.01]		[3.14]	
Incentives for service providers		0.071	-0.117					0.240**	-0.222***
		[1.49]	[1.10]					[2.39]	[2.64]
Program has soft skills training			-0.728**			-0.473***			-1.830***
			[2.42]			[2.83]			[7.70]
Outcome characteristics									
Employment outcome			0.008			-0.004			-0.038
(base=earnings outcome)			[0.55]			[0.57]			[0.86]
Estimated unadj. diff. in means			-0.146			-0.039**			-0.035**
			[1.66]			[2.03]			[2.30]
Measured over one year			0.027			0.063***			0.13
after exit from program			[0.65]			[6.14]			[0.84]
Target/evaluation group									
Low income / disadvantaged			0.793***			0.594***			2.675***
			[3.30]			[4.00]			[13.96]
Male			-0.014			0.015			0.022
(base=male and female combined)			[0.72]			[1.34]			[1.58]
Female			-0.043*			-0.019			-0.032
(base=male and female combined)			[2.11]			[1.58]			[0.63]
Younger Participants			0.039			-0.004			-0.027
			[0.89]			[0.38]			[0.70]
Type of implementer									
(base=jointly public&private)									
Government only			-0.844***			-0.422***			-1.699***
			[3.27]			[3.11]			[9.70]
Private sector only			-0.259***			-0.158			-0.794***
			[3.15]			[1.60]			[5.06]
Constant	0.033	0.062	-0.099	0.033***	0.135***	-0.265**			
	[1.43]	[0.65]	[0.40]	[6.80]	[2.91]	[2.27]			
R2	0.07	0.13	0.64	.	.	.	.	.	.
N	1,299	788	489	1,298	788	489	1,715	841	456
Number of interventions:	53	27	16	53	27	16	60	27	16
Number of studies:	54	37	27	54	37	27	60	36	26

Notes: Regressions weighted by inverse of number of observations coming from each intervention and errors clustered at the intervention level. Marginal effects evaluated at the variable means reported for probit regressions. Research designs refer to: RCT = Randomized Controlled Trial; RDD = Regression Discontinuity Design; DID = (Regression-adjusted) Difference-in-Differences; Cross-Section C-A = Covariate-adjustment (matching- or regression-based) with cross-sectional data; Panel C-A = Covariate-adjustment (matching- or regression-based) with panel data; Combined/other = Combination of methods or other methods. For variable definitions see appendix. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 7 Meta-regression results: low and middle-income country sample

Low- and middle-income countries	Weighted Least Squares Hedges' g regressions			Random Effects SMD regressions			PSS probit regressions		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Main intervention category									
(base=skills training)									
Entr. Prom.	0.046	0.052	0.04	0.02	0.050**	0.005	0.079	0.05	-0.026
	[0.81]	[1.03]	[0.92]	[1.17]	[2.18]	[0.17]	[0.61]	[0.35]	[0.38]
Empl. Serv.	-0.055	0.04	-0.009	-0.029	0.021	-0.08	-0.136	0.172	-0.197
	[1.16]	[0.81]	[0.11]	[0.87]	[0.53]	[1.54]	[1.08]	[1.24]	[1.14]
Subs. Empl.	-0.026	-0.123	0.103**	0.006	-0.077**	0.118**	0.057	-0.231	-0.037
	[0.71]	[1.70]	[2.56]	[0.24]	[2.56]	[2.17]	[0.46]	[1.51]	[0.27]
Evaluation features									
Log Evaluation Sample Size		-0.028*	-0.031		-0.020***	0.007		0.049	0.101**
		[1.83]	[1.55]		[2.90]	[0.67]		[1.03]	[2.39]
Publication Peer-Reviewed		-0.015	0.004		0.035**	0.037*		0.117	-0.035
		[0.36]	[0.09]		[2.41]	[1.70]		[1.11]	[0.31]
Intention-to-treat estimator		-0.038	-0.04		-0.018	-0.01		-0.007	-0.019
		[0.84]	[1.26]		[1.06]	[0.53]		[0.08]	[0.34]
Variance	-0.695	-1.162**	-0.689						
	[1.37]	[2.27]	[1.08]						
Research Design *(base=RCT)*									
DiD		0.133			0.041			0.107	
		[1.70]			[0.64]			[0.37]	
Cross-sectional C.A.		0.1	0.070*		0.061**	0.062		0.203	0.341*
		[1.46]	[1.77]		[2.46]	[1.60]		[1.30]	[1.84]
Combined/other		-0.005	0.061		0.052***	0.067***		-0.021	0.272***
		[0.09]	[1.42]		[2.77]	[2.65]		[0.09]	[2.79]
Program design features									
Additional services		0.032	0.151*		0.054**	0.224***		-0.119	0.575***
		[0.55]	[1.89]		[2.37]	[4.77]		[0.89]	[3.48]
Participant profiling		0.089**	0.088**		0.100***	0.125***		0.160*	0.330***
		[2.34]	[2.30]		[4.94]	[5.00]		[1.67]	[4.26]
Participant engagement mechanism		0.098*	0.041		0.050**	0.007		0.367***	0.153
		[2.03]	[1.09]		[2.32]	[0.23]		[2.84]	[1.40]
Incentives for service providers		0.043	-0.034		0.061***	-0.024		0.118	-0.024
		[0.82]	[0.88]		[2.99]	[0.81]		[0.73]	[0.21]
Program has soft skills training			0.006			-0.062**			-0.506***
			[0.08]			[2.10]			[3.20]
Outcome characteristics									
Employment outcome			0.033			-0.021*			0.095
(base=earnings outcome)			[1.17]			[1.69]			[1.42]
Estimated unadj. diff. in means			0.180***			0.117			0
			[4.25]			[1.17]			
Measured over one year			0.039*			-0.002			0.191***
after exit from program			[1.79]			[0.14]			[3.06]
Target/evaluation group									
Low income / disadvantaged			0.03			0.039			-0.044
			[0.99]			[1.58]			[0.38]
Male			-0.044			-0.036**			-0.035
(base=male and female combined)			[1.14]			[1.98]			[0.50]
Female			0.027			0.009			0.014
(base=male and female combined)			[1.07]			[0.69]			[0.26]
Younger Participants			-0.014			0.040*			-0.008
			[0.38]			[1.91]			[0.05]
Type of implementer									
(base=jointly public&private)									
Government only			0.067			0.120**			0.204
			[1.04]			[2.35]			[1.38]
Private sector only			0.041			0.082***			0.446***
			[0.68]			[2.71]			[3.16]
Constant	0.106***	0.232*	0.218*	0.085***	0.130**	-0.044			
	[3.98]	[1.88]	[1.98]	[13.82]	[2.19]	[0.57]			
R2	0.02	0.28	0.29	.	.	.			
N	701	577	496	701	577	496	1,217	596	513
Number of interventions:	44	30	22	44	30	22	45	30	22
Number of studies:	42	32	25	42	32	25	44	30	24

