LELAM-TVET4INCOME

ETH zürich DMTEC Chair of Education Systems

The impact of vocational training interventions on youth labor market outcomes. A meta-analysis.

Authors:
Andrea Ghisletta
Johanna Kemper
Jonathan Stöterau

Working Papers, No. 20, July 2021

ETH Zurich
Department of Management and Technology (MTEC)
Chair of Education Systems (CES)
Leonhardstrasse 21
8092 Zurich, Switzerland

Financed by:

Swiss Programme for Research on Global Issues for Development

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Agency for Development and Cooperation SDC

The impact of vocational training interventions on youth labor market outcomes: A meta-analysis.

Andrea Ghisletta* Johanna Kemper† Jonathan Stöterau‡

Abstract

We summarize more than 1,600 effect size estimates from 89 studies evaluating the impact of vocational training on youth labor market outcomes globally. Two-thirds of effect sizes derive from experimental evaluations. Using Hedge's g as outcome measure, we employ robust variance estimation (RVE) to account for statistically dependent effect sizes. Univariate meta-regressions show a more than two-times larger impact of training on labor market outcomes (g=0.12***) than in related meta-analyses. Multivariate meta-regressions suggest that: 1) pure in-classroom or workplace-based training is less effective than combining both; 2) training affects outcomes through human capital accumulation in high income countries, while it acts a signaling and screening device in low- and middle income countries; 3) involving non-public actors in the design or implementation of interventions yields larger effect sizes; 4) the impact of training on conditional, e.g. formal employment or wages, is lower than on unconditional outcomes; 5) intention-to-treat impact estimates are relatively lower; 6) effect sizes are smaller in male and female sub-samples than in mixed samples.

Keywords: Youth employment, vocational skills, training, ALMP, impact evaluation, meta-analysis

JEL Codes: I24, I25, I26, J13, J24, J44, J46

*University of Basel, Switzerland.
†Chair of Education Systems, Department of Management and Technology (MTEC), ETH Zurich
‡World Bank & Humboldt University Berlin, Germany.

1. Introduction

Globally, young people belong to the most disadvantaged and vulnerable groups in the labor market. The ILO (2020a) estimates that in 2019, youth of age 15-24 were three times as likely as adults to be unemployed. The youth employment challenge is especially acute in lower- and middle income countries (LMICs), where youth are more likely to be underemployed, to work in the informal sector in low-paid jobs with little access to social protection or working rights (ILO, 2020b). The Covid-19 pandemic has recently aggravated the youth employment challenge in LMICs *and* high-income countries (HICs) (ILO, 2020c). High levels of youth unemployment and underemployment are not only a waste of economic resources, but have also been connected to emigration, social unrest, or political upheaval (World Bank, 2019). Helping young people getting jobs and decent employment is hence a major policy priority among governments and international donor organizations. Among the various policy options, vocational training interventions have been a key measure to improve labor market outcomes of youth outside the formal education system. Between 2002 and 2012, the World Bank and its client governments invested nearly USD 1 billion per year in such training Blattman and Ralston (2015).

Given the policy relevance of training interventions for youth, a significant number of impact evaluations have been conducted in the recent decades to assess their effectiveness. However, systematic analysis to synthesize this evidence remains scarce. On the one hand, several meta-analyses do not assess training interventions separately but within a larger sample of active labor market programs (ALMPs) (Kluve et al., 2019; Escudero et al., 2018; Vooren et al., 2019). On the other hand, the reviews that analyze training interventions separately (Tripney and Hombrados, 2013), besides the overall impact of ALMPs (Card et al., 2018), do this very briefly and none focuses particularly on interventions targeted at youth.[1] Thus, there is yet limited systematic evidence about the specific factors that make youth-targeted training interventions in LMICs and HICs effective.

Our meta-analysis fills the gap in the literature by synthesizing recent evidence from impact evaluations of youth-targeted training interventions conducted in LMICs and HICs. We focus on interventions that provide vocational training to youth (aged 15-35) outside the formal education system with the goal to improve employment or earnings. While all included interventions provide some vocational training, they may also provide soft-skills or business skills training, as well as employment services, subsidized employment or entrepreneurship promotion. With regard to study design, we only include studies that conduct a counterfactual impact evaluation based on either experimental or quasi-experimental methods. We include peer-reviewed articles, published working papers and other grey literature.

With our analysis, we make six important contributions. First, our is the first study to focus on youth-targeted training interventions globally, which allows us to describe and analyse their effect heterogeneity and success factors in more detail. Second, for our study, we build on the sample of Kluve et al. (2019), which includes 55 studies evaluating training interventions published up to 2014. We extend their sample with studies published

[1] More details on all related studies are summarized in Table 10 in Appendix B.

between 2015 and June 2020 by adding 34 new and updating 10 studies. Resulting in a total sample of 89 studies, covering 97 different interventions, and a total of 1651 effect size estimates, of which 1054 relate to employment and 597 to earnings. Third, due to the large increase of experimental evaluations in recent years, more than half (65%) of effect size estimates in our sample derive from experimental evaluations. Fourth, we systematically analyze a broad range of factors that may drive the impact heterogeneity between impact estimates. Namely, effect size, study and participant sample characteristics, intervention context, setup and design. Fifth, the large sample allows us to conduct separate analyses for the impact of training on employment or earnings, for the sample of experimental studies and for LMICs and HICs.

Finally, employing the Standardized Mean Difference (SMD) Hedges' g as an outcome measure to make outcome variables comparable across all studies, we estimate univariate and multivariate meta-regression models using random-effects robust-variance estimation (RVE). RVE addresses the problem of statistically dependent effect sizes, which is typically ignored by standard estimation methods used for meta-analysis in economics.

Results from univariate models indicate that training interventions have statistically significant positive impact on all outcomes (SMD=0.12***). Thereby, the impact of training on employment outcomes (SMD=0.11***) is lower than on earnings (SMD=0.12***). Consistent with findings in the literature, the average impact of training is larger LMICs (SMD=0.13***) than HICs (SMD=0.10***). Using only experimental studies yields impacts close to that for HICs (SMD=0.10***). Overall, the magnitude of these unconditional averages is larger than the one reported in Kluve et al. (2019) (SMD=0.05***) and comparable to Tripney and Hombrados (2013) (e.g. SMD=0.13**).[2] All findings stay significant and are robust when controlling for reporting bias, though all decrease in magnitude. Further sample-splits show that the effect of training is highest in MICs, while its is much lower in HIcs and LICs. This finding suggests a hump-shaped effect of training when differentiated by income level. Which could imply a lack of job opportunities to make training productive at low and decreasing returns to training at high levels of economic development.

Multivariate meta-regressions accounting for impact estimate heterogeneity reveal substantial differences between HICs and LMICs. Results yield six key findings. First, we find that interventions with only in-classroom or workplace-based training are less effective than interventions that combine both.[3] This result is particularly pronounced for the sample of LMICs. Hence, assuming that classroom-based training is more adequate to teach general skills and workplace-based training to learn specific, job-relevant skills, combining both learning places may provide the right mix of skills that are immediately productive and do not out-date too quickly (Wolter and Ryan, 2011). Second, using the aggregate sample shows that less intensive and hence shorter ($<$400 hours) training interventions provide for higher impact estimates. Dis-aggregating this effect reveals that it is driven by the sample of LMICs. In contrast, in HICS, medium (400-800 hours) and longer ($>$800 hours) training interventions lead to more favorable outcomes. Together with the finding that medium-

[2]Other related studies use different outcomes measures and are therefore not reported here.

[3]Due to a relatively lower number of observations for workplace-only interventions, the results are less stable than for the in-classroom only interventions.

(1-2 years after intervention exit) to long-term ($>$ 2 years) impacts are higher in HICs and short-term ($<$ 1 year) impacts higher in LMICs, the above results could imply a human capital effect of training in HICs and a signaling and screening effect of training in LMICs, helping to integrate youth in the labor market or firms to hire a better selection of workers. Third, we find larger effect sizes for interventions that involve actors from the private sector (e.g. employer associations), NGOs or (inter-) national organizations in the design or implementation of interventions relative to government actors only. The effect on implementation design is more robust in statistical terms. Involving these actors in intervention design may increase relevancy of skills, while involving them in intervention implementation could imply better monitoring mechanisms and standards. Interestingly, the finding on implementation does not hold in the context of LMICs where it would be most binding.[4] Fourth, the impact of training on conditional outcomes, such as formal employment or wages, is lower than on unconditional outcomes, namely employment or earnings. This effect is more pronounced for the intensive than the extensive margin of employment. It is more robust for LMICs than HICs in statistical terms, which may be due due to a lower incidence of wage and formal employment opportunities in LMICs. In addition, the impact of training in LMICs is lower on earnings. This is in line with the fact that youth in LMICs are more often underemployed rather than unemployed (Fields, 2004). Hence, the impact of training runs through the intensive margin of earnings rather than the extensive margin. Fifth, intention-to-treat (ITT) impact estimates are lower than estimates based on the average treatment effect on the treated (ATT) or (local) average treatment effect ((L)ATE). The ITT is particularly relevant for policy, since it captures the effect of training in the presence of non-compliance, an all-day problem in the implementation of interventions. The finding of lower ITT-based impact estimates is particularly robust in the context of LMICs Finally, across all samples and specifications, our results show that effect sizes are generally smaller in male *and* female sub-sample analyses as opposed to samples including males and females together. However, the finding on males is more robust in terms of statistical significance.

Running multivariate regressions using WLS as a robustness test confirms our main findings. Further, restricting our multivariate analysis to the sample of experimental interventions for the aggregate sample as a robustness check confirms only two of our main findings. In contrast, using only the sample of experimental interventions in LMICs confirms most of our main findings.

The remainder of this paper is organized as follows. The second chapter provides theoretical background and associated hypotheses we will test in the empirical analysis. The third chapter describes the search, selection, and coding of primary studies, and highlights characteristics of included studies and interventions. The fourth chapter describes the sample of included studies, the fifth chapter the effect size measure and methodology used to estimate univariate and multivariate meta-regressions. The sixth chapter provides the results of the meta-analysis, while chapter seven concludes discussing the findings.

[4] Kluve et al. (2019) conclude that public-private collaborations in program implementation require a strong institutional set-up, which may not always be in place in LMICs.

2. Theoretical background and hypotheses

Understanding which factors make vocational training interventions successful is of key importance for future policy design and research. In this section, we discuss the theoretical background for the main hypotheses we aim to test in this paper. We set our discussion in the context of two mechanisms through which vocational training is expected to improve labour market outcomes. First, a *human-capital mechanism*, whereby training increases the (hard- and soft-) skills of participants and thus their productivity (Becker, 1962). Second, a *signaling and screening mechanism*, where training is used to overcome information asymmetries: either by workers as a device to signal their higher innate productivity to potential employers (Spence, 1978) or by employers as a device to screen workers for their productivity (Autor, 2001; Brunello and De Paola, 2004; Muehlemann et al., 2010). We group our main hypothesis in five dimensions in order to structure our theoretical discussion and empirical analysis: (1) effect size and study characteristics, (2) participant sample characteristics, (3) intervention context, (4) setup, and (5) design.

2.1 Effect size and study characteristics

As part of our empirical analysis, we assess a variety of factors related to the characteristics of primary studies and the effect sizes we coded.

(+/-) Outcome construct: We consider two aggregate categories of labor market outcomes: employment- and earnings- related outcomes (see Section A.4 for more details). The reason being that the impact of training on the extensive margin of employment may not be substantially different for people in treatment or control group, but on the intensive margin of employment (number of working hours or productivity), as captured by earnings-related outcomes. This effect may be particularly pronounced in LMICs (Fields, 2004).

(-) Treatment effect parameter: Primary studies typically report impact estimates either as the (Local) Average Treatment Effect ((L)ATE), the Average Treatment Effect on the Treated (ATT) or the Intention-to-Treat Effect (ITT). While the ATE measures the impact of training between treatment and control group, and the ATT for individuals in the treatment group, the ITT measures the effect of *offering* training, i.e. the effect on training on both, compliers and non-compliers.[5] Training interventions commonly observe non-compliance rates of around 10-30% (McKenzie, 2017). By accounting for non-compliance, we expect that ITT estimates are smaller in magnitude relative to estimates reported as (L)ATE or ATT.

(+/-) Evaluation design: We distinguish between experimental (RCTs), quasi-experimental designs or natural experiments (more information in Section A). As RCTs are considered the "gold standard" of evaluation methods, we expect experimental evaluation designs to provide less biased estimates.

(+/-) Publication status: Publication bias refers to the selection of empirical results, e.g. by authors or journal editors, based on the direction or magnitude of the estimated effect, its statistical significance or a combination of both Card et al. (2018). Journals may have an interest to publish high quality studies, which

[5]The LATE represents the ATE *only* for compliers.

may be more likely to report statistically significant results or surprising results that confirm/counter some prior belief or theoretical assumption (school-of-thought bias)(Andrews and Kasy, 2019). In their meta-analysis, Kluve et al. (2019) find that peer-reviewed articles report smaller effect size estimates on average. We test for publication bias by comparing estimates of studies published in peer-reviewed journals with those that are not (e.g. working papers) and by including a measure accounting for the precision of effect size estimates in our regression models (see Section 6.2 for more details on the latter).

(-) Conditional vs. unconditional outcomes: We expect unconditional outcomes (e.g. earnings, employment) to exhibit larger treatment effects than outcomes being conditioned on other primary outcomes (e.g. formal employment, wages). Unconditional outcomes generally capture more aspects of the intervention impact than conditional ones. For example, while income captures the extensive and intensive margin of employment, wages are measured conditional on being employed, which restricts the impact to either run through the number of working hours or productivity. So far, empirical reviews of ALMPs report little evidence for a differential effect between both kinds of outcomes (Kluve et al., 2019; Card et al., 2018).

(+/-) Follow-up timing: The impact of training on labour market outcomes may depend on the timing when outcomes are measured. This is of key interest for our analysis, as it reveals the effectiveness of training interventions over time. A large share of studies in the literature only collected short- or medium-term follow-up data, which may lead to a systematic underestimation of the true impact of training interventions. This "time profile" of training effectiveness may also indicate through which channel training impacts labour market outcomes: the human capital or signaling mechanism. If human capital accumulation is the key mechanism, the positive effect of training should increase with follow-up duration as productivity increases need time to materialize (Osikominu, 2016).[6] If signaling and screening are the key mechanisms, effect sizes may be larger in the short run, as training participation acts as a device for workers to signal their innate productivity or for firms to screen workers (Brunello and De Paola, 2004; Muehlemann et al., 2010).[7] While some studies find long-term effects of training interventions (e.g. Card et al. (2018); Escudero et al. (2018), others find that the initial positive impact of training disappears after a rather short time (e.g. Hirshleifer et al. (2016).

2.2 Participant sample characteristics

Vocational training interventions may be more effective only for particular groups of participants.

(+/-) Gender: There is a long-standing debate whether vocational training interventions are more or less effective for female participants (McKenzie, 2017). On the one hand, women often have lower initial human capital levels and labor force participation rates in many (developing) countries. If returns to human capital

[6] Alternatively, trainees may lower job-search intensity and may not be employed during the training period, leading to the so-called "lock-in effect" (Ham and Lalonde, 1996) in the short and hence higher effects in the long run. Of course, longer interventions exhibit deeper lock-in effects (Osikominu, 2016).

[7] Alternative explanations for larger impacts of training in the short run are that training increases trainees self-confidence, their search efforts and reduces their reservation wage in the short run (Crépon et al., 2012); Acevedo et al. (2020) argue that training may increase participants' expectations about earning possibilities, which increases their job-search intensity in the short run.

are decreasing, women should benefit relatively more from training (Chakravorty and Bedi, 2019). On the other hand, relative to men, women may be particularly (financially and non-financially) constrained to participate training interventions or to take up post-intervention employment opportunities, due to e.g. domestic work or cultural barriers, and hence miss out on potential benefits (Cho et al., 2013). The overall evidence is mixed. Some reviews find that training interventions have larger impacts on women (Tripney and Hombrados, 2013; Escudero et al., 2018; Psacharopoulos, 2018), while other studies reject this finding (McKenzie, 2017).

(+/-) Age: On the one hand, the impact of training could be larger for younger participants if returns to human capital are decreasing, assuming that older participants generally have accumulated more education (Vaillancourt, 1995; Montenegro and Patrinos, 2014). On the other hand, the impact of training could be larger for older participants, as they may have accumulated more work experience, making additional training more productive more immediately (Montenegro and Patrinos, 2014). Empirical evidence is mixed. For example, Montenegro and Patrinos (2014) show that the returns to schooling and labor market experience are complements, but also show that returns to human capital are decreasing.

2.3 Intervention context

The impact of training interventions likely depends on the overall context in which they are implemented. In our analysis, we follow Kluve et al. (2019) by focusing on differences in the impact of training across categories of country income levels.

(+/-) Country income level: The impact of training on labor market outcomes may be very different in LMICs and HICs. One the one hand, it is reasonable to expect larger impacts in LMICs. First, since baseline education levels are generally lower in LMICs, the impact of training should be larger– provided that returns to human capital are decreasing (Chakravorty and Bedi, 2019). Second, the positive effect of training as a signaling device may be more pronounced in LMICs since labor markets are generally less formalized in terms of e.g. providing certificates (Bassi and Nansamba, 2020). On the other hand, training may be more effective in HICs for two reasons. First, labor markets in HICs often offer more post-intervention employment opportunities than LMICs (Sumberg et al. (2020) refer to this as the "missing jobs crisis" in LMICs). Second, interventions in HICs may be of higher quality, more human-capital-intensive and more formalized (c.f. Kluve et al. (2012)), providing for relatively better outcomes. Kluve et al. (2019) show that ALMPs in general (including training) have larger effects in LMICs. Psacharopoulos (2018) show that returns to human capital investments are highest in countries with the lowest per capita income. However, if training delivered in HICs and LMICs is systematically different, its impact may also be different (Escudero et al., 2018).

2.4 Intervention setup

A variety of actors are typically involved in the design and implementation of interventions. We test whether intervention effectiveness differs depending on the actors involved. In our main analysis, we compare interventions that are solely run by public actors to interventions that did not involve any non-public actors, such as national (e.g. donors)- or international (e.g. World Bank) Non-Governmental Organizations (NGOs) and private-sector actors (e.g. firms or business associations).[8]

(+/-) Implementation setup: National or international NGOs and private sector actors may implement interventions more efficiently and effectively, since they are subject to stronger accountability mechanisms and must provide that training is cost effective (Becker, 1962; Acemoglu and Pischke, 1998; Wolter and Ryan, 2011), providing for better labor market outcomes.

(+) Design setup Involving national or international NGOs may improve the quality of intervention design, since they bring in their expertise, while involving private sector actors may increase relevancy of skills (Wolter and Ryan, 2011).

2.5 Intervention design

In this section, we summarize different factors that account for heterogeneity in intervention design.

(+) Complementary interventions: A significant share of interventions combine training interventions with other interventions such as employment services (e.g. job matching), subsidized employment or entrepreneurship promotion (e.g. start-up grants) (see Section 4). These complementary interventions aim to address the multiple needs and constraints faced by a heterogeneous group of beneficiaries (Kluve et al., 2012). Previous meta-analysis found larger effect sizes for ALMPs which combine a comprehensive set of interventions (Kluve et al., 2019).

(+/-) Classroom vs. workplace training:

Relative to workplace training, classroom-based training, may be more adequate to teach general skills, which are relatively more transferable and become obsolete relatively slower over time than skills taught in a workplace-based setting, where firms have an incentive to provide specific skills (Becker, 1962; Wolter and Ryan, 2011). On the other hand, classroom-based training may fail to deliver vocational skills that are demanded on the labor market (Wolter and Ryan, 2011).[9]

(+/-) Training duration: Longer and usually more human-capital-intensive training interventions (in terms of hours/days/months) do not necessarily imply larger impacts on labor market outcomes. From a human-

[8] This has two reasons. First, because the majority of interventions involve the government (see Section 4). Second, the sample size did not credibly allow for a more detailed sub-sample analysis of difference across non-public actors.

[9] The empirical literature evaluating the relative advantage of general versus vocational training suggests that, after an initial relative advantage of vocational education, general education leads to relatively better labor market outcomes in the long run (e.g. Hanushek et al. (2017)). However, the time-horizon of the studies included in the meta-analysis is too short to test this. In addition, the training evaluated in this meta-analysis is never purely general in nature.

capital perspective, longer and more intense training interventions should lead to relatively better labor market outcomes, since they have a larger impact on worker productivity. The same effect would be expected from a signaling or screening perspective, as longer and more intense interventions could signal higher worker productivity (Fitzenberger et al., 2010; Lechner et al., 2011; Kluve et al., 2012).[10] Analyzing the labor market impacts of short-term, job-search-oriented and long-term, human-capital-intensive training, Osikominu (2013) finds that while short-term training lowers unemployment spells of participants, long-term training participants have an about 40% higher exit rate from unemployment than non-participants.

(+) Non-vocational skills: Many training interventions add non-vocational skills to the curricula – in particular (i) non-cognitive (soft- or life-) skills and (ii) business or entrepreneurship skills, which are considered as particularly important for (youth) labor market outcomes in recent years (Heckman and Kautz, 2012; Ignatowski, 2017). However, evidence about their effectiveness remains mixed (Kluve et al., 2019; Adhvaryu et al., 2018; Groh et al., 2016; Acevedo et al., 2020). However, there is yet little evidence about the effectiveness interventions that explicitly combine vocational and business skills (e.g. Kluve et al. (2019).

(+) Certification: If training impacts labor market outcomes through the signaling or screening mechanism, then a stronger and more credible signal should enhance the impact of training (Spence, 1978)– especially in LIMICs, where certification of training is less common (Bassi and Nansamba, 2020). One way is to provide certificates for participation, which ideally include training content or even individual test scores.

(+/-) Monetary and non-monetary support to beneficiaries: Drop-out is a key issue for many ALMPs (Kluve et al., 2019). Financial constraints could prevent individuals to participate or complete training. Monetary support (e.g. stipend, transport allowance, salary) can alleviate these constraints. However, monetary support could also reduce training effectiveness if it lowers motivation and effort to actively participate in training and benefit from it (Blattman and Ralston, 2015). Further, cultural norms and socio-economic barriers can affect training participation - in particular for marginalized groups or women (Cho et al., 2013). Non-monetary support (e.g. child care, catering, transport) could lower these barriers. Empirical evidence is mixed (Crépon and Premand, 2019; Maitra and Mani, 2017; Kluve et al., 2019).

3. Sample construction

In the first part of this section, we describe the unit of analysis of our meta-regression that minimizes the amount of dependencies between effect sizes. In the second part, we describe the inclusion and exclusion criteria for primary studies in depth.

[10] At the same time, such interventions may aggravate potential lock-in effects, as these provide that participants search less intensively for jobs during intervention participation (Osikominu, 2013; Lechner et al., 2011).

3.1 Unit of analysis

In order to derive valid statistics in meta-analyses, effect size estimates reported by primary studies should be independent of each other.[11] However, this is rarely the case. Meta-analyses involve a multilevel data structure– with effect size estimates at the lowest and most commonly primary studies or interventions (typically implying one data source by intervention) at the highest level. Effect sizes are typically correlated within and across these different levels (Tipton, 2013).[12] Besides using estimation methods that account for such multilevel dependencies (see Section 5.2), it is important to choose a Unit of Analysis (UOA), or cluster that minimizes dependencies between the different nested levels.

In our meta-analysis, we choose the Unit of Analysis (UoA) as cohort participating in a certain intervention, hence the *intervention x cohort level*.[13] We chose this UoA, since most dependencies between effect sizes arise at the level of interventions due to multiple outcomes being measured on the same sample of individuals. Typically, one primary study nests one or more UoAs, since each study evaluates at least one intervention with one participating cohort. However, there are cases when one UoA nests two or more studies: if these evaluate the same intervention for the same cohort (e.g. Job Cops in the USA), all studies are grouped in an unique UoA. Another special case happens if one or more studies evaluate one intervention delivered to different cohorts but estimate also a pooled effect of all cohorts: in this case we produce an UoA for each cohort and a separated one for the pooled cohorts. In remainder of the paper, we refer to our chosen the *intervention x cohort level* UoA as "intervention".

3.2 Study inclusion criteria

We define study inclusion criteria along five dimensions, following the Cochrane Guidelines for Systematic Reviews (Higgins et al., 2019): Participants, Interventions, Comparison group, Outcomes and Study characteristics (PICOS).[14]

Participants: We include training interventions targeted at youth, typically aged between 14 to 35 years.
Interventions: First, interventions must provide vocational/technical skills as the main component. They can complement this with other skills training components (e.g. soft or life skills) or with additional non-training services (e.g. employment services). Second, they key goal of interventions should be to improve labor market outcomes of beneficiaries. Third, we only include training interventions outside the (formal) education system, which are typically part of ALMPs.

[11] Specifically, dependent effect sizes can lead to an under-estimation of standard errors on the average effect and thus to false conclusions about the precision of the estimate.

[12] For example, dependencies can arise if a single primary study reports estimates for several outcome constructs, or if several primary studies evaluate the impact of the same intervention for different cohorts Tipton (2013).

[13] We distinguish between programs, interventions and their components. Interventions and their components are nested within *programs*, such as "Juventud y Empleo" in the Dominican Republic or "Job Corps" in the United States. One program can comprise multiple training *interventions*. Each intervention can provide several additional *components* to beneficiaries (e.g. support services or skills). Each intervention can be delivered to multiple *cohorts* of participants (e.g. over time, sub-groups). Each intervention can be evaluated by multiple *studies* (e.g. the same Job Corps intervention cohort is evaluated by several papers).

[14] Section A in Appendix A describes the PICOS in detail.

Comparison group: Studies must measure outcomes against a comparison group that has been constructed using methods of counterfactual impact evaluation (see below). Comparison groups can receive no intervention or receive the intervention at a later point in time (e.g. pipeline or wait-list design). Hence we exclude studies that report only relative impact estimates between different interventions.
Outcomes: Studies must report at least one primary outcome measure of interest. We follow Kluve et al. (2019) and distinguish two main types outcome measures: (i) employment-related outcomes such as employment probability or labor market participation rate; (ii) earnings-related outcome constructs like earnings or wages. Outcomes can be measured conditional or unconditional on up-stream/intermediate outcomes (e.g. earnings conditional on employment). In most cases, we only coded the impact estimates reported in the main regression tables of studies but not those included in the appendix.
Study characteristics: Studies are only eligible if they employ counterfactual impact evaluation designs. Namely, using either experimental (RCTs) or namely quasi-experimental research designs (Difference-in-Differences (DiD), instrumental variable (IV) designs, matching, regression discontinuity designs (RDDs)). The results of a study need to be published either in peer-reviewed journals or other publication forms, such as working papers, reports of international organizations or of NGOs, PhD theses. If several versions (e.g. working paper and journal publication)a of a study had been published, we only coded the most recent one. We only include publications available in English, French or Spanish. We consider studies that were published between 1990 to June 2020.

3.3 Search, selection and coding of studies

Search process: We followed two approaches to find relevant literature for our meta-analysis. First, we used the existing database of youth-focused ALMP studies in Kluve et al. (2019) which includes studies until 2014. From the full sample, we identified studies that matched our inclusion criteria. We then assessed whether studies included in Kluve et al. (2019) had been published or updated since then. If updates were available, we coded these and only included the most recent published version. Second, we extended the database of studies from Kluve et al. (2019) with studies that had been published since 2015 and/or where not included in Kluve et al. (2019).[15] In our search for primary studies, we combined search terms that were used in related systematic reviews (Tripney and Hombrados, 2013; Kluve et al., 2019) with additional training-related terms that we selected based on their frequency in "key papers", i.e. papers that were included in Tripney and Hombrados (2013); Kluve et al. (2019); McKenzie (2017). We only included English search terms. Our search included 7 general literature databases (e.g. Google Scholar, Web of Science), 13 specialized databases (e.g. IDEAS/RePEc), online search engines (e.g. Google), and one specific database for grey literature (OpenGrey). While most databases allowed us to use combined keyword search strings with Boolean operators, these search strings had to be adapted for each database.[16] We complemented

[15] In our search, we followed the guidelines of the MAER-Network (Havránek et al., 2020) and the Campbell Collaboration (Collaboration et al., 2014; Kugley et al., 2017).

[16] A full list of the databases, the precise combination of keywords employed and the number of results and downloaded studies is available upon request.

the primary search by using three different methods. First, by a manual search of online repositories and relevant websites (e.g. conference websites). Second, we screened reference lists of primary studies or previous literature reviews for studies (e.g. of Tripney and Hombrados (2013); Kluve et al. (2019); McKenzie (2017)). Third, we used 59 studies that were included in Kluve et al. (2019); Tripney and Hombrados (2013) or McKenzie (2017) to search for additional studies using forward and backward citations in Google Scholar.

Search and selection of results: The primary search of databases was conducted between end January and June 2020. All studies found in the primary and complementary search were downloaded and exported in a library for further screening. Our primary search resulted in about 5,870 studies, the complementary search in about 500 additional ones. After removing duplicates, the primary and complementary search resulted in about 4,288 studies. We screened these studies based on title and abstract, which resulted in a final number of 158 potentially eligible studies. The final selection based on the full text finally resulted in 34 additional studies beyond those included in Kluve et al. (2019). Of the 113 studies included in Kluve et al. (2019), 55 matched our eligibility criteria. We identified and coded updated versions for 10 of these studies. The search and selection process resulted in a final set of 89 eligible studies.

Coding of effect sizes: For the coding of effect sizes, study and intervention characteristics, we built on the existing tools and definitions developed by Kluve et al. (2019).[17] We improved some classifications and definitions in the coding tool, based on the authors' experience from the initial systematic review. Further, we added several variables that allow us to assess our main hypotheses. The final coding tool contained a total of 151 different variables. These changes required a partial re-coding of studies obtained from Kluve et al. (2019). The papers were coded by one main coder. The coding was revised and discussed by a second coder. To avoid misunderstandings during coding and for transparency reasons, the authors used a coding manual, which was initially developed by Kluve et al. (2019) and further elaborated by the authors.

4. Sample description

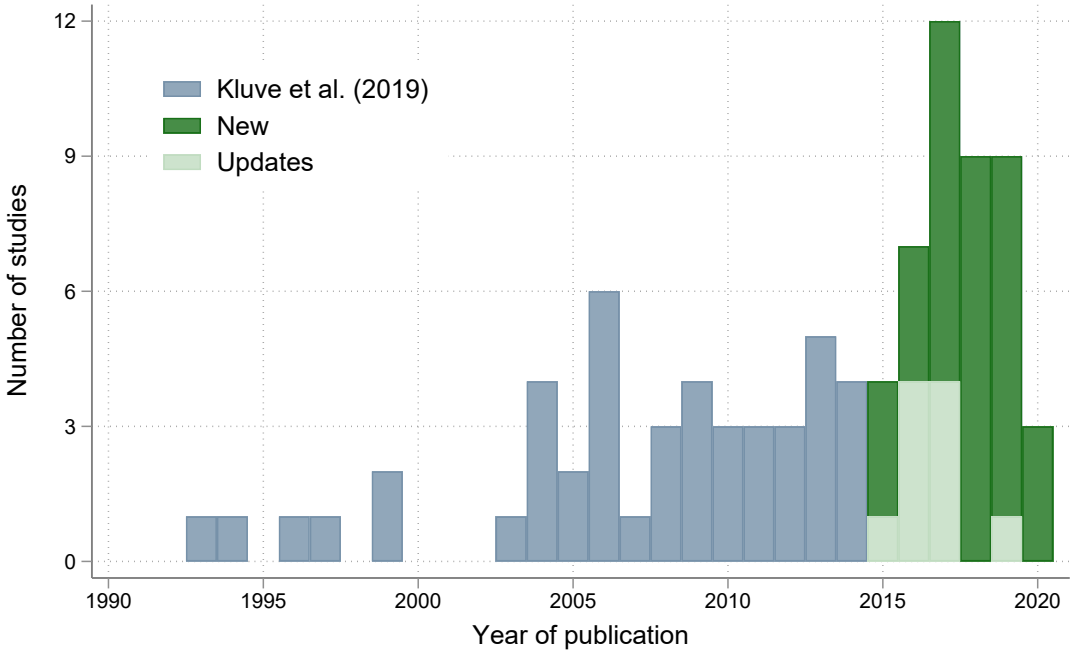
We extract 1690 treatment effects evaluating 97 different interventions in 89 studies with sufficient information to code effect sizes and their standard errors. On average, we extract 17 effect sizes per intervention, with a minimum of 1 and a maximum of 162 effect sizes. Table 1 summarizes the characteristics of studies and interventions included in our sample.

Regarding study characteristics, described in the left column in Table 1, seven key observations stand out. First, we are able to expand the base of studies evaluating training interventions included in Kluve et al. (2019) with 10 updated versions and 34 new studies. Almost half of included studies were published after 2015. The growth of evidence over recent years is also observable in Figure 1. Second, almost two-thirds (65%) of studies employ experimental evaluation designs. The share of experimental studies is particularly large in our additional sample of recently published studies and hence have not been included in any previous reviews. While about 51% of the studies published before 2015 were experimental, this share increases to

[17] Described in more detail in Kluve et al. (2017).

70% for the period 2015-2020. Third, we achieve a balanced mix between peer-reviewed academic journal publications (43%) and grey literature (e.g. working papers, technical reports). Fourth, with respect to the policy-relevant impact of training, we observe that slightly more half of studies report the effect of offering training to participants (ITT), while the majority of studies (69%) report the effect of training participation (i.e. ATE, ATET or LATE).

Figure 1: Included studies by publication year



Blue bars: studies included in Kluve et al. (2019) that match our inclusion criteria (55 studies). Light green bars ("Update"): updates of studies included in Kluve et al. (2019) (10 studies). Dark green bars ("New"): newly coded studies not included in Kluve et al. (2019) (34 studies).

Fifth, a significant share of studies (44%) estimate the longer-term effects of training ($>$ 2 years), representing more than one-third of coded effect sizes. This is a much larger share than in previous meta-analyses. For example, Card et al. (2018) sample includes only 16% of effect sizes beyond $>$ 2 years. However, evidence shows that the effect of human capital intensive interventions takes time to materialize (Card et al., 2018; McKenzie, 2017). Hence, short-term treatment effect estimates may not suffice to judge the effectiveness of training interventions. Sixth, employment-related outcomes dominate the number of studies (90%) and coded effect size estimates (64%). Most of these employment-related outcomes measure impacts on the extensive margin (employment probability) rather than the intensive margin (hours worked). At the same time, we observe that 84% of studies (additionally) provide estimates for earnings-related outcomes. This is a notable feature of the literature as it generally considered more difficult to earnings-related collect data. Finally, we see that a significant share of studies conducts sub-sample analysis by gender and/or age. This allows our meta-analysis to summarize effect size heterogeneity across potential target

groups across a large set of studies. Taken together, we believe our sample presents a broad distribution of effect sizes across outcome constructs, follow-up timing and sub-samples. This sample allows for a detailed analysis of effect size heterogeneity with respect to study characteristics.

We turn to the right column of Table 1, which summarizes the characteristics of the included 97 interventions. We observe a certain concentration in the distribution of interventions across regions and country-income groups. About a third of interventions was implemented in HICs, almost half in MICs, and a sixth in LICs. Figure 4 in the Appendix shows a detailed map of the geographic distribution of evaluated interventions. We see that interventions in North- or Latin America present about half of our sample.[18] However, it cannot be inferred whether this reflects either the predominance of training interventions or of evaluation efforts in these countries. The majority of interventions are part of national-level programs (56%) and actors from public sector organization often play a role in design (60%) and implementation (84%). Nonetheless, we also observe that actors from the private sector, national or international NGOs are involved in the design of a significant share of interventions (40%), to a lower extent in the implementation (17%). With regard to design, we clearly see that most interventions go beyond the traditional classroom-based training or having a single focus on vocational skills. First, around 50% of interventions complement training with other service components. There is a clear focus on employment services (39%), while few provide employment subsidies or entrepreneurship promotion. Second, more than three-quarters of interventions provide not only vocational skills but also soft skills (48%) or business skills (27%). Finally, only 26% of interventions deliver training only in a classroom setting. The majority (56%) of interventions combine both classroom and workplace components, reflecting the increased popularity of this approach. Many training are rather intense in terms of hours. Almost half of interventions comprise more than 800 training hours (43%) and only 27% last less than 400 hours.

Finally, a large share of interventions (72%) provide monetary or non-monetary support to beneficiaries (e.g. scholarships) – reflecting that intervention designers acknowledge the relevance of such support measures as incentives for participation (Kluve et al., 2019). At the same time, less than half of interventions report that they provide certificates to beneficiaries upon completion. This appears rather low given that certificates can be a simple, cost-effective device to improve labour market prospects (Bassi and Nansamba, 2020).

[18] Note that Central America and the Caribbean geographically belong to North America.

Table 1: Characteristics of included studies and interventions

	Studies	%Share		Interventions	%Share
Publication period:			*Country income level:*		
1990-2000	6	7	High income	36	37
2001-2005	7	9	Middle income	48	50
2006-2010	17	19	Low income	13	13
2011-2015	19	21	*Continent:*		
2015-2020	40	45	Africa	15	16
Publication status:			Asia	17	18
Peer-reviewed publication	38	43	Europe	16	17
Working paper	24	27	North America	22	23
Technical report	23	26	South America	27	28
Other publication type	4	5	*Scale of intervention:*		
Evaluation design:			National	54	56
Experimental	54	61	Regional	23	24
Quasi-experimental	35	39	Local	19	20
Estimated parameter:			*Actors- design:*		
Intention-to-treat effect (ITT)	50	56	Private/NGO involved	39	40
Others (ATE, TOT, LATE, etc.)	61	69	Governmment involved	58	60
Timing of follow-up measurement:			*Actors- implementation:*		
Short-term follow-up (<1 year)	62	70	Private/NGO involved	16	17
Medium-term follow-up (1-2 years)	35	39	Government involved	81	84
Long-term follow-up (>2 years)	39	44	*Additional service components:*		
Outcome category:			Employment services	38	39
Employment	80	90	Subsidized employment	5	5
Income	75	84	Entrepreneurship promotion	4	4
Outcome construct:			*Additional skill components:*		
Employment probability	76	85	Business skills	26	27
Participation rate	3	3	Soft skills	47	49
Hours worked	32	36	*Duration:*		
Unemployment duration	3	3	Short (<400 h)	26	27
Quality of employment	11	12	Medium (400-800 h)	29	30
Earnings	65	73	Long (>800 h)	42	43
Wage	30	34	*Training delivery:*		
Analysis sample: Gender			Classroom only	25	26
Male	36	40	Workplace only	18	19
Female	50	56	Combined	54	56
Female and male together	70	79	*Additional design elements:*		
Analysis sample: Age			(Non-)Monetary support	70	72
Younger (<25y)	70	79	Certificate	46	47
Older (≥25y)	23	26	**Total**	97	100
Total	89	100			

Note: Percentages correspond to the respective share of studies (left column) and interventions (right column) within each category. These shares do not necessarily add up to 100 since the same study or intervention can contribute to more than one characteristic (e.g. studies reporting multiple outcomes or interventions providing monetary support and a training certificate).

5. Effect sizes and meta regression models

In the first part of this chapter, we describe our outcome measure, the standardized mean difference (SMD) Hedge's g in more detail. In the second part, we explain why we employ random-effects robust variance estimation (RVE) to estimate simple and multiple meta-regressions and contrast it with weighted least squares estimation (WLS) typically used for meta-analysis in economics in the third part. In the last part of this chapter, we introduce the concept of publication bias and how we test for it.

5.1 Effect size computation

To summarize effect sizes across primary studies we compute Hedge's g for each treatment effect estimate, which is the small-sample bias-corrected version of Cohen's d (Hedges, 1981) as shown in equation (1). The particularly of using SMDs as outcome measure it that it not only allows conclusions about the statistical significance and direction of the effect of training, but also captures the magnitude of the treatment effect of training in a unit-less manner.

$$g = \frac{Y_t - Y_c}{S_p} * \ (1 - \frac{3}{4 * df - 1}) \tag{1}$$

Where Y_t and Y_c are the mean outcomes in the treatment and control group, respectively. The difference between Y_t and Y_c captures the treatment effect of training. Degrees-of-freedom are defined as $df = n_t + n_c - 2$, where n_t, n_c represent the sample sizes of treatment and the control group. S_p is the pooled standard deviation.[19]

Hedge's g thus requires information about the mean and standard deviation of the outcome variable for treatment, control and pooled sample. Since only a few papers report these metrics, we often approximated them through the following formula provided by Borenstein et al. (2009):

$$S_p = SE * \sqrt{\frac{n_c * n_t}{n_c + n_t}} \tag{2}$$

where SE is the standard error of the regression coefficient of the outcome variable.

5.2 Random-effects robust variance estimation (RVE)

As mentioned in 3.1, one common problem in meta-analysis are statistically dependent effect sizes (i.e. nonzero covariances between effect size pairs). To be able to derive valid statistics, effect sizes must be independent of each other. Many meta-analyses in economics typically assume statistical independence in the residuals of the statistically dependent effect sizes after controlling for characteristics at the level where

[19] More precisely, $S_p = \sqrt{\frac{((n_c-1)*S_c^2+(n_t-1)*S_t^2)}{(n_t+n_c-2)}}$; where S_t, S_c represent the standard deviation of outcomes in the treatment and control group before treatment.

dependencies arise– typically the study or, as in our case, the intervention-level– by means of multivariate meta-regression. However, this requires information about the covariance structure of effect size estimates within the at the level where dependencies arise, which is rarely reported in primary studies (Tipton, 2013). Failing to account for this covariance structure may lead to underestimating the standard errors in meta regression and thus to over-proportionally high Type I errors (López-López et al., 2017). As the level where dependencies arise in our analysis is the intervention level, we refer to this level in what follows. Robust variance estimation (RVE) for meta-regression deals with statistically dependent effect sizes in that it approximates their variance–covariance matrix. This can be done by using the cross products of residuals at the intervention level to estimate a crude average of the covariances in the variance–covariance matrix. Averaging is done since estimating each element of the variance–covariance matrix between effect size pairs would require more data than normally available. For this procedure to work, the must be independent of one another, while effect size estimates within interventions can be statistically dependent (Hedges et al., 2010).

Even though RVE does not make any requirement on weights used for estimation, Hedges et al. (2010) argue that approximately inverse variance weights are most efficient. They offer two methods for deriving simple, approximately inverse variance weights for RVE: correlated and hierarchical effects weights. In the hierarchical effects case, dependencies arise due to multiple studies studying independent samples of the same intervention. In the correlated effects case, dependencies arise due to multiple outcomes measured on the same sample of individuals (Tipton, 2013). In the context of our meta-analysis, the *correlated effects model* seems the most adequate, since most of the time multiple outcomes are measured on the same sample of individuals. Instead of using inverse variance weights, which would give more weight to clusters with a large number of estimates, the correlated effects model employs weights w_{ij} for each effect size i within each intervention j that accounts for variation within and between clusters and the number of estimates per cluster.[20]

Since the degree of heterogeneity across primary studies in our sample is quite high, e.g. in terms of participant samples, geographic contexts, it is hard to argue that there is one common true effect across all included clusters. Hence, the random effects model seems the most adequate for our analysis, as opposed to a fixed affect model assuming the source of variance to be exclusively due to measurement error within each intervention. Hence, the random effect RVE model explicitly models between-intervention heterogeneity *and* within-intervention measurement error:

$$g_{ij} = \alpha + x'_{ij}\beta + z'_j\gamma + \nu_j + \varepsilon_{ij} \tag{3}$$

where g_{ij} represents the i-th effect size estimate within intervention j. α is the mean of the distribution of true effects across interventions. x'_{ij} is a vector of covariates that vary between and within the intervention, z'_j

[20]More precisely, $w_{ij} = \frac{1}{\{(V_{\bullet j}+\tau^2)[1+(k_j-1)\rho]\}}$. Where $V_{\bullet j}$ the mean of the within-intervention sampling variances for each intervention j, τ^2 is the estimate of the between-intervention variance component (in a random effects model), k_j is the number of effect sizes within each intervention j, and ρ is a constant measuring within-intervention correlation between effect sizes.

is a vector of covariates that vary only between interventions. $\nu_j \sim N(0, \tau^2)$ is the intervention-level random effect and $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$ is the residual of the i-th effect size estimate within intervention j. Here, τ^2 is the between-intervention variance in true effects, which is unknown and has to be estimated from the data using the method of moments. We use the weights w_{ij} to account for the correlation of estimates within interventions. Thereby, we use the default value of $\rho = 0.8$ (Tipton, 2013).

5.3 Weighted least squares (WLS)

While our main analysis relies on the RVE approach, we acknowledge different views on the appropriate method that imply a set of different assumptions. To test the sensitivity of our results using RVE, we re-estimate our main regression models by weighted least squares (WLS) advocated by Stanley and Doucouliagos (2012, 2017). In contrast to RVE, WLS does not account for statistically dependent effect sizes within our unit of observation, namely interventions. López-López et al. (2017) show that "standard methods" that ignore dependencies between effect sizes, including the kind of WLS we estimate as robustness check, underestimate the standard errors of meta-regression coefficients and provide for highly inflated type I error rates. We weight the WLS regressions by the inverse of the number of effect size observations contributed by each intervention. This way, we guarantee that clusters with large numbers of effect size estimates are not over-proportionally weighted. As in the RVE model, we cluster standard errors in the WLS model at the level of our unit of observation, interventions.

5.4 Publication bias

Selective reporting of findings is a key challenge for meta-analyses. Publication bias refers to the selection of results, e.g. by authors or journal editors, based on the direction of the estimated effect, its statistical significance or combination of both (Card et al., 2018). A common method to inspect publication bias are funnel plots. Funnel plots show the relationship between the effect size and the precision of the effect size estimate (the inverse standard error). In these plots, less precise estimates – plotted lower down the y-axis – are typically scattered more widely around the true effect. In the absence of publication bias, the standard error of an estimate should be orthogonal to the reported effect sizes and the plot should be symmetric around the true effect size. In presence of a positive publication bias, one would expect a skew toward the right for less precise estimates (funnel asymmetry) (Stanley and Doucouliagos, 2012).

We also test for publication bias by means of regressions, thereby following Stanley and Doucouliagos (2012). In a first step, we include the estimates' standard error SE_{ij} in the univariate RVE meta-regression model (equation 4) to conduct the "Funnel Asymmetry Test" (FAT), testing the presence of publication bias (null $H_0^I : \gamma_{FAT} = 0$). In the same estimation, we conduct the "Precision Effect Test" (PET), testing for the presence of a genuine effect beyond publication bias (null $H_0^{II} : \alpha_{PET} = 0$).

$$g_{ij} = \alpha_{PET} + SE_{ij}\gamma_{FAT} + \nu_j + \varepsilon_{ij} \tag{4}$$

Following Stanley and Doucouliagos (2012), we then estimate the Precision-Effect-Estimate with Standard Error (PEESE) model to get unbiased coefficients in case of rejected FAT-PET tests. The PEESE includes the estimates' variance SE_{ij}^2 in the univariate RVE meta-regression model (equation 5).

$$g_{ij} = \alpha_{PEESE} + SE_{ij}^2\gamma + \nu_j + \varepsilon_{ij} \tag{5}$$

6. Results

We extract 1690 treatment effects from 89 studies with sufficient statistical information to code effect sizes (Hedges' g) and their standard errors. These relate to impact estimates from 76 interventions and 97 intervention x cohort clusters (see Section 3 for details). On average, we extract 17 effect sizes per cluster, with a minimum of 1 and a maximum of 162 effect sizes for a single cluster. We exclude outliers with an Hedges' g below -1 or above 1 or an inverse standard error below 1 or above 100, following common procedure in meta-analyses (e.g. Kluve et al. (2019); Askarov and Doucouliagos (2020)).[21] This trimming reduces the sample to 1651 effect size estimates and results in slightly smaller average effect sizes (see Table 10 in B). In the following, we present results of univariate and multivariate meta-regressions.

6.1 Univariate meta-regressions

Table 2 displays estimates from univariate RVE random-effects meta-regression.[22] The results show that the overall impact of training on labor market outcomes is positive and statistically significant at the 1% level ($g = 0.116$ SD). The mean effect size is larger for earnings-related outcomes ($g = 0.120$) than for

Table 2: Univariate meta-regression

	Pooled	Employment	Earnings	LMICs	HICs	RCTs
Average effect	0.116***	0.107***	0.120***	0.125***	0.104***	0.104***
	(0.01)	(0.015)	(0.020)	(0.020)	(0.025)	(0.020)
95% Confidence Interval	0.086,0.147	0.078,0.137	0.081,0.159	0.086,0.163	0.055,0.152	0.064,0.144
Estimates	1651	1054	597	1002	649	1075
Interventions	97	96	77	61	36	42
Reports	89	80	75	50	39	54

Note: This table shows estimation results of the effect of vocational training on pooled labor market outcomes, sample-splits for employment and earning outcomes, LMICs and HICs, as well as RCTs separately. Models were estimated using Robust Variance Estimation (RVE), correlated random effects models (Tipton, 2013), setting Rho at 0.8. Standard errors are clustered at the intervention level. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

[21] Outliers are a common issue in meta-analyses, typically arising from reporting errors or incompatible effect size measures (Askarov and Doucouliagos, 2020).

[22] Summary effect sizes from the RVE random-effects meta-regressions are slightly larger than raw means in the lower panel of Appendix Table 10, which do not account for correlations between effect sizes. However, our main results remain the same in both analyses.

employment outcomes ($g = 0.107$ SD), which is in line with our hypotheses that earnings is a generally a more sensitive outcome measure. Furthermore, effect sizes are larger on average in LMICs ($g = 0.125$) than in HICS ($g = 0.104$). Finally, effect sizes based on experimental evaluation methods (RCTs) are lower on average but still highly statistically significant ($g = 0.104$).

Compared to previous meta-analysis that include skills training interventions, these average effect sizes are significantly larger. Kluve et al. (2019) report an average of $g = 0.05(C.I. = 0.02 - 0.07)$ for vocational training, while Card et al. (2018) report an average of around $g = 0.06$ in the medium- and long-term. Only Tripney and Hombrados (2013) report comparable estimates (SMD=0.13**).[23]

Table 3: Univariate Meta-Regression by income level

	Pooled	Employment	Earnings	RCTs
Panel I: HICs				
Average effect	0.104***	0.099***	0.071*	0.074
	(0.025)	(0.021)	(0.041)	(0.057)
95% Confidence Interval	0.055,0.152	0.058,0.140	-0.010,0.151	-0.038,0.185
Estimates	649	376	273	311
Interventions	36	36	18	8
Studies	39	39	38	35
Panel II: LMICs				
Average effect	0.125***	0.113***	0.134***	0.113***
	(0.020)	(0.021)	(0.023)	(0.021)
95% Confidence Interval	0.086,0.163	0.072,0.154	0.090,0.178	0.072,0.154
Estimates	1002	678	324	764
Interventions	61	60	59	34
Studies	57	55	54	37
Panel III: MICs				
Average effect	0.132***	0.117***	0.156***	0.121***
	(0.023)	(0.024)	(0.028)	(0.025)
95% Confidence Interval	0.087,0.177	0.070,0.164	0.101,0.211	0.073,0.170
Estimates	899	611	288	701
Interventions	48	47	46	25
Studies	46	44	43	28
Panel IV: LICs				
Average effect	0.094***	0.097**	0.067***	0.089**
	(0.033)	(0.038)	(0.019)	(0.038)
95% Confidence Interval	0.029,0.158	0.022,0.172	0.029,0.105	0.013,0.164
Estimates	103	67	36	63
Interventions	13	13	13	9
Studies	11	11	11	9

Note: This table shows estimation results of the effect of vocational training on pooled labor market outcomes, for employment and earning outcomes, as well as RCTs separately. Panels I-IV show separate regressions for HICs, LMICs, MICs and LICs. Models were estimated using random effects Robust Variance Estimation (RVE) (Tipton, 2013), setting Rho at 0.8 and using interventions as unit of analysis. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

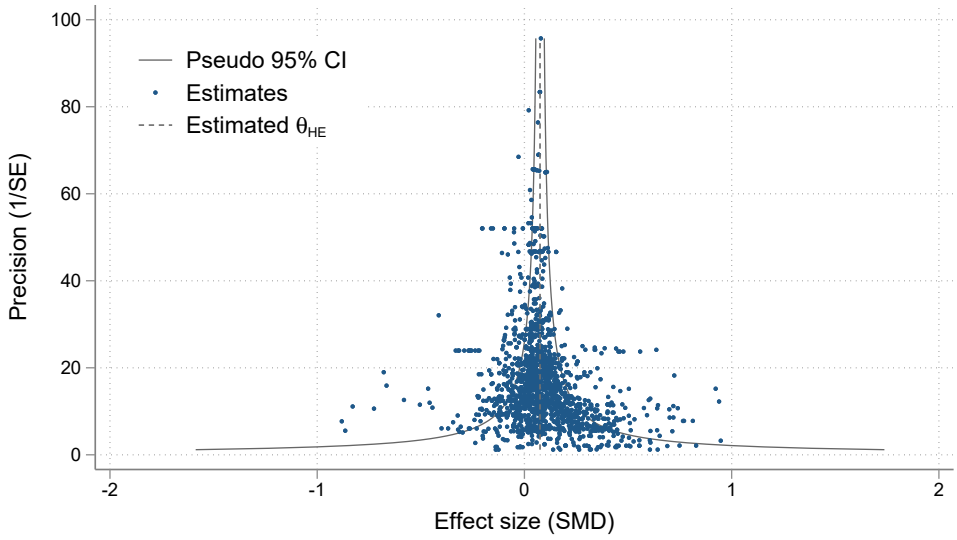
[23] Other related studies use different outcomes measures and are therefore not reported here.

We further conduct univariate regressions for sample-splits by income level displayed in Table 3. Namely, for HICs, and for MICs and LICs taken together and separately. Across all specifications, the effect of training is highest in MICs, in particular for earnings outcomes, while its is much lower in HIcs and LICs. This finding suggests a hump-shaped effect of training when differentiated by income level. Which could imply a lack of job opportunities to make training productive at lower– and decreasing returns to training at higher levels of economic development.

6.2 Assessing and accounting for publication bias

Figure 2 provides a funnel plot for the pooled sample of effect size estimates in our sample. The dashed line refers to the unweighted mean effect size ($g = 0.094$, see Table 10 in Appendix B). For reference, the figure includes the boundaries of the 5% statistical significance level. As expected, more precise estimates are generally more closely centered around the estimated true mean effect size. Figure 2 suggests that the distribution of reported effect size estimates is fairly symmetrical, with the cloud of estimates being slightly skewed a towards the lower right area of the plot and suggesting modest reporting bias mostly due to low-precision studies reporting a positive effect of training.

Figure 2: Funnel Plot



Note: Figure 2 provides a funnel plot for the pooled sample of effect size estimates in our sample. The dashed line refers to the unweighted mean effect size ($g = 0.094$). The figure includes the boundaries of the 5% statistical significance level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

Table 4 shows the results of the regression-based test for publication bias for the pooled sample (see Section 6.2 for more details). Results from the FAT-PET test suggest the presence of publication bias in the pooled sample, but also of a genuine effect beyond this reporting bias. Following Stanley and Doucouliagos (2012), we hence estimate the PEESE model to account for reporting bias.

Table 4: Univariate, FAT-PET and PEESE meta-regressions

	Univariate	FAT-PET	PEESE
Average effect	0.116***	0.049**	0.105***
	(0.01)	(0.021)	(0.016)
SE	-	0.872***	-
		(0.208)	
SE^2	-	-	1.257***
			(0.486)
Estimates	1651	1651	1651
Interventions	97	97	97
Studies	89	89	89

Note: This table shows estimation results of the effect of vocational training on pooled labor market outcomes using univariate, FAT-PET and PEESE models accounting. Models were estimated using random effects Robust Variance Estimation (RVE) (Tipton, 2013), setting Rho at 0.8 and using interventions as unit of analysis. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

Table 5 shows the respective PEESE meta-regression results for the univariate sub-samples in Table 2. The resulting estimated mean effect size are only slightly smaller than in the baseline model that does not account for publication bias (e.g. from 0.116 SD to 0.105 SD in the pooled sample).[24]

Table 5: PEESE meta-regressions

	Pooled	Employment	Earnings	LMIC	HIC	RCT
Average effect	0.105***	0.097***	0.103***	0.112***	0.096***	0.088***
	(0.016)	(0.015)	(0.024)	(0.021)	(0.026)	(0.022)
Estimates	1651	1054	597	1002	649	1075
Interventions	97	96	77	61	36	42
Reports	89	80	75	50	39	54

Note: This table shows estimation results of the effect of vocational training on pooled labor market outcomes, sample-splits for employment and earning outcomes, LMICs and HICs, as well as RCTs separately. Models were estimated using random effects Robust Variance Estimation (RVE) (Tipton, 2013), setting Rho at 0.8 and using interventions as unit of analysis. All models account for reporting bias using SE_{ij}^2, the sampling variance of effect size estimate i in intervention j. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

[24]Table 13 in the Appendix shows the PEESE results for sample slits for HICs, and for MICs and LICs separately and combined.

6.3 Heterogeneity in average effect sizes

Table 6 provides an initial assessment of effect size heterogeneity across different types of interventions and study characteristics. These were generated using univariate RVE meta-regressions controlling for publication bias by including estimates' sampling variance. One key observation is that average effect sizes are positive and statistically significant for almost all sub-samples. Insignificant summary effects are mostly due to the small number of interventions. One exception is the small and insignificant average effect estimated for outcomes that are measured conditionally on some other primary outcome (e.g. the impact on earnings conditional on being employed). This effect is discussed later in section 6.4 and is tied to rather inelastic intensive margins as well to the possibility for selection of low-earning workers into the labor force. Another exception is the small and insignificant average effect of classroom-based training interventions ($g = 0.032$). This result links directly to the common critique against such more traditional interventions.

Table 6: Bivariate meta-analysis: All outcomes controlling for publication bias, RVE correlated effects model

Variable	Effect	95% CI	#Est. / #Int.	Variable	Effect	95% CI	#Est. / #Int.
Total				**Participant characteristics**			
Average effect	.105	[.073, .137]	1651 / 97	Males	.110	[.037, .182]	314 / 32
Outcome construct				Females	.109	[.042, .177]	465 / 45
Employment probability	.110	[.074, .146]	771 / 92	Males and females	.114	[.078, .151]	872 / 78
Participation rate	.014	[-.157, .186]	18 / 2	Younger Participants (<25y)	.100	[.060, .139]	1305 / 71
Hours worked	.045	[.002, .087]	152 / 42	Older participants (>25y)	.103	[.054, .151]	346 / 29
Unemployment duration	.018	[-.679, .715]	11 / 3	**Training place**			
Employment quality	.076	[.006, .147]	102 / 14	Classroom and workplace	.130	[.082, .177]	933 / 54
Earnings	.011	[.058, .161]	462 / 68	Classroom only	.032	[-.015, .079]	549 / 25
Wages	.052	[.002, .103]	135 / 33	Workplace only	.128	[.043, .213]	169 / 18
Outcome characteristics				**Training duration**			
Dependent employment	.098	[.063, .134]	904 / 77	Hours: Short (<400h)	.091	[.004, .179]	364 / 17
Self-employment	.045	[-.011, .100]	103 / 20	Hours: Medium (400-800h)	.086	[.009, .163]	310 / 29
Self- and dependent employment	.115	[.073, .157]	644 / 50	Hours: Long (>800h)	.130	[.083, .178]	928 / 42
Formal employment context	.078	[.027, .129]	676 / 57	**Training surplus**			
Informal employment context	.039	[-.444, .541]	49 / 4	No certification	.059	[.025, .094]	645 / 51
Formal and informal	.121	[.083, .160]	926 / 67	Certification	.152	[.098, .207]	1006 / 46
Study characteristics				No business skills	.093	[.063, .123]	1256 / 72
Unconditional outcome	.112	[.078, .145]	1386 / 92	Business skills	.126	[.042, .209]	395 / 26
Conditional outcome	.020	[-.031, .071]	212 / 33	No soft skills	.100	[.063, .136]	808 / 50
Not Intention-to-treat effect	.119	[.080, .158]	830 / 75	Soft skills	.109	[.052, .166]	843 / 47
Intention-to-treat effect	.058	[.032, .083]	821 / 47	No extra services[25]	.069	[.034, .104]	770 / 39
Peer-reviewed	.085	[.038, .132]	594 / 33	Extra services	.132	[.082, .182]	881 / 58
Working paper	.136	[.086, .186]	466 / 32	No participation incentives	.065	[.018, .111]	313 / 28
Eval./Techn. report	.083	[.017, .149]	413 / 37	Participation incentives	.121	[.080, .161]	1338 / 70
Other publication	.042	[-.447, .531]	178 / 3	**Actors characteristics**			
Evaluation characteristics				Design: no private or ngo	.066	[.033, .099]	762 / 39
Design experimental	.088	[.042, .133]	1075 / 42	Design: private or ngo	.139	[.088, .191]	889 / 58
Natural experiment	.008	[-.055, .072]	181 / 10	Implem.: no private or ngo	.044	[-.002, .090]	232 / 16
Quasi experiment	.118	[.055, .181]	395 / 46	Implem.: private or ngo	.117	[.080, .155]	1419 / 81
Low-/middle income country	.112	[.070, .154]	1002 / 61				
High income country	.096	[.042, .149]	649 / 36				
Short-term follow-up	.080	[.029, .131]	758 / 75				
Medium-term follow-up	.105	[.054, .156]	337 / 40				
Long-term follow-up	.103	[.047, .159]	531 / 29				

Note: This table shows estimation results of the effect of vocational training on different study or intervention characteristics. Models were estimated using random effects Robust Variance Estimation (RVE) (Tipton, 2013), setting Rho at 0.8 and using interventions as unit of analysis. All models account for reporting bias using SE_{ij}^2, the sampling variance of effect size estimate i in intervention j. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

6.4 Multivariate meta-regression

In comparison to univariate meta-regression, multivariate meta-regression allows to control for other factors that may drive heterogeneity in effect size estimates. This is important since moderator variables may be

[25] In the text, "extra services" denote the provision of employment services, subsidized employment and/or entrepreneurship promotion.

jointly correlated with effect size magnitude, in which case unconditional averages from the univariate models may be misleading. This would not be surprising, since the impact of any given intervention is expected to depend on the interaction of design features with beneficiaries' characteristics and country context. In our multivariate meta-regressions, we assess to what extent factors grouped in the following five dimensions introduced in Section 2 account for effect size heterogeneity: (1) effect size and study characteristics; (2) participant sample characteristics; (3) context; (4) intervention setup; (5) intervention design.

We start by assessing the aggregate sample across all interventions and all outcome measures and sample-splits by employment and earning outcomes in Table 7. Table 8 summarizes the results for HICs and LMICs.

We start by looking at the relevance of study and effect size characteristics. In the specifications I-III we observe that training has a slightly larger average impact on earnings than on employment outcomes. The sample-split by income level (specifications X-XV) reveals that the larger impact on earnings is driven by the sample of LMICs. Particularly in LMICs, many people cannot afford not to work, even for short periods (Fields, 2004). Therefore, it seems intuitive that the impact of training in LMICs is more visible on the intensive rather than the extensive margin of employment. This finding is broadly in line with that of Kluve et al. (2019). Coefficient estimated from ITT-estimates are smaller, on average, than estimates from ATE/ATT or LATE models. This finding is driven by estimates from LMICs, suggesting that low take-up (or high drop-out) rates are a particular challenge in developing countries. In fact, the coefficient estimate for LMICs is consistent with drop-out rates of 15-30% reported by McKenzie (2017).

We do not find that effect size estimates from experimental studies are significantly lower than from quasi-experimental on average. This is in contrast to results from the univariate RVE model 6.3, suggesting that the use of experimental evaluation methods correlates with other study- or intervention-related characteristics. In particular, Table 8 shows the coefficient for RCT-based studies is significantly negative in the HIC sub-sample . Our results suggest that contradictory findings in previous meta-analysis of related literature may be driven by difference in studies from different country-income levels.

For most specifications we find that peer-reviewed articles report effect sizes of (marginally) smaller magnitude, except for the sample of LMICS. This confirms results of related meta-analyses (Kluve et al., 2019; Card et al., 2018).However, the effect is statistically significant only for the sub-sample of HICs, which could be either an indication for a higher publication bias in the literature on HICs or the effect of a better quality of published studies on HICs.

Consistent across all estimations, effect sizes are also smaller in magnitude if the outcome variable is measured conditional on some other outcome (e.g. formal employment or wages conditional on being employed). The estimates are statistically significant in most specifications, except for the sub-sample for employment and HICs. On the one hand, this implies that training has a relatively lower impact on wages

[25] In particular, Card et al. (2018); Evans and Yuan (2020); Escudero et al. (2018) find not differences between experimental and quasi-experimental studies, whereas Kluve et al. (2019) report slightly smaller effect size estimates.

than earnings, while the difference between the employment probability and conditional outcomes such as formal employment or hours worked is less pronounced. On the other hand, it implies that the lower impact of training on conditional outcomes is mainly driven by the sub-sample of LMICs. This latter finding could be one consequence of a relatively lower impact of training on wage and formal employment due to a lower incidence of wage and formal employment opportunities in LMICs. In addition, this may also provide that these outcomes are less likely reported by studies in LMICs and employment less often implying formal employment.

Secondly, we assess differences in the impact of training according to when and for which sub-sample it is measured.

Consistent with Kluve et al. (2019); Card et al. (2018), the impact of training for the aggregate sample is higher in the medium (1-2 years after intervention exit) and long term ($\geq$ 2 years) relative to the short term ($\leq$ 1 year). The sample-split by outcome types suggests higher effects for employment outcomes in the medium- and longer-term and higher short-term impacts for earnings outcomes. However, these estimates are not statistically significant. While Card et al. (2018) find higher long-term impacts of training programs on the employment probability, our results show that it may be harder to impact earnings in the long-term. Differentiating by country income level, the short-term impact of training seems higher in LMICs, while medium- to long-term impacts of training are higher in HICs. Even though these estimates are not statistically significant, this could point at training having a stronger signalling/screening impact in LMICs and a stronger human capital effect in HICs.

Across all samples and specifications, our results show that effect sizes are generally smaller in male *and* female sub-sample analyses as opposed to samples including males and females together.[26] The result for males is statistically significant for the aggregate sample, employment outcomes and LMICs sub-sample, while it is never statistically significant for females. Results indicate that the effect of training is lower for younger participants ($<$ 25 years). This result speaks in favor of returns to training and labor market experience being complements Montenegro and Patrinos (2014), making additional training more productive more immediately.

We turn to assess heterogeneity across the context and setup of interventions. Surprisingly and in contrast to results from univariate regressions, multivariate regressions suggest that training interventions are slightly more effective in HICs. The sample-split by outcome measure shows that this effect is driven by the impact of training on employment, while the impact on earnings larger in LMICs (as already observed previously). However, none of these findings is statistically significant. This finding is also in contrast to the ALMP meta-analysis of Kluve et al. (2019), who find lower effects for HICs. This could have two reasons: either

[26]Note that the negative directions for both genders with respect to the pooled-genders result is due to the sample composition. Indeed, for 46 interventions we don't have gender disaggregation, but only the pooled effect, and for 15 interventions we have results for only one gender. For 4 interventions we only have the gender-disaggregated results without a pooled result, while for 31 interventions we have the gender disaggregation and the pooled effect. The gender comparison is thus influenced by the lack of information on some interventions. Estimating the variable-complete regression on the subsample of 35 interventions with a disaggregation for both genders, we obtain a slightly negative effect for males and a slightly positive for females.

the results in Kluve et al. (2019) were mainly driven by interventions other than training; or our sample of more recently published studies presents an higher impact in HICs than in LMICs. Further, we find larger effect sizes for interventions that involved actors from the private sector (e.g. employer associations), NGOs or (inter-) national organizations in the design or implementation of interventions. Except for the effect of involving private sector actors, NGOs or (inter-) national organizations in the implementation of interventions in LMICs, where the result is negative, though not statistically significant. The finding on intervention design could imply that having private sector actors, NGOs or (inter-) national organizations on board increases relevancy of skills in training interventions. While the finding on implementation could be the consequence of having better monitoring mechanisms and standards involved if NGOs or international organizations participate intervention implementation. Interestingly, this does not hold in the context of LMICs where it would be most binding. This result supports similar findings by Kluve et al. (2019), who conclude that public-private collaborations require a strong institutional set-up which may not always be the case in LMICs.

We now turn to the relevance of specific intervention design features – delivery mechanism, intensity, and additional components. First, we find no clear evidence in the aggregate sample that interventions which provide additional labor market services– namely employment services, subsidized employment and/or entrepreneurship promotion– to beneficiaries perform significantly better. This is surprising, given that it is a common finding in related meta-analysis of ALMPs (Escudero et al., 2018; Kluve et al., 2019). It is also commonly highlighted by researchers and policymakers as a cost-efficient means to improve the impact of interventions (e.g. McKenzie (2017)). One reason for our finding may be that the interventions in our sample mainly deliver training rather than multiple service components to equal parts. Second, with regard to the delivery mechanism, we find evidence for the common critique that interventions with only in-classroom or workplace-based training are less effective than interventions that combine both. The results on in-classroom training interventions are statistically significant for the aggregate sample and sub-samples for earnings and employment. Though less robust in terms of direction of the effect and statistical significance, the same tendency holds for training that is only workplace-based, especially when looking at the sub-sample of earnings outcomes. One reason for the less robust finding for workplace-only interventions could be that the number of interventions with this kind of training in our sample is two-thirds lower than for classroom-base interventions. Third, the results of the aggregate sample show that more intensive and hence longer training interventions provide for lower impact estimates of training. However, differentiating the effect of training intensity by income setting reveals that in HICS, medium (400-800 hours) and longer ($>$800 hours) training interventions lead to more favorable outcomes than shorter ($<$400 hours) interventions. This result is statistically significant. In contrast, in LMICs, more intensive and longer training interventions lead to lower impact estimates. Taking the findings for HICs and LMICs together, suggest the importance of human capital effect of training in HICs and of signalling and screening in LMICs. Especially when complemented with the finding of higher medium- and long-term effects of training HICs and short-term higher effects in LMICs.

Finally, we turn to the relevance of adding further components to the core vocational training module, e.g. teaching soft or business skills or providing support and certificates. Results for the aggregate sample suggest that interventions including a soft-skills training component perform slightly better on average. This is not the case for interventions that include a business training component, where results are mixed. Dis-aggregating the effects by outcome type and income level reveals that soft-skills training provides for slightly higher impact estimates in all sub-samples, while business training provides for slightly higher impact estimates for earning outcomes and in LMICs and lower estimates for employment outcomes and in HICs.[27] We find that training interventions providing certificates correlate with larger average effect size estimates. This effect is most pronounced in the LMICs sample. This supports our assumption made in Section 2, that certificates may have the largest impact in LMICS where much more people may not have any certificate proving skills beyond the primary school level. Providing monetary or non-monetary support to beneficiaries, provides for lower impact estimates. This effect is driven by the sub-samples for employment outcomes and LMICs, where the latter could indicate a lower impact of training due to a lower motivation to participate in training through e.g. higher opportunity costs to participate, cultural norms or socio-economic barriers.

[27] Note that this result may be driven by interventions that aim to enhance incomes from self-employment rather than employment per se, which are mainly implemented in LMICs.

Table 7: All outcomes, Employment & Earnings outcomes, RVE correlated effects model, controlling for publication bias

	I	II	III	IV	V	VI	VII	VIII	IX
	Aggregate outcomes			**Employment**			**Earnings**		
SMD's sampling variance	1.088**	1.050**	0.984**	1.125***	1.062***	0.915**	1.855	1.874	1.91
	(0.446)	(0.410)	(0.406)	(0.405)	(0.386)	(0.373)	(1.692)	(1.540)	(1.678)
Employment construct	ref.	ref.	ref.	-	-	-	-	-	-
Income construct	0.021	0.018	0.021						
	(0.022)	(0.021)	(0.019)						
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.066**	-0.042	-0.043	-0.040	-0.019	-0.031	-0.047	-0.033	-0.017
	(0.033)	(0.032)	(0.028)	(0.031)	(0.031)	(0.030)	(0.038)	(0.037)	(0.027)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	0.001	0.007	0.016	-0.015	-0.015	0.002	-0.027	-0.007	-0.012
	(0.039)	(0.039)	(0.038)	(0.038)	(0.037)	(0.036)	(0.051)	(0.051)	(0.056)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.018	-0.014	-0.019	-0.019	-0.015	-0.024	-0.014	-0.002	0.005
	(0.025)	(0.025)	(0.026)	(0.025)	(0.028)	(0.029)	(0.035)	(0.034)	(0.032)
Unconditional outcome	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Conditional outcome	-0.063*	-0.080**	-0.090***	-0.018	-0.038	-0.057	-0.086***	-0.107***	-0.120***
	(0.033)	(0.034)	(0.034)	(0.037)	(0.036)	(0.037)	(0.029)	(0.029)	(0.031)
Measurement: <1 year from program end	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)	0.009	0.009	0.006	0.008	0.006	-0.005	-0.006	-0.022	-0.017
	(0.030)	(0.031)	(0.031)	(0.027)	(0.029)	(0.029)	(0.044)	(0.042)	(0.037)
Long-term follow-up(>2 years)	0.012	0.013	0.009	0.031	0.036	0.032	-0.037	-0.047	-0.055*
	(0.029)	(0.032)	(0.031)	(0.027)	(0.030)	(0.029)	(0.030)	(0.034)	(0.032)
Both male and female participants	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Measurement: Male participants	-0.037	-0.044*	-0.058**	-0.049*	-0.055**	-0.071***	-0.037	-0.033	-0.044
	(0.026)	(0.027)	(0.026)	(0.025)	(0.027)	(0.027)	(0.035)	(0.035)	(0.032)
Measurement: Female participants	-0.028	-0.026	-0.032	-0.037	-0.032	-0.034	-0.009	-0.019	-0.033
	(0.031)	(0.032)	(0.032)	(0.034)	(0.035)	(0.034)	(0.035)	(0.036)	(0.038)
Older participants (>25y)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)	-0.025	-0.037	-0.038	-0.018	-0.029	-0.029	-0.002	-0.013	-0.011
	(0.028)	(0.029)	(0.029)	(0.029)	(0.030)	(0.029)	(0.038)	(0.035)	(0.037)
Low-/middle income country	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
High income country	0.020	0.031	0.021	0.014	0.027	0.008	-0.027	-0.025	-0.015
	(0.041)	(0.044)	(0.040)	(0.039)	(0.044)	(0.039)	(0.047)	(0.055)	(0.046)
Imple: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Imple: Private sector or NGO	0.047*	0.038	-0.004	0.043*	0.034	0.005	0.040	0.022	-0.048
	(0.028)	(0.031)	(0.042)	(0.026)	(0.033)	(0.046)	(0.039)	(0.038)	(0.053)
Design: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Design: Private sector or NGO	0.094***	0.097**	0.084**	0.064*	0.075*	0.073**	0.090**	0.087*	0.069
	(0.035)	(0.038)	(0.034)	(0.036)	(0.040)	(0.036)	(0.042)	(0.047)	(0.048)
Does not provide extra services		ref.	ref.		ref.	ref.		ref.	ref.
Provides extra services		0.002	-0.007		-0.006	-0.018		0.037	0.027
		(0.032)	(0.034)		(0.031)	(0.035)		(0.039)	(0.039)
Training in classroom and at workplace		ref.	ref.		ref.	ref.		ref.	ref.
Training only in classroom		-0.091**	-0.114**		-0.070*	-0.089**		-0.130**	-0.161**
		(0.040)	(0.048)		(0.040)	(0.044)		(0.057)	(0.068)
Training only at workplace		-0.034	0.002		0.008	0.039		-0.127**	-0.093
		(0.041)	(0.048)		(0.039)	(0.054)		(0.056)	(0.057)
Training duration: short (<400 h)		ref.	ref.		ref.	ref.		ref.	ref.
Training duration: medium (400-800 hours)		-0.033	-0.041		-0.051	-0.055		-0.004	-0.032
		(0.041)	(0.040)		(0.041)	(0.040)		(0.051)	(0.048)
Training duration: long (>800 hours)		-0.005	-0.019		0.003	-0.004		-0.041	-0.077*
		(0.035)	(0.038)		(0.038)	(0.042)		(0.040)	(0.044)
Program does not provide business skills			ref.			ref.			ref.
Program provides business and entrp. skills training			-0.005			-0.056			0.066
			(0.041)			(0.043)			(0.048)
Program does not provide soft skills			ref.			ref.			ref.
Program provides soft skills training			0.057*			0.053			0.06
			(0.031)			(0.034)			(0.037)
Program does not provide certification			ref.			ref.			ref.
Program provides certification			0.069			0.063			0.071
			(0.046)			(0.051)			(0.054)
No participation incentives			ref.			ref.			ref.
Participation incentives			-0.026			-0.03			0.003
			(0.039)			(0.038)			(0.054)
Constant	0.060	0.103**	0.128**	0.068	0.100**	0.134**	0.091*	0.160***	0.174***
	(0.048)	(0.050)	(0.057)	(0.042)	(0.050)	(0.060)	(0.053)	(0.059)	(0.057)
Estimates	1651	1651	1651	1054	1054	1054	597	597	597
Studies	89	89	89	80	80	80	75	75	75
Interventions	97	97	97	96	96	96	77	77	77

Note: This table shows estimation results of the effect of vocational training on aggregate labor market outcomes (columns I-III) and employment- (columns IV-VI) and earnings-related outcomes (columns VII-IX) separately. Models were estimated using Robust Variance Estimation (RVE), correlated random effects models (Tipton, 2013), setting Rho at 0.8. All models account for publication selection bias using SE^2_{ij}, the sampling variance of effect size estimate i in intervention j. Standard errors are clustered at the *interventionXcohort* level. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

Table 8: HICs & LMICs, RVE correlated effects model, controlling for publication bias

	X	XI	XII	XII	XIV	XV
		HICs			**LMICs**	
SMD's sampling variance	0.721	0.598	0.472	1.320**	1.292**	1.232**
	(0.769)	(0.616)	(0.547)	(0.519)	(0.518)	(0.517)
Employment construct	ref.	ref.	ref.	ref.	ref.	ref.
Income construct	-0.016	-0.032	-0.023	0.045*	0.044*	0.041*
	(0.036)	(0.036)	(0.040)	(0.024)	(0.023)	(0.022)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.096	-0.049	-0.06	-0.076**	-0.065*	-0.052
	(0.058)	(0.057)	(0.065)	(0.038)	(0.036)	(0.034)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	-0.099*	-0.129*	-0.121	0.044	0.037	0.028
	(0.053)	(0.070)	(0.075)	(0.051)	(0.049)	(0.048)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.104***	-0.054	-0.021	0.006	0.008	0.001
	(0.040)	(0.055)	(0.074)	(0.023)	(0.027)	(0.030)
Unconditional outcome	ref.	ref.	ref.	ref.	ref.	ref.
Conditional outcome	-0.116	-0.102*	-0.088	-0.066**	-0.083***	-0.088***
	(0.076)	(0.056)	(0.076)	(0.026)	(0.029)	(0.031)
Measurement: <1 year from program end	ref.	ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)	0.062	0.064	0.079	-0.032	-0.049	-0.054
	(0.059)	(0.065)	(0.070)	(0.029)	(0.031)	(0.033)
Long-term follow-up(>2 years)	0.069	0.076	0.093	-0.005	-0.016	-0.015
	(0.053)	(0.053)	(0.063)	(0.031)	(0.036)	(0.033)
Both male and female participants	ref.	ref.	ref.	ref.	ref.	ref.
Measurement: Male participants	-0.056	-0.039	-0.042	-0.045	-0.063*	-0.065*
	(0.034)	(0.037)	(0.033)	(0.035)	(0.032)	(0.034)
Measurement: Female participants	-0.024	-0.007	-0.006	-0.041	-0.046	-0.038
	(0.036)	(0.042)	(0.040)	(0.037)	(0.037)	(0.036)
Older participants (>25y)	ref.	ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)	0.014	-0.074	-0.09	-0.028	-0.035	-0.009
	(0.064)	(0.078)	(0.090)	(0.034)	(0.036)	(0.040)
Imple: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Imple: Private sector or NGO	0.074*	0.035	0.012	-0.013	-0.052	-0.077
	(0.041)	(0.047)	(0.089)	(0.040)	(0.055)	(0.063)
Design: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Design: Private sector or NGO	0.157***	0.114	0.129	0.063*	0.090*	0.057
	(0.059)	(0.085)	(0.087)	(0.033)	(0.049)	(0.058)
Does not provide extra services		ref.	ref.		ref.	ref.
Provides extra services		0.091	0.05		0.004	0.000
		(0.090)	(0.109)		(0.036)	(0.041)
Training in classroom and at workplace		ref.	ref.		ref.	ref.
Training only in classroom		-0.068	-0.076		-0.087	-0.122
		(0.059)	(0.082)		(0.065)	(0.087)
Training only at workplace		-0.094	-0.048		-0.046	-0.015
		(0.075)	(0.117)		(0.050)	(0.064)
Training duration: short (<400 h)		ref.	ref.		ref.	ref.
Training duration: medium (400-800 hours)		0.098**	0.097		-0.080	-0.082*
		(0.048)	(0.065)		(0.053)	(0.048)
Training duration: long (>800 hours)		0.119**	0.119*		-0.015	-0.034
		(0.048)	(0.071)		(0.055)	(0.057)
Program does not provide business skills			ref.			ref.
Program provides business and entrp. skills training			-0.058			0.025
			(0.129)			(0.042)
Program does not provide soft skills			ref.			ref.
Program provides soft skills training			0.052			0.026
			(0.068)			(0.042)
Program does not provide certification			ref.			ref.
Program provides certification			0.038			0.08
			(0.097)			(0.069)
No participation incentives			ref.			ref.
Participation incentives			0.041			-0.04
			(0.057)			(0.064)
Constant	0.062	0.050	0.011	0.117**	0.204***	0.220***
	(0.066)	(0.085)	(0.118)	(0.053)	(0.063)	(0.074)
Estimates	649	649	649	1002	1002	1002
Studies	36	36	36	61	61	61
Interventions	39	39	39	50	50	50

Note: This table shows estimation results of the effect of vocational training on aggregate labor market outcomes in HICs (columns X-XII) and LMICs-(columns XIII-XV) separately. Models were estimated using Robust Variance Estimation (RVE), correlated random effects models (Tipton, 2013), setting Rho at 0.8. All models account for publication selection bias using SE_{ij}^2, the sampling variance of effect size estimate i in intervention j. Standard errors are clustered at the *interventionXcohort* level. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

6.5 Robustness checks

We first provide WLS estimates of our main univariate and multivariate regressions. Then, we show results of our main specifications using only the sample of experimental studies.

6.5.1 Univariate WLS models

As a robustness check, we re-run our univariate regressions using WLS regressions. Results can be found in Table 11 in the Appendix . The results are qualitatively the same as for the RVE model across all samples. However, in line with previous meta-analysis that compare both approaches (Kaiser and Menkhoff, 2019), average effect sizes estimated from WLS models are smaller than in RVE models. The reason is that WLS models place more weight on larger studies, assuming that each estimate relates to a single true effect. Given the dependency structure of our data, we consider the RVE model more plausible in our setting.

6.5.2 Multivariate WLS models

Using unrestricted WLS model to test the robustness of the RVE estimations does not change our main conclusions from the multivariate regressions, i.e. the magnitude, sign and statistical significance of the effects does not change substantially. Some estimates become more significant in statistical terms. This may reflect the fact that by ignoring dependencies between effect sizes, WLS underestimates the standard errors of meta-regression coefficients, leading to higher type I errors López-López et al. (2017). Tables 17 to 22 in the Appendix report the results of WLS regressions.

6.5.3 Restricting analysis to RCT sample

As a first robustness check, we restrict our pooled sample to the sub-sample of studies that employ experimental evaluation methods (see Table 14 in the Appendix). Reassuringly, we find that almost all coefficient estimates are similar in sign and magnitude to that of the full sample in Table 7. However, practically all results are not statistically significant. With the exception of lower effects of training on conditional outcomes, the finding of the aggregate sample that longer and more intensive interventions provide for lower impact estimates and positive effects of business skills training. One reason for the different results in terms of statistical significance may be that our effective sample size is more than halved.

To further check potential heterogeneity between estimates from experimental studies, we look into differences by income level. For the sample of HICs, there are only 8 interventions, which is too few degrees of freedom for our multivariate regressions. Therefore, we only analyze earnings and employment outcomes for the sample of experimental studies from LMICs and even for the sample of MICs.[28] Results can be found in Tables 15 and 16. Regressions using earnings as outcome variable confirm some of our main findings. First, that interventions with only in-classroom or workplace-based training are less effective

[28] Unfortunately, the sample on LICs only has with 11 interventions also too few degrees of freedom to run separate multivariate regressions. Therefore, we stick to the sample of MICs for further dis-aggregation.

than interventions that combine both. Second, that less intensive and hence shorter training interventions (<400 hours) provide for higher impact estimates than medium (400-800 hours) but especially than longer (>800 hours) training interventions in LMICs and MICs. Third, short-term (> 1 year) impacts of training are higher in LMICs and MICs relative to medium- (1-2 years after intervention exit) to long-term (< 2 years) impacts. Taking the last tow findings together suggests that participation in training in LMICs and MICs may primarily act as a signaling and screening device, helping to integrate youth in the labor market or firms to hire a better selection of workers. Fourth, estimates are lower on conditional (e.g. formal employment or wages) than unconditional outcomes. Fifth, effect sizes are generally smaller in male *and* female sub-sample analyses as opposed to samples including males and females together. Sixth, in the context of LMICs, young ($\leq$ 25) profit relatively less than older training participants. Overall, the results are more robust in terms of statistical significance for the sample of MICs, which implies that the results found in Table 15 are mainly driven by the sample of MICs.

7. Conclusion

In this meta-analysis, we analyze the impact of vocational training interventions on youth labor market outcomes in LMICs and HICs. Thereby, we focus on vocational training that is typically part of Active Labor Market Programs (ALMPs), i.e. non-formal education interventions. To quantify the impact of vocational training interventions, we employ the Standardized Mean Difference (SMD), in particular Hedges' g, which captures the direction and magnitude of the treatment effect relative to the comparison mean. We use random-effects robust-variance estimation (RVE), which addresses the common problem of meta-analyses, namely statistically dependent effect sizes, by explicitly modelling between-study heterogeneity *and* within-study measurement error. We run regressions for the aggregate sample of studies for all outcomes pooled together, for employment and earnings, for experimental studies and for LMICs and HICs separately. As a robustness check, we compare the results of the RVE models with that of an unrestricted weighted least squares (WLS) model advocated by Stanley and Doucouliagos (2012, 2017).

Results from univariate models indicate that training interventions have an overall positive and statistically significant impact on all outcomes (SMD=0.12***). Thereby, the impact of training on employment outcomes (SMD=0.11***) is lower than on earnings (SMD=0.12***). Consistent with findings in the literature, the average impact of training is large in Lower-and Middle-Income Countries (LMICs) (SMD=0.13***) than in High-Income Countries (HICs) (SMD=0.10***). A sample split using only experimental studies confirms the magnitude and statistical significance of the effect (SMD=0.10***). All findings stay significant are robust when controlling for reporting bias, though all decrease in magnitude.

Besides computing the average effect, we use multivariate meta-regressions to evaluate factors driving the impact heterogeneity of training interventions. Results reveal substantial differences between HICs and LMICs. Results provide six key findings. First, we find that interventions with only in-classroom or workplace-based training are less effective than interventions that combine both. Second, the results of the

aggregate sample show that more intensive and hence longer training interventions provide for lower impact estimates of training. However, differentiating the effect of training intensity by income setting reveals that the overall effect is driven by LMICs, while in HICS, medium (400-800 hours) and longer (>800 hours) training interventions lead to more favorable outcomes than shorter (<400 hours) interventions. Evidence points at larger medium- (1-2 years after intervention exit) to long-term (> 2 years) impacts of training in HICs and larger short-term (< 1 year) impacts of training in LMICs. The latter two findings together suggest a signaling and screening effect of training in LMICs and a human capital effect in HICs. Third, we find larger effect sizes for interventions that involve actors from the private sector (e.g. employer associations), NGOs or (inter-) national organizations in the design or implementation of interventions relative to government actors only. The effect on implementation design is more robust in statistical terms. Fourth, the impact of training on conditional outcomes, such as formal employment or wages, is lower than on unconditional outcomes, namely employment or earnings. This effect is more pronounced for the intensive than extensive margin of employment. It is more robust for LMICs than HICs in statistical terms, which may be due due to a lower incidence of wage and formal employment opportunities in LMICs. In addition, in LMICs, we find evidence of a larger effect on earnings than on employment. This is in line with the fact that youth in LMICs are more often underemployed rather than unemployed (Fields, 2004). Fifth, intention-to-treat (ITT) impact estimates are lower than estimates based on the average treatment effect on the treated (ATT) or (local) average treatment effect ((L)ATE). The finding of lower ITT-based impact estimates is particularly robust in the context of LMICs Finally, across all samples and specifications, our results show that effect sizes are generally smaller in male *and* female sub-sample analyses as opposed to samples including males and females together. However, the finding on males is more robust in terms of statistical significance.

Running multivariate regressions using WLS as a robustness test confirms our main findings. Further, restricting our multivariate analysis to the sample of experimental interventions for the entire sample as a robustness check only confirms our finding on lower estimates for conditional outcomes and that more intense and longer training interventions lead to lower estimates. In contrast, using only the sample of LMICs confirms most of our main findings.

The effectiveness of vocational training interventions has been doubted by scholars and policy makers (Blattman and Ralston, 2015). These doubts are mainly based on two points of criticism. First, the labor market impacts of vocational training interventions found in the literature, with magnitudes close to zero up to a fifth of a standard deviation, are often perceived as too small and have lead governments and donors to doubt the effectiveness training interventions. Second, that training interventions are not cost-effective if calculations are based on private returns, not to speak of the social costs of training. Raising the question whether training interventions are the best use of limited public funds (Blattman and Ralston, 2015; McKenzie, 2017).

Regarding the first criticism, four points stand out. First, the magnitudes of unconditional averages in our meta-analysis are larger than close to zero up to a fifth of a standard deviation (e.g. as reported in

Kluve et al. (2019) (SMD=0.05***)). Our estimates are more comparable to the meta-analysis of Tripney and Hombrados (2013) (e.g. SMD=0.13**). Second, the impact estimates found by our study, between [10.4;12.5] standard deviations, are in fact small in *absolute terms* if compared to impact estimates in behavioural sciences (Cohen, 2013). According to Cohen (2013), effect sizes of 0.10 are considered small, of 0.25 medium and of 0.40 large. However, Cohen (2013) besides categorizing the size of impact estimates in *absolute terms*, he also emphasizes considering these in *relative terms*, e.g. relative to the effects of comparable fields in the literature. Relative to other ALMP measures, vocational training has been found to be one of the most effective ALMP interventions, besides entrepreneurship training (Card et al., 2018; Escudero et al., 2018; Kluve et al., 2019). Third, contrasting the returns of training interventions with the returns to schooling in general, the effects are within the same order of magnitude (McKenzie, 2017). In addition, the recent literature on education interventions is a natural comparison. In a recent large-scale meta-analysis Kraft (2020) provides benchmarks based on a review of 1,942 effect sizes from 747 RCTs evaluating education interventions with standardized test outcomes. In this review, 0.05 is considered small, 0.10 is the median effect size, and anything above 0.20 is considered larges. Fourth, while the impact of training interventions may be considered as small, more than two thirds of studies in our sample only report short-term outcomes ($<$1 year), while only about two thirds report medium- to long-term outcomes (1-2 and $>$ 2 years). This may lead to a significant underestimation of the true (long-term) impact of training (see Section 2 for more details). In fact, the meta-analyses of Card et al. (2018) and Kluve et al. (2019) find larger long-run impacts of ALMPs, including training.

Against the second criticism, while some simple cost-benefit studies have proven the ineffectiveness of some interventions (McKenzie, 2017), other studies (Maitra and Mani, 2017) and more sophisticated cost-benefit analyses show that training interventions are cost-effective- especially in the long run (Lammers and Kok, 2019). In addition, given the comparatively low share of long-term estimates in the literature, the cost-effectiveness of vocational training may be underestimated. Further, assuming that training interventions can to some extent be regarded as "second-chance" investments in human capital for youth- especially from disadvantaged backgrounds. If its true that private returns of training interventions are comparable to private returns from schooling, and assuming that costs of training interventions are in general not substantially higher than costs of schooling in general, then arguing training interventions being ineffective is questionable. If training interventions are partly seen as educational programs for young people, it might good to increase their quality. And to have a long-term perspective in implementing training interventions rather than stopping them after one or two pilot-rounds, which is often the case.

References

Acemoglu, D. and Pischke, J. S. (1998). Why do firms train? theory and evidence. *The Quarterly Journal of Economics*, 113(1):79–119.

Acevedo, P., Cruces, G., Gertler, P., and Martinez, S. (2020). How vocational education made women better off but left men behind. *Labour Economics*, page 101824.

Adhvaryu, A., Kala, N., and Nyshadham, A. (2018). The skills to pay the bills: Returns to on-the-job soft skills training. Technical report, National Bureau of Economic Research.

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

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.

Askarov, Z. and Doucouliagos, H. (2020). A meta-analysis of the effects of remittances on household education expenditure. *World Development*, 129:104860.

Autor, D. H. (2001). Why do temporary help firms provide free general skills training? *The Quarterly Journal of Economics*, 116(4):1409–1448.

Bassi, V. and Nansamba, A. (2020). Screening and Signaling Non-Cognitive Skills: Experimental Evidence from Uganda. [Online; accessed 10. Nov. 2020].

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, 70(5, Part 2):9–49.

Blattman, C. and Ralston, L. (2015). Generating employment in poor and fragile states: Evidence from labor market and entrepreneurship programs.

Borenstein, M., Hedges, L., Higgins, J., and Rothestein, H. (2009). Chapter 20: meta-regression. *Introduction to meta-analysis*, pages 187–203.

Brunello, G. and De Paola, M. (2004). Market failures and the under-provision of training.

Card, D., Kluve, J., and Weber, A. (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

Chakravorty, B. and Bedi, A. S. (2019). Skills Training and Employment Outcomes in Rural Bihar. *The Indian Journal of Labour Economics*, 62(2):173–199.

Cho, Y., Kalomba, D., Mobarak, A. M., and Orozco, V. (2013). Gender differences in the effects of vocational training: Constraints on women and drop-out behavior. *World Bank Research Working Paper 6545*.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Collaboration, C. et al. (2014). Campbell systematic reviews: Policies and guidelines. *Campbell Systematic Reviews*, 1.

Crépon, B., Ferracci, M., and Fougère, D. (2012). Training the unemployed in france: How does it affect unemployment duration and recurrence? *Annals of Economics and Statistics*, (107/108):175–199.

Crépon, B. and Premand, P. (2019). Direct and Indirect Effects of Subsidized Dual Apprenticeships. IZA Discussion Papers 12793, Institute of Labor Economics (IZA).

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109):629–634.

Escudero, V., Kluve, J., López Mourelo, E., and Pignatti, C. (2018). Active labour market programmes in latin america and the caribbean: Evidence from a meta-analysis. *The Journal of Development Studies*, pages 1–18.

Evans, D. K. and Yuan, F. (2020). How big are effect sizes in international education studies. *Center for Global Development, Working Paper*, 545.

Fields, G. S. (2004). A guide to multisector labor market models.

Fitzenberger, B., Osikominu, A., and Paul, M. (2010). The heterogeneous effects of training incidence and duration on labor market transitions. *ZEW-Centre for European Economic Research Discussion Paper*, (10-077).

Groh, M., Krishnan, N., McKenzie, D., and Vishwanath, T. (2016). The impact of soft skills training on female youth employment: Evidence from a randomized experiment in jordan. *IZA Journal of Labor & Development*, 5(1):9.

Ham, J. C. and Lalonde, R. J. (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica*, 64(1):175–205.

Hanushek, E. A., Schwerdt, G., Woessmann, L., and Zhang, L. (2017). General education, vocational education, and labor-market outcomes over the lifecycle. *Journal of Human Resources*, 52(1):48–87.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and van Aert, R. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451–464.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128.

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, 1(1):39–65.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Hirshleifer, S., McKenzie, D., Almeida, R., and Ridao-Cano, C. (2016). The impact of vocational training for the unemployed: Experimental evidence from turkey. *The Economic Journal*, 126(597):2115–2146.

Ignatowski, C. (2017). What works in soft skills development for youth employment. a donor's perspective. *Youth Employment Funders Group (YEFG)*.

ILO (2020a). Global Employment Trends for Youth 2020: Technology and the future of jobs. [Online; accessed 15. May 2020].

ILO (2020b). World Employment and Social Outlook. [Online; accessed 7. Apr. 2020].

ILO (2020c). Youth and covid-19: impacts on jobs, education, rights and mental well-being: survey report 2020.

Kaiser, T. and Menkhoff, L. (2019). Financial education in schools: A meta-analysis of experimental studies. *Economics of Education Review*, page 101930.

Kluve, J., Puerto, S., Robalino, D., Romero, J. M., Rother, F., Stoeterau, J., Weidenkaff, F., and Witte, M. (2017). Interventions to improve the labour market outcomes of youth: A systematic review of training, entrepreneurship promotion, employment services and subsidized employment interventions. *Campbell Systematic Reviews*, 13(1):1–288.

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

Kluve, J., Schneider, H., Uhlendorff, A., and Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2):587–617.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.-M. K., Hammerstrøm, K., and Sathe, N. (2017). Searching for studies: A guide to information retrieval for campbell. *Campbell Systematic Reviews*.

Lammers, M. and Kok, L. (2019). Are active labor market policies (cost-) effective in the long run? evidence from the netherlands. *Empirical Economics*, pages 1–28.

Lechner, M., Miquel, R., and Wunsch, C. (2011). Long-run effects of public sector sponsored training in west germany. *Journal of the European Economic Association*, 9(4):742–784.

López-López, J. A., Van den Noortgate, W., Tanner-Smith, E. E., Wilson, S. J., and Lipsey, M. W. (2017). Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A monte carlo simulation. *Research synthesis methods*, 8(4):435–450.

Maitra, P. and Mani, S. (2017). Learning and earning: Evidence from a randomized evaluation in India. *Labour Economics*, 45:116–130.

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

Montenegro, C. E. and Patrinos, H. A. (2014). *Comparable estimates of returns to schooling around the world*. The World Bank.

Muehlemann, S., Pfeifer, H., Walden, G., Wenzelmann, F., and Wolter, S. C. (2010). The financing of apprenticeship training in the light of labor market regulations. *Labour economics*, 17(5):799–809.

OECD (2020). Recognition of non-formal and informal learning. [Online; accessed 21. Nov. 2020].

Osikominu, A. (2013). Quick Job Entry or Long-Term Human Capital Development? The Dynamic Effects of Alternative Training Schemes. *Review of Economic Studies*, 80(1):313–342.

Osikominu, A. (2016). The dynamics of training programs for the unemployed. *IZA World of Labor*.

Psacharopoulos, G. (2018). Returns to education: A further international update and implications. *Discounting and Environmental Policy*, page 63.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics*, pages 281–306. Elsevier.

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

Stanley, T. D. and Doucouliagos, H. (2017). Neither fixed nor random: weighted least squares meta-regression. *Research synthesis methods*, 8(1):19–42.

Sumberg, J., Fox, L., Flynn, J., Mader, P., and Oosterom, M. (2020). Africa's 'youth employment'crisis is actually a 'missing jobs' crisis. *Development Policy Review*.

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research synthesis methods*, 4(2):169–187.

Tripney, J. S. and Hombrados, J. G. (2013). Technical and vocational education and training (TVET) for young people in low-and middle-income countries: a systematic review and meta-analysis. *Empirical Research in Vocational Education and Training*, 5(1):3.

Vaillancourt, F. (1995). The private and total returns to education in canada, 1985. *Canadian Journal of Economics*, pages 532–554.

Vooren, M., Haelermans, C., Groot, W., and Maassen van den Brink, H. (2019). The effectiveness of active labor market policies: a meta-analysis. *Journal of Economic Surveys*, 33(1):125–149.

Wolter, S. C. and Ryan, P. (2011). Apprenticeship. In *Handbook of the Economics of Education*, volume 3, pages 521–576. Elsevier.

World Bank (2019). *World Development Report 2019*. Washington, DC: World Bank.

Yeyati, E. L., Montané, M., Sartorio, L., et al. (2019). What works for active labor market policies? a meta analysis. Technical report, Universidad Torcuato Di Tella.

A. Inclusion criteria (PICOS)

A.1 (P) Participants: Youth

We apply two main inclusion features with regards to participants: First, studies must have investigated interventions that are designed for youth or primarily target them. Most often, this refers to women or men aged between 15 and 35. If reports do not explicitly state an age criteria, we consider whether the authors refer to the intervention target group as youth or young. We exclude programs that do not explicitly target youth, even if reports provide youth sub sample analyses. Second, youth participants must have exited the formal education system, either with or without a (primary/secondary) degree. We therefore exclude after-school or holidays (e.g. summer-holidays) interventions. However, this distinction was not always straightforward as in the case of the "adolescent development clubs" implemented in Uganda and evaluated by Alfonsi et al. (2020), that could be frequented also by young girls currently enrolled in school. We do not apply any further criteria on participants or targeting mechanisms of interventions (e.g. unemployed, low skilled, disadvantaged).

A.2 (I) Interventions: training interventions

We apply three inclusion criteria with regards to interventions: First, interventions must provide vocational/technical skills training as the main component. This basic vocational skills training component may be combined with other skills training components – e.g. basic skills training (mathematics, language), non-cognitive skills (soft or life skills), business skills – or with other services, such as employment subsidies, entrepreneurial promotion or employment services (e.g. job counselling, job search assistance, mentoring or job placement). But the vocational training component should be considered the main component of the intervention. Either account for at least 50% of the duration of the intervention, or be mentioned by the authors as a vocational training intervention. This means that we exclude pure entrepreneurial skills training programs.[29] Second, the main goal/objective of the intervention should be to improve the immediate labor market situation of youth – e.g. their employment status or earnings. A large part of included training interventions are therefore conducted in the framework of public Active Labor Market Programs (ALMPs). Third, interventions must be set outside the (formal) education system. This is in line with the inclusion criteria on the intervention's target group (youth who completed education) and intended goal (improved labor market outcomes). Hence, we exclude trainings that are typically within the framework of national education systems, such as Technical Vocational Education and Training (TVET). Such programs often expect that beneficiaries continue their education following participation. Thus, they are not comparable in terms of outcome measures, characteristics (e.g. duration) and participants (e.g. age).

Beyond these criteria, we consider a broad range of design options for trainings. We do not exclude interventions based on their duration, intensity or other design features. We include interventions irrespective

[29] Furthermore, such interventions commonly aim to improve self-employment outcomes, which we exclude as outcomes from our analysis.

of their geographic location – i.e. we include interventions from all countries and regions (urban, rural) – or of their implementation scale (local, regional, national). We include trainings irrespective of the learning model they follow.[30] We consider training irrespective of their delivery place. We include interventions delivered primarily at the workplace (e.g. on-the-job trainings, (informal) apprenticeship or internships), interventions delivered primarily in classroom settings, as well as interventions combining classroom- and workplace components.[31]

A.3 (C) Comparison group

We include only studies that follow a counterfactual evaluation approach to estimate the impact of an intervention. Thus studies must compare outcomes of beneficiaries ("treatment group") relative to a comparison group. Comparison groups have to be established following a counterfactual research design (see section A.5). Eligible comparison groups receive no intervention nor receive the intervention later (e.g. pipeline or waitlist design). We exclude impact evaluations that report *only* relative impact estimates between different interventions. Note, however, that the comparison group might be exposed to policies or programs other than the evaluated intervention.

A.4 (O) Outcomes

Eligible studies must report at least one measure of one of the primary labor market outcomes of interest. We follow Kluve et al. (2019) and distinguish two main types of outcome constructs. First, employment-related outcomes – e.g. employment probability, labor market participation rate, hours worked and quality of employment (e.g. (fixed) contract, benefits). Second, earnings-related outcome constructs like individual earnings and wages. We do not apply specific criteria for these outcome constructs (e.g. thresholds for considering an individual as employed). Rather, we follow the definitions applied in the respective primary studies. We consider outcomes both in case of dependent- and self-employment, as well as in cases where both situations are evaluated together. We consider outcomes irrespective of whether they are measured conditional or unconditional on other (upstream or intermediary) outcomes. One example is whether earnings are measured conditional or unconditional on being employed. The conditional measured earnings would include zero values for those who are unemployed, whereas conditional measures exclude these individuals from the sample (thus leading to higher average incomes).

A.5 (S) Study characteristics

We consider several key study characteristics in our inclusion criteria. First, studies must be completed. This means that data collection should be finalized and that the study results are published in reports either as an article in a peer-reviewed journal, or in any other publication form ("grey literature"), e.g. official working papers, reports of international organizations or of NGOs and PhD theses. Second, studies must implement

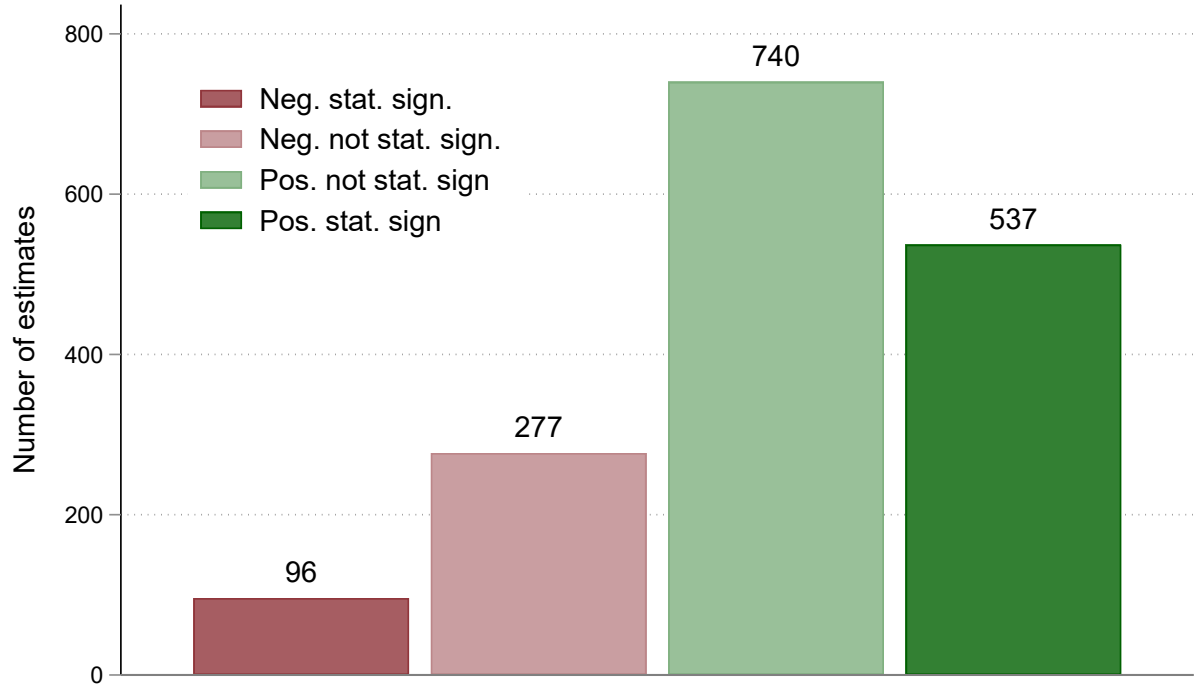
[30] OECD (2020) identifies three types of learning models: "informal", "non formal" and "formal" learning.

[31] We define the interventions as "combined" or "dual" when the share of classroom- or workplace delivery does not exceed 75%.

an experimental or quasi-experimental counterfactual research design. This includes: (i) randomized experiments, (ii) methods for causal inference under unconfoundedness (matching), and (iii) selection on unobservables (instrumental variables, regression discontinuity design, difference-in-differences). Third, we include studies written in English, French or Spanish. Fourth, we include studies that were *first* published between 1990 and June 2019. For each of these studies, we include the most recent available version up to June 2020. Therefore, a study published for the first time in January 2020 would not be included in the meta-analysis. On the contrary, a study published in January 2020 in a peer-reviewed journal and being previously available as a working paper in the 1990-2019 period, would be included in its peer-reviewed version.

B. Additional tables and figures

Figure 3: Direction and significance of included estimates



Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100. Statistical significance at the 5% level.

Table 9: Literature review existing meta-analyses

Paper	# of studies/ interventions	Focus	Sample	Share of training interventions	Share employing experimental design	Measure	Effect of training interventions
		Evaluate the following type(s) of intervention(s):	*Geographic region, age-based target group, published &/or unpublished studies, publication period*			*Measure for magnitude &/or measure sign & direction of effect*	
Tripney and Hombrados (2013)	26/20	training interventions	LMIC & HICs, all ages, published & unpublished studies, 2000-2011	100%	11.50%	Magnitude: Hedge's g	Paid employment: Hedge's g=0.134** Formal employment: Hedge's g=0.199** Monthly earnings: Hedge's g=0.127** Self-employment earnings: Hedge's g=0.025** Weekly hours worked: Hedge's g=0.043**
Escudero et al. (2018)	51/53	ALPMs	Latin America & the Carribean, all ages, published & unpublished studies, 1980-2011	75%	16.60%	Sign& direction: binary variable for positive and statistically significant (PSS) treatment effect at a 5% sign. level	Stat. sign. positive effect on formal employment and earnings
Card et al. (2018)	207/n.n.	ALPMs	80% HICs, 10% LICs, 10% MICs, all ages, published & unpublished studies, 1980-Oct.2014	50%	19%	Magnitude: prob(employed); sign& direction: PSS (see cell above for explanation)	Stat. sign. positive effect on probability to be employed & labor market outcomes (using PSS)
Vooren et al. (2019)	57/n.n.	ALPMs	OECD countries, all ages, only published studies, 1990-2017	38.50%	7.30%	Magnitude: Hedge's g	Effect on labor market outcomes: Hedge's g= -0.01; disentangling this effect by follow-up timing, the initially negative effect turns positive at 24 months.
Yeyati et al. (2019)	73/102	ALPMs	LMIC & HICs, all ages, published & unpublished studies, 2000-2018	45.10%	100%	Sign&direction: PSS (see above for explanation)	Stat. sign. positive effect on labor market outcomes
Kluve et al. (2019)	113/107	ALPMs	HICs (N=65) & LMICs (N=48), youth-targeted ALMPs (youth aged 15-35), published & unpublished studies, 1990-2014	51%	47%	Magnitude: Hedge's g; sign& direction: PSS (see above for explanation)	Effect on labor market outcomes: Hedge's g= 0.05***; relative to other ALMPs, skills training interventions seem to be the most effective

Figure 4: Evaluated training interventions by country

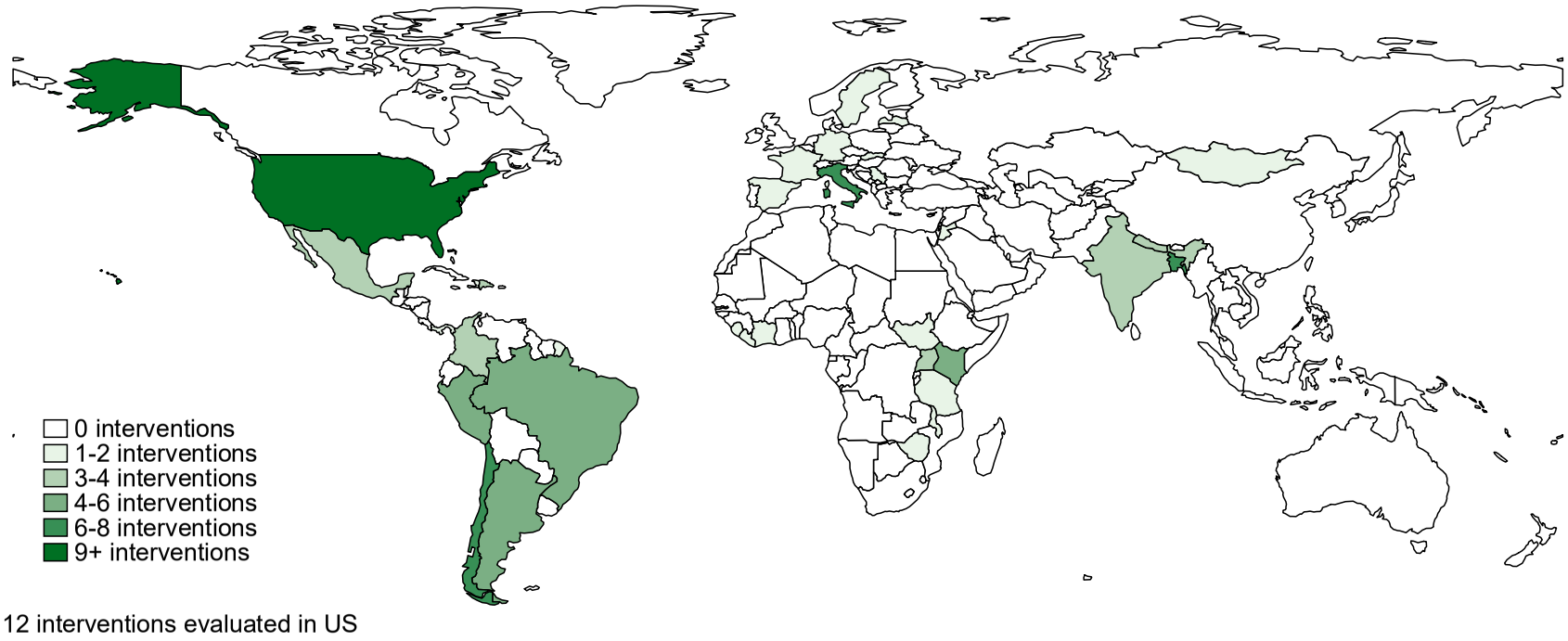


Table 10: Summary statistics of the SMD, with and without censoring

Description	Mean	95% CI	Min / Max	#Est / #Int	Direction of Estimates (abs. / rel.)			
					NSS	NnSS[32]	PnSS	PSS
Without Censoring								
Pooled	0.103	[0.093 - 0.113]	-1.141 / 1.995	1690 / 97	106 / 6%	284 / 17%	757 / 45%	542 / 32%
Employment	0.114	[0.101 - 0.127]	-0.881 / 1.995	1075 / 96	57 / 5%	162 / 15%	508 / 47%	348 / 32%
Earnings	0.082	[0.066 - 0.098]	-1.141 / 1.811	615 / 77	49 / 8%	122 / 20%	249 / 41%	194 / 32%
With Censoring								
Pooled	0.094	[0.086 - 0.103]	-0.881 / 0.947	1651 / 97	96 / 6%	277 / 17%	740 / 45%	537 / 33%
Employment	0.104	[0.093 - 0.115]	-0.881 / 0.947	1054 / 96	50 / 5%	161 / 15%	498 / 47%	345 / 33%
Earnings	0.077	[0.064 - 0.089]	-0.332 / 0.923	597 / 77	46 / 8%	116 / 19%	242 / 41%	192 / 32%

Note: # Est= Number of estimates; # Int= Number of interventions; NSS= Negative statistically significant; NnSS= Negative not statistically significant; PnSS= Positive not statistically significant; PSS= Positive statistically significant. Statistical significance at the 5% level.

C. Robust Variance Estimation (RVE) in detail

More formally, to explain effect size i in study j, that is, T_{ij}, by p study-level covariates $X_{1ij}...X_{pij}$, the following multivariate meta-regression could be specified:

$$T_{ij} = \beta_1 X_{1ij} + \beta_2 X_{1ij} + ... + \beta_p X_{pij} + \varepsilon_{ij} \tag{6}$$

The Weighted Least Squares (WLS) estimate of $\beta = (\beta_1, ..., \beta_p)$ can be estimated using

$$b = (X'WX)^{-1}(X'WT) \tag{7}$$

with W matrix of weights and X the design matrix. The variance of the estimate b of β can be written as

$$V(b) = (X'WX)^{-1}(X'W\Sigma WX)(X'WX)^{-1} \tag{8}$$

where Σ is the variance–covariance matrix of the effect size estimates. While the diagonal values of Σ are the variances of effect sizes, the off-diagonal elements are the covariances between effect size pairs, which are assumed to be zero in case a study included in a meta-regression model contributes *less* than one effect size estimate to the model. In case a study contributes *more* than one effect size estimate to the meta-regression model, this assumption is no longer appropriate (Tripney and Hombrados, 2013). RVE overcomes the problem of non-zero covariances between effect size pairs in two steps. First, each of the included m studies must be independent, even if effect size estimates within studies can be statistically

[32]Few observations with zero value were added to the negative not statistically significant category.

dependent. Hence, 8 can be re-written (Tripney and Hombrados, 2013):

$$V(b) = (\sum_{j=1}^{m} X_j' W_j X_j)^{-1} (\sum_{j=1}^{m} X_j' W_j \Sigma_j W_j X_j) (\sum_{j=1}^{m} X_j' W_j X_j)^{-1} \tag{9}$$

If fixed or random effects models are used to estimate equation 2, the elements of Σ_j are modeled as known (fixed effects models) or partially known and partially estimated (random effects models).

Second, since the problem of unknown covariance matrices Σ_j in multivariate meta-regression cannot be overcome by estimating each element of Σ_j in equation 9, as the number of estimates required would exceed the number of data elements, RVE estimates an average of the linear combinations involving Σ_j, by reformulating equation 9 as:

$$V(b) = (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j X_j)^{-1} (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j \Sigma_j W_j X_j) (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j X_j)^{-1} \tag{10}$$

This can be done by using the cross products of residuals within study j as a crude estimate of the covariances in Σ_j, i.e. to replace Σ_j in equation 10 by $e_j e_j'$, where $e_j = T_j \check{} X_{jb}$, which is the estimated residual vector for study j. Even if $e_j e_j'$ is a rather crude estimate of Σ_j, it is good enough that

$$V(b) = (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j X_j)^{-1} (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j e_j e_j' W_j X_j) (\frac{1}{m}\sum_{j=1}^{m} X_j' W_j X_j)^{-1} \tag{11}$$

converges in probability to the correct value as $m \to \infty$ (Hedges et al., 2010).

To estimate $V(b)$ in equation 10 and to increase efficiency of estimates, we need to calculate the weights W_j. Even if RVE does not impose any requirement on the weights W_j, Hedges et al. (2010) mention that approximately inverse variance weights are statistically the most efficient ones. However, since calculating exact weights requires knowing the covariance structure of Σ_j to invert Σ_j and since estimating all variance parameters is computationally heavy, Hedges et al. (2010) offer two methods for deriving simple, approximately inverse variance weights, which do not involve the within-study covariances: correlated effects and hierarchical effects weights.

D. WLS Univariate Regressions & Testing for reporting bias using WLS

Table 11 shows the results of univariate WLS regressions. Overall, the magnitude of the estimated effect is only about half of that when using RVE. The average effect for the pooled estimation has about the same magnitude as the effect of training interventions reported by (Kluve et al., 2019).

Table 11: WLS Univariate meta-regression

	Pooled	Employment	Earnings	LMIC	HIC	RCT
Average effect	0.054***	0.061***	0.045***	0.076***	0.038***	0.056***
	(0.008)	(0.008)	(0.010)	(0.007)	(0.006)	(0.011)
	0.038,0.070	0.044,0.077	0.024,0.066	0.062,0.090	0.025,0.050	0.034,0.078
Estimates	1651	1054	597	1002	649	1075
Interventions	97	96	77	61	36	42
Reports	89	80	75	50	39	54

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Weight: inverse number of estimates per intervention
* p<0.10, ** p<0.05, *** p<0.01

In a first step, we test for the presence of reporting bias (null $H_0^I : \gamma = 0$) by means of the Egger Test (Egger et al., 1997), or so-called "Funnel Asymmetry Test" (FAT) (Stanley and Doucouliagos, 2012). Therefore, we include the standard error of estimates in the univariate meta-regression model (model 4). Note that this test has low power. Therefore, the We also test for reporting bias by means of regressions, thereby following Stanley and Doucouliagos (2012). In a first step, we test for the presence of reporting bias (null $H_0^I : \gamma = 0$) by means of the "Funnel Asymmetry Test" (FAT). Therefore, we include the standard error of estimates in the univariate meta-regression model (model 4). In the same estimation, we conduct a "Precision Effect Test" (PET), testing for the presence of a genuine effect beyond reporting bias (null $H_0^{II} : \alpha_{PET} = 0$).

$$g_{ij} = \alpha_{PET} + SE_{ij}\gamma_{FAT} + \nu_j + \varepsilon_{ij} \tag{12}$$

Where SE_{ij} represents the standard error of effect size estimate i in study j.

As suggested by Stanley and Doucouliagos (2012) if H_0^{II} can be rejected, we estimate a Precision-Effect-Estimate with Standard Error (PEESE) model, which includes the variance instead of the standard error to account for publication bias (model 5). According to Stanley and Doucouliagos (2012), model 5 provides a more precise estimate of the true empirical effect than model 4 *iff the null* $H_0^{II} : \alpha_{PET} = 0$ *cannot be rejected*.

$$g_{ij} = \alpha_{PEESE} + SE_{ij}^2\gamma + \nu_j + \varepsilon_{ij} \tag{13}$$

Table **??** shows the results of FAT and PET tests. Reporting bias seems to be present in all regressions (FAT-test), while a genuine effect beyond reporting bias exists for all outcomes (PET-test). The magnitude of the reporting bias suggests a selection effect for more positive results. Compared to the estimates of the univariate meta-regression shown in Table 2, the magnitude of average effects when controlling for reporting bias in the PEESE are slightly lower but still statistically significant for all outcomes.

Table 12: WLS Funnel Asymmetry Test (FAT), Precision Effect Test (PET) and Precision Effect Estimate with Standard Errors (PEESE)

	Pooled	Employment	Earnings	LMIC	HIC	RCT
FAT-PET						
SE	0.734***	0.757***	0.646**	0.616**	0.552	0.517
	(0.230)	(0.225)	(0.283)	(0.294)	(0.360)	(0.323)
Average effect	0.026**	0.030**	0.021**	0.047**	0.020**	0.037*
	(0.012)	(0.015)	(0.010)	(0.018)	(0.009)	(0.018)
PEESE						
SE^2	2.334**	2.746***	1.626	3.066**	1.361*	2.889*
	(1.053)	(0.966)	(1.005)	(1.298)	(0.700)	(1.641)
Average effect	0.049***	0.054***	0.042***	0.066***	0.036***	0.050***
	(0.008)	(0.009)	(0.010)	(0.009)	(0.006)	(0.012)
Estimates	1651	1054	597	1002	649	1075
Interventions	97	96	77	61	36	42
Reports	89	80	75	50	39	54

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Weight: inverse variance

* p<0.10, ** p<0.05, *** p<0.01

Table 13: Sample splits for HICs and LMICs, RVE Precision Effect Estimate with Standard Errors (PEESE)

	All	Employment	Earnings	RCT
		HICs		
Publ Bias (SMD VAR)	1.229	1.565	0.696	-0.903
	(1.648)	(1.754)	(1.094)	(4.230)
	-2.001,4.460	-1.873,5.002	-1.448,2.841	-9.193,7.387
Average effect	0.096***	0.090***	0.065	0.077
	(0.026)	(0.022)	(0.042)	(0.064)
	0.044,0.147	0.047,0.133	-0.018,0.148	-0.048,0.202
Estimates	649	376	273	311
Interventions	36	36	18	8
Reports	39	39	38	35
		LMICs		
Publ Bias (SMD VAR)	1.220**	1.033**	5.374***	1.855***
	(0.551)	(0.522)	(1.842)	(0.579)
	0.140,2.300	0.010,2.057	1.764,8.983	0.719,2.990
Average effect	0.112***	0.102***	0.094***	0.094***
	(0.021)	(0.022)	(0.022)	(0.023)
	0.071,0.153	0.059,0.145	0.050,0.138	0.050,0.139
Estimates	1002	678	324	764
Interventions	61	60	59	34
Reports	57	55	54	37
		MICs		
Publ Bias (SMD VAR)	1.445	1.316	5.132*	1.252
	(0.951)	(0.865)	(2.645)	(1.024)
	-0.418,3.308	-0.379,3.011	-0.051,10.316	-0.755,3.259
Average effect	0.119***	0.104***	0.115***	0.109***
	(0.025)	(0.026)	(0.031)	(0.030)
	0.069,0.168	0.053,0.156	0.055,0.175	0.051,0.168
Estimates	899	611	288	701
Interventions	48	47	46	25
Reports	46	44	43	28
		LICs		
Publ Bias (SMD VAR)	1.293*	1.005	4.073	2.307*
	(0.707)	(0.683)	(3.071)	(1.363)
	-0.092,2.678	-0.335,2.344	-1.947,10.093	-0.364,4.979
Average effect	0.074**	0.080**	0.048**	0.064
	(0.032)	(0.038)	(0.022)	(0.039)
	0.010,0.137	0.006,0.154	0.005,0.090	-0.014,0.141
Estimates	103	67	36	63
Interventions	13	13	13	9
Reports	11	11	11	9

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

* p<0.10, ** p<0.05, *** p<0.01

E. Multivariate regressions using only the sample of experimental studies

Table 14: Pooled outcomes, Results of RCTs only, RVE correlated effects model, controlling for publication bias

	I	II	III	IV	V	VI
SMD's sampling variance	1.975***	2.318***	2.116***	1.895***	2.063***	2.063***
	(0.574)	(0.709)	(0.625)	(0.488)	(0.543)	(0.543)
Employment construct	ref.	ref.	ref.	ref.	ref.	ref.
Income construct	0.026	0.040	0.048	0.028	0.018	0.018
	(0.031)	(0.035)	(0.033)	(0.025)	(0.022)	(0.022)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.045	-0.024	-0.040	-0.040	-0.035	-0.035
	(0.044)	(0.040)	(0.036)	(0.034)	(0.028)	(0.028)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	0.017	0.026	0.011	0.008	0.021	0.021
	(0.040)	(0.034)	(0.028)	(0.032)	(0.030)	(0.030)
Unconditional outcome		ref.	ref.	ref.	ref.	ref.
Conditional outcome		-0.090***	-0.094***	-0.097***	-0.107***	-0.107***
		(0.027)	(0.027)	(0.032)	(0.034)	(0.034)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		0.010	0.023	0.035	0.039	0.039
		(0.056)	(0.054)	(0.052)	(0.049)	(0.049)
Long-term follow-up(>2 years)		-0.024	-0.005	-0.009	-0.015	-0.015
		(0.023)	(0.027)	(0.033)	(0.034)	(0.034)
Both male and female participants		ref.	ref.	ref.	ref.	ref.
Measurement: Male participants		-0.054	-0.044	-0.052	-0.038	-0.038
		(0.034)	(0.030)	(0.035)	(0.034)	(0.034)
Measurement: Female participants		-0.030	-0.033	-0.054	-0.041	-0.041
		(0.037)	(0.037)	(0.042)	(0.038)	(0.038)
Older participants (>25y)		ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)		-0.006	-0.017	-0.053	-0.041	-0.041
		(0.050)	(0.048)	(0.035)	(0.037)	(0.037)
Low-/middle income country			ref.	ref.	ref.	ref.
High income country			-0.062	-0.029	-0.028	-0.028
			(0.052)	(0.072)	(0.069)	(0.069)
Imple: Private/NGO not involved				ref.	ref.	ref.
Imple: Private sector or NGO				0.041	0.002	0.002
				(0.045)	(0.051)	(0.051)
Design: Private/NGO not involved				ref.	ref.	ref.
Design: Private sector or NGO				0.028	-0.000	-0.000
				(0.053)	(0.060)	(0.060)
Does not provide extra services				ref.	ref.	ref.
Provides extra services				0.073*	0.091	0.091
				(0.041)	(0.056)	(0.056)
Training in classroom and at workplace				ref.	ref.	ref.
Training only in classroom				-0.040	-0.032	-0.032
				(0.050)	(0.054)	(0.054)
Training only at workplace				-0.017	-0.005	-0.005
				(0.050)	(0.054)	(0.054)
Training duration: short (<400 h)				ref.	ref.	ref.
Training duration: medium (400-800 hours)				-0.053	-0.047	-0.047
				(0.053)	(0.052)	(0.052)
Training duration: long (>800 hours)				-0.074	-0.101**	-0.101**
				(0.047)	(0.050)	(0.050)
Program does not provide business skills					ref.	ref.
Program provides business and entrp. skills training					0.077*	0.077*
					(0.044)	(0.044)
Program does not provide soft skills					ref.	ref.
Program provides soft skills training					-0.015	-0.015
					(0.043)	(0.043)
Program does not provide certification					ref.	ref.
Program provides certification					0.040	0.040
					(0.050)	(0.050)
No participation incentives					ref.	ref.
Participation incentives					0.049	0.049
					(0.048)	(0.048)
Constant	0.101**	0.115*	0.145**	0.142	0.092	0.092
	(0.040)	(0.069)	(0.063)	(0.092)	(0.095)	(0.095)
Estimates	1075	1075	1075	1075	1075	1075
Interventions	54	54	54	54	54	54
Studies	42	42	42	42	42	42

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: all, Outcome: SMD, Estimation: robumeta_corr, pub. bias covar: EES_SMD_VAR, cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 15: Results of RCTs only for LMICs, earnings and employment outcomes, RVE correlated effects model, controlling for publication bias

	I	II	III	IV	V	VI
	Earnings			Employment		
SMD's sampling variance	4.003**	3.735**	3.926**	1.998***	1.850***	2.002***
	(1.966)	(1.808)	(1.874)	(0.560)	(0.483)	(0.462)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.026	-0.031	-0.029*	-0.006	-0.025	-0.039
	(0.030)	(0.025)	(0.017)	(0.043)	(0.043)	(0.042)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	0.021	0.027	0.044	-0.000	-0.002	0.020
	(0.034)	(0.038)	(0.035)	(0.031)	(0.047)	(0.044)
Unconditional outcome	ref.	ref.	ref.	ref.	ref.	ref.
Conditional outcome	-0.057**	-0.087***	-0.084***	-0.123***	-0.122**	-0.121***
	(0.023)	(0.031)	(0.026)	(0.040)	(0.047)	(0.046)
Measurement: <1 year from program end	ref.	ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)	-0.030	-0.031	-0.010	-0.039	-0.038	-0.034
	(0.036)	(0.030)	(0.032)	(0.033)	(0.036)	(0.036)
Long-term follow-up(>2 years)	-0.035	-0.049*	-0.042	-0.011	-0.008	-0.013
	(0.025)	(0.027)	(0.026)	(0.024)	(0.038)	(0.032)
Both male and female participants	ref.	ref.	ref.	ref.	ref.	ref.
Measurement: Male participants	-0.061	-0.088**	-0.055	-0.065*	-0.062	-0.031
	(0.037)	(0.045)	(0.040)	(0.037)	(0.053)	(0.052)
Measurement: Female participants	-0.046	-0.060	-0.032	-0.072*	-0.083	-0.057
	(0.036)	(0.039)	(0.035)	(0.043)	(0.055)	(0.051)
Older participants (>25y)	ref.	ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)	-0.114**	-0.096**	-0.075	-0.037	-0.053	-0.076
	(0.056)	(0.045)	(0.046)	(0.052)	(0.057)	(0.062)
Imple: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Imple: Private sector or NGO	0.018	0.009	0.019	-0.002	0.021	0.014
	(0.022)	(0.076)	(0.076)	(0.025)	(0.076)	(0.078)
Design: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Design: Private sector or NGO	0.015	0.030	-0.021	0.052	0.024	-0.024
	(0.038)	(0.063)	(0.056)	(0.032)	(0.091)	(0.082)
Does not provide extra services		ref.	ref.		ref.	ref.
Provides extra services		0.065	0.105**		0.068	0.120*
		(0.052)	(0.046)		(0.067)	(0.067)
Training in classroom and at workplace		ref.	ref.		ref.	ref.
Training only in classroom		-0.087*	-0.045		-0.016	0.029
		(0.051)	(0.043)		(0.073)	(0.073)
Training only at workplace		-0.099	-0.096		-0.001	-0.013
		(0.064)	(0.066)		(0.068)	(0.085)
Training duration: short (<400 h)		ref.	ref.		ref.	ref.
Training duration: medium (400-800 hours)		-0.012	0.005		-0.035	-0.037
		(0.070)	(0.063)		(0.079)	(0.081)
Training duration: long (>800 hours)		-0.093**	-0.121***		-0.083	-0.107
		(0.043)	(0.042)		(0.078)	(0.088)
Program does not provide business skills			ref.			ref.
Program provides business and entrp. skills training			0.066			0.035
			(0.040)			(0.068)
Program does not provide soft skills			ref.			ref.
Program provides soft skills training			-0.061***			-0.100
			(0.023)			(0.063)
Program does not provide certification			ref.			ref.
Program provides certification			0.009			-0.037
			(0.035)			(0.070)
No participation incentives			ref.			ref.
Participation incentives			0.073*			0.064
			(0.041)			(0.073)
Constant	0.199**	0.250**	0.151	0.135**	0.177	0.200
	(0.083)	(0.125)	(0.151)	(0.068)	(0.149)	(0.193)
Estimates	231	231	231	533	533	533
Reports	34	34	34	33	33	33
Interventions	32	32	32	31	31	31

Note: This table shows estimation results of the effect of vocational training on earnings (columns I-III) and employment in LMICS (columns IV-VI) separately. Models were estimated using Robust Variance Estimation (RVE), correlated random effects models (Tipton, 2013), setting Rho at 0.8. All models account for publication selection bias using SE_{ij}^2, the sampling variance of effect size estimate i in intervention j. Standard errors are clustered at the $interventionXcohort$ level. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

Table 16: Results of RCTs only for MICs, earnings and employment outcomes, RVE correlated effects model, controlling for publication bias

	I	II	III	IV	V	VI
	Earnings			Employment		
SMD's sampling variance	2.989	2.779	2.517	2.047**	2.364***	2.410***
	(2.216)	(2.316)	(2.145)	(0.812)	(0.634)	(0.661)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.004	-0.012	-0.004	0.018	-0.004	-0.001
	(0.041)	(0.016)	(0.011)	(0.053)	(0.036)	(0.039)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.033	-0.076**	-0.048	-0.056*	-0.127	-0.125
	(0.039)	(0.037)	(0.040)	(0.032)	(0.077)	(0.083)
Unconditional outcome	ref.	ref.	ref.	ref.	ref.	ref.
Conditional outcome	-0.060**	-0.085***	-0.066***	-0.115***	-0.133**	-0.159***
	(0.026)	(0.024)	(0.021)	(0.040)	(0.054)	(0.058)
Measurement: <1 year from program end	ref.	ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)	-0.063	-0.026	-0.029	-0.053	-0.025	-0.018
	(0.039)	(0.024)	(0.026)	(0.036)	(0.028)	(0.026)
Long-term follow-up(>2 years)	-0.069**	-0.049**	-0.057**	-0.017	0.010	0.014
	(0.030)	(0.019)	(0.023)	(0.026)	(0.028)	(0.029)
Both male and female participants	ref.	ref.	ref.	ref.	ref.	ref.
Measurement: Male participants	-0.029	-0.083*	-0.054	-0.031	-0.040	-0.037
	(0.049)	(0.047)	(0.045)	(0.039)	(0.059)	(0.060)
Measurement: Female participants	-0.025	-0.087**	-0.053	-0.042	-0.082	-0.067
	(0.043)	(0.035)	(0.037)	(0.038)	(0.077)	(0.064)
Older participants (>25y)	ref.	ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)	-0.152**	-0.184***	-0.183**	-0.042	-0.078	-0.027
	(0.060)	(0.063)	(0.076)	(0.057)	(0.107)	(0.153)
Imple: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Imple: Private sector or NGO	0.019	-0.055	-0.014	0.014	-0.008	0.066
	(0.048)	(0.066)	(0.071)	(0.030)	(0.120)	(0.148)
Design: Private/NGO not involved	ref.	ref.	ref.	ref.	ref.	ref.
Design: Private sector or NGO	-0.036	0.072	0.051	0.025	0.135	0.280
	(0.037)	(0.088)	(0.114)	(0.037)	(0.123)	(0.238)
Does not provide extra services		ref.	ref.		ref.	ref.
Provides extra services		-0.035	-0.046		-0.095	-0.226
		(0.072)	(0.094)		(0.097)	(0.191)
Training in classroom and at workplace		ref.	ref.		ref.	ref.
Training only in classroom		-0.203***	-0.127**		-0.096	-0.081
		(0.059)	(0.053)		(0.119)	(0.098)
Training only at workplace		-0.075*	-0.040		0.053	0.121
		(0.044)	(0.050)		(0.075)	(0.138)
Training duration: short (<400 h)		ref.	ref.		ref.	ref.
Training duration: medium (400-800 hours)		-0.052	-0.038		-0.052	-0.019
		(0.056)	(0.071)		(0.101)	(0.118)
Training duration: long (>800 hours)		-0.146***	-0.145***		-0.145	-0.179
		(0.044)	(0.053)		(0.109)	(0.116)
Program does not provide business skills			ref.			ref.
Program provides business and entrp. skills training			-0.022			-0.120
			(0.035)			(0.117)
Program does not provide soft skills			ref.			ref.
Program provides soft skills training			-0.055			0.024
			(0.037)			(0.070)
Program does not provide certification			ref.			ref.
Program provides certification			0.012			0.030
			(0.054)			(0.100)
No participation incentives			ref.			ref.
Participation incentives			0.101***			0.078
			(0.039)			(0.068)
Constant	0.297***	0.496***	0.366*	0.149*	0.299	0.083
	(0.112)	(0.162)	(0.201)	(0.084)	(0.285)	(0.304)
Estimates	211	211	211	490	490	490
Reports	25	25	25	24	24	24
Interventions	24	24	24	23	23	23

Note: This table shows estimation results of the effect of vocational training on earnings (columns I-III) and employment in MICS (columns IV-VI) separately. Models were estimated using Robust Variance Estimation (RVE), correlated random effects models (Tipton, 2013), setting Rho at 0.8. All models account for publication selection bias using SE_{ij}^2, the sampling variance of effect size estimate i in intervention j. Standard errors are clustered at the *interventionXcohort* level. */**/*** denotes statistical significance at the 10%/5%/1%-level. Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100.

F. WLS Multivariate Regressions

Table 17: All outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5	Spec6
Employment construct	ref.	ref.	ref.	ref.	ref.	ref.
Income construct	0.008	0.018	0.015	0.013	0.012	0.014
	(0.022)	(0.023)	(0.023)	(0.021)	(0.020)	(0.019)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.070**	-0.061*	-0.066**	-0.073**	-0.051*	-0.053**
	(0.033)	(0.032)	(0.031)	(0.030)	(0.028)	(0.025)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	0.015	0.014	0.006	0.000	0.007	0.012
	(0.035)	(0.037)	(0.040)	(0.037)	(0.036)	(0.034)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.024	-0.017	-0.018	-0.017	-0.014	-0.018
	(0.028)	(0.026)	(0.025)	(0.023)	(0.023)	(0.024)
Unconditional outcome		ref.	ref.	ref.	ref.	ref.
Conditional outcome		-0.064**	-0.066**	-0.055*	-0.074**	-0.082***
		(0.028)	(0.027)	(0.029)	(0.029)	(0.029)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		-0.003	0.006	-0.002	-0.004	-0.004
		(0.030)	(0.030)	(0.027)	(0.026)	(0.026)
Long-term follow-up(>2 years)		-0.007	0.002	0.000	-0.001	-0.002
		(0.025)	(0.026)	(0.027)	(0.029)	(0.028)
Both male and female participants		ref.	ref.	ref.	ref.	ref.
Measurement: Male participants		-0.029	-0.029	-0.036	-0.041	-0.053**
		(0.023)	(0.023)	(0.027)	(0.026)	(0.025)
Measurement: Female participants		0.003	-0.000	-0.021	-0.020	-0.026
		(0.031)	(0.030)	(0.031)	(0.030)	(0.029)
Older participants (>25y)		ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)		-0.007	-0.003	-0.035	-0.044*	-0.041
		(0.028)	(0.029)	(0.026)	(0.026)	(0.025)
Low-/middle income country			ref.	ref.	ref.	ref.
High income country			-0.035	0.032	0.041	0.034
			(0.035)	(0.039)	(0.040)	(0.036)
Imple: Private/NGO not involved				ref.	ref.	ref.
Imple: Private sector or NGO				0.049*	0.036	0.000
				(0.026)	(0.028)	(0.036)
Design: Public/NGO not involved				ref.	ref.	ref.
Design: Private sector or NGO				0.103***	0.103***	0.091***
				(0.033)	(0.034)	(0.031)
Does not provide extra services					ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services					0.011	-0.001
					(0.029)	(0.031)
Training in classroom and at workplace					ref.	ref.
Training only in classroom					-0.089**	-0.106**
					(0.036)	(0.041)
Training only at workplace					-0.047	-0.013
					(0.036)	(0.042)
Training duration: short (<400 h)					ref.	ref.
Training duration: medium (400-800 hours)					-0.039	-0.049
					(0.037)	(0.036)
Training duration: long (>800 hours)					-0.013	-0.033
					(0.032)	(0.035)
Program does not provide business skills						ref.
Program provides business and entrp. skills training						-0.003
						(0.037)
Program does not provide soft skills						ref.
Program provides soft skills training						0.051*
						(0.027)
Program does not provide certification						ref.
Program provides certification						0.066
						(0.040)
No participation incentives						ref.
Participation incentives						-0.012
						(0.035)
Constant	0.146***	0.156***	0.169***	0.076*	0.125***	0.137***
	(0.024)	(0.031)	(0.035)	(0.044)	(0.046)	(0.051)
Estimates	1651	1651	1651	1651	1651	1651
Reports						
Interventions						
Adjusted R-squared	0.021	0.028	0.032	0.076	0.102	0.116

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: all, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 18: Employment outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5	Spec6
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.057*	-0.052	-0.055*	-0.063*	-0.044	-0.055*
	(0.031)	(0.032)	(0.032)	(0.032)	(0.031)	(0.030)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	0.002	0.001	-0.005	-0.006	-0.006	0.006
	(0.034)	(0.036)	(0.038)	(0.037)	(0.036)	(0.034)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.017	-0.017	-0.019	-0.015	-0.017	-0.026
	(0.029)	(0.027)	(0.027)	(0.025)	(0.026)	(0.028)
Unconditional outcome		ref.	ref.	ref.	ref.	ref.
Conditional outcome		-0.014	-0.014	-0.008	-0.022	-0.036
		(0.036)	(0.036)	(0.037)	(0.036)	(0.036)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		0.000	0.007	-0.002	-0.009	-0.016
		(0.025)	(0.027)	(0.025)	(0.026)	(0.026)
Long-term follow-up(>2 years)		0.009	0.016	0.013	0.016	0.017
		(0.025)	(0.026)	(0.026)	(0.028)	(0.027)
Both male and female participants		ref.	ref.	ref.	ref.	ref.
Measurement: Male participants		-0.036	-0.037	-0.041	-0.049	-0.064**
		(0.026)	(0.026)	(0.029)	(0.029)	(0.029)
Measurement: Female participants		-0.002	-0.004	-0.020	-0.020	-0.022
		(0.035)	(0.035)	(0.036)	(0.035)	(0.033)
Older participants (>25y)		ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)		-0.008	-0.005	-0.032	-0.039	-0.032
		(0.027)	(0.029)	(0.027)	(0.027)	(0.027)
Low-/middle income country			ref.	ref.	ref.	ref.
High income country			-0.025	0.027	0.042	0.025
			(0.035)	(0.039)	(0.041)	(0.037)
Imple: Private/NGO not involved				ref.	ref.	ref.
Imple: Private sector or NGO				0.046*	0.040	0.017
				(0.026)	(0.031)	(0.040)
Design: Public/NGO not involved				ref.	ref.	ref.
Design: Private sector or NGO				0.075**	0.087**	0.084**
				(0.034)	(0.037)	(0.033)
Does not provide extra services					ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services					-0.004	-0.022
					(0.030)	(0.034)
Training in classroom and at workplace					ref.	ref.
Training only in classroom					-0.065*	-0.076*
					(0.038)	(0.041)
Training only at workplace					-0.006	0.025
					(0.035)	(0.047)
Training duration: short (<400 h)					ref.	ref.
Training duration: medium (400-800 hours)					-0.065*	-0.071*
					(0.039)	(0.038)
Training duration: long (>800 hours)					-0.009	-0.025
					(0.037)	(0.041)
Program does not provide business skills						ref.
Program provides business and entrp. skills training						-0.052
						(0.039)
Program does not provide soft skills						ref.
Program provides soft skills training						0.047
						(0.030)
Program does not provide certification						ref.
Program provides certification						0.060
						(0.045)
No participation incentives						ref.
Participation incentives						-0.013
						(0.036)
Constant	0.142***	0.150***	0.160***	0.085**	0.124**	0.139**
	(0.024)	(0.029)	(0.033)	(0.041)	(0.047)	(0.055)
Estimates	1054	1054	1054	1054	1054	1054
Reports						
Interventions						
Adjusted R-squared	0.015	0.013	0.015	0.041	0.064	0.081

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: empl, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 19: Earning outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5	Spec6
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.028	-0.029	-0.032	-0.039	-0.029	-0.016
	(0.038)	(0.037)	(0.038)	(0.035)	(0.034)	(0.026)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	-0.027	-0.035	-0.044	-0.049	-0.037	-0.042
	(0.044)	(0.048)	(0.050)	(0.047)	(0.044)	(0.044)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.033	0.014	0.001	-0.010	0.004	0.011
	(0.039)	(0.032)	(0.032)	(0.033)	(0.031)	(0.029)
Unconditional outcome		ref.	ref.	ref.	ref.	ref.
Conditional outcome		-0.123***	-0.123***	-0.101***	-0.125***	-0.138***
		(0.029)	(0.028)	(0.028)	(0.028)	(0.029)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		-0.016	-0.004	-0.017	-0.026	-0.016
		(0.044)	(0.045)	(0.044)	(0.036)	(0.031)
Long-term follow-up(>2 years)		-0.052	-0.031	-0.030	-0.039	-0.048
		(0.037)	(0.036)	(0.035)	(0.037)	(0.034)
Both male and female participants		ref.	ref.	ref.	ref.	ref.
Measurement: Male participants		-0.026	-0.020	-0.037	-0.027	-0.038
		(0.033)	(0.031)	(0.034)	(0.032)	(0.028)
Measurement: Female participants		0.038	0.033	0.006	-0.004	-0.017
		(0.034)	(0.034)	(0.035)	(0.035)	(0.034)
Older participants (>25y)		ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)		0.016	0.024	-0.007	-0.018	-0.015
		(0.040)	(0.039)	(0.034)	(0.031)	(0.032)
Low-/middle income country			ref.	ref.	ref.	ref.
High income country			-0.083*	-0.025	-0.023	-0.012
			(0.045)	(0.044)	(0.048)	(0.040)
Imple: Private/NGO not involved				ref.	ref.	ref.
Imple: Private sector or NGO				0.037	0.009	-0.057
				(0.033)	(0.032)	(0.042)
Design: Public/NGO not involved				ref.	ref.	ref.
Design: Private sector or NGO				0.101***	0.098**	0.085**
				(0.038)	(0.037)	(0.037)
Does not provide extra services					ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services					0.047	0.034
					(0.036)	(0.035)
Training in classroom and at workplace					ref.	ref.
Training only in classroom					-0.117**	-0.148***
					(0.051)	(0.056)
Training only at workplace					-0.144***	-0.113**
					(0.050)	(0.049)
Training duration: short (<400 h)					ref.	ref.
Training duration: medium (400-800 hours)					-0.007	-0.034
					(0.046)	(0.040)
Training duration: long (>800 hours)					-0.054	-0.097**
					(0.036)	(0.038)
Program does not provide business skills						ref.
Program provides business and entrp. skills training						0.068
						(0.043)
Program does not provide soft skills						ref.
Program provides soft skills training						0.054*
						(0.031)
Program does not provide certification						ref.
Program provides certification						0.073
						(0.046)
No participation incentives						ref.
Participation incentives						0.008
						(0.046)
Constant	0.168***	0.176***	0.195***	0.118**	0.200***	0.210***
	(0.039)	(0.041)	(0.045)	(0.046)	(0.050)	(0.051)
Estimates	597	597	597	597	597	597
Reports						
Interventions						
Adjusted R-squared	0.020	0.085	0.111	0.154	0.256	0.301

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: earn, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 20: RCTs outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5	Spec6
Employment construct	ref.	ref.	ref.	ref.	ref.	ref.
Income construct	0.009	0.020	0.030	0.016	0.009	0.009
	(0.027)	(0.029)	(0.026)	(0.021)	(0.019)	(0.019)
Non ITT	ref.	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.046	-0.030	-0.047	-0.056*	-0.055**	-0.055**
	(0.039)	(0.037)	(0.032)	(0.029)	(0.026)	(0.026)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	0.022	0.023	0.004	0.003	0.013	0.013
	(0.036)	(0.030)	(0.026)	(0.024)	(0.024)	(0.024)
Unconditional outcome		ref.	ref.	ref.	ref.	ref.
Conditional outcome		-0.082***	-0.089***	-0.088***	-0.098***	-0.098***
		(0.024)	(0.023)	(0.026)	(0.028)	(0.028)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		0.004	0.020	0.030	0.033	0.033
		(0.047)	(0.046)	(0.041)	(0.039)	(0.039)
Long-term follow-up(>2 years)		-0.028	-0.005	-0.011	-0.014	-0.014
		(0.022)	(0.024)	(0.028)	(0.027)	(0.027)
Both male and female participants		ref.	ref.	ref.	ref.	ref.
Measurement: Male participants		-0.028	-0.017	-0.026	-0.015	-0.015
		(0.036)	(0.033)	(0.035)	(0.034)	(0.034)
Measurement: Female participants		-0.011	-0.016	-0.041	-0.030	-0.030
		(0.035)	(0.035)	(0.035)	(0.032)	(0.032)
Older participants (>25y)		ref.	ref.	ref.	ref.	ref.
Younger Participants (<25y)		-0.031	-0.043	-0.068**	-0.061**	-0.061**
		(0.043)	(0.038)	(0.027)	(0.029)	(0.029)
Low-/middle income country			ref.	ref.	ref.	ref.
High income country			-0.079*	-0.048	-0.049	-0.049
			(0.047)	(0.059)	(0.054)	(0.054)
Imple: Private/NGO not involved				ref.	ref.	ref.
Imple: Private sector or NGO				0.054	0.023	0.023
				(0.034)	(0.038)	(0.038)
Design: Public/NGO not involved				ref.	ref.	ref.
Design: Private sector or NGO				0.032	0.012	0.012
				(0.043)	(0.046)	(0.046)
Does not provide