Notes: Regressions weighted by inverse of number of observations coming from each intervention and errors clustered at the intervention level. Marginal effects evaluated at the variable means reported for probit regressions. Research designs refer to: RCT = Randomized Controlled Trial; DID = (Regression-adjusted) Difference-in-Differences; Cross-Section C-A = Covariate-adjustment (matching- or regression-based) with cross-sectional data; Combined/other = Combination of methods or other methods. For variable definitions see appendix. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Figures

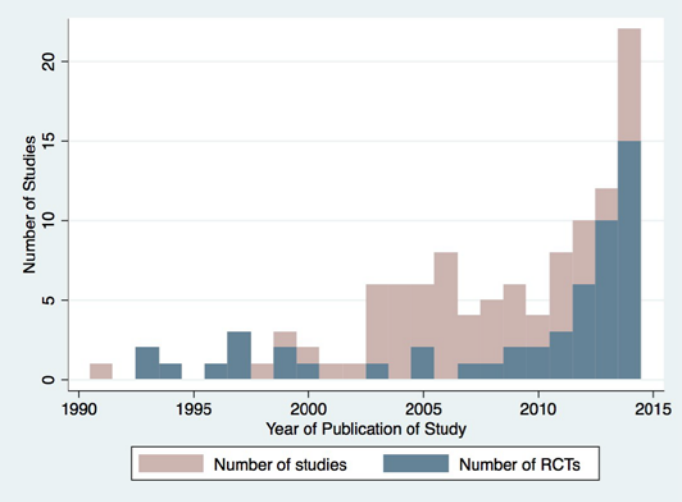

FIGURE 1: NUMBER OF STUDIES BY YEAR OF PUBLICATION

Endnotes

[1] Note that, by this definition of track, we are assuming that each intervention within a program has separate groups of participants that do not overlap. On the other hand, if a program has a single track that includes two different services, whereby the same participants take both training and employment services, this would be considered a single intervention consisting of two different types.

[2] The coding tool and coding manual can be retrieved from the authors on request.

[3] The review defines "main category of intervention" as the largest and predominant intervention type within a program. If several intervention types are equally distributed across the target population (i.e., an individual is exposed to more than one intervention type with the same level of intensity), the main category of intervention is classified as comprehensive.

[4] We consider a treatment effect as statistically significant if it has a p-value from a two-tailed test of less than 0.05.

[5] By redundancy we mean providing additional information about a group that is not needed for the desired level of aggregation. For example, if the goal is to create program aggregates for all participants, then male and female sub-group estimates may be dropped. On the other hand, if the goal is to create an aggregate for females for each program, then pooled estimates would be dropped.

[6] In this table, as in our main regressions, we winsorize standardized mean differences (Hedges' g) at the 1% and 99% level, and drop outliers above 0.75 or a standard error above 0.75. Our results are robust to various measures of dealing with outliers.

[7] To provide a more representative picture we have combined employment and unemployment probabilities.

[8] Quality of employment captures intervention effects on outcomes such as attaining a fixed contract and receiving benefits.

[9] Aside from those listed above other skills include problem solving, successful work habits, social skills, self-esteem, personal planning, interpersonal relationships, leadership, time management, positive thinking, organization, financial management, competencies, anger management, negotiation, motivation, ethics, accountability, problem resolution, life coping skills, taking feedback, and respecting diversity. Because of a lack of standard definitions for skills categories some of the skills may overlap.

[10] In Appendix 5, we assess whether we observe any systematic positive bias in reported effect size estimates that could be related to publication bias (the so-called file-drawer problem). Since we find some indication for publication bias, we discuss our measure to adjust for this and provide further robustness test of our main estimates.

[11] An alternative explanation is that programs that are already successful in the short-run are more likely to be followed-up in the long run -- i.e. sample selection, likely driven by research interest in successful programs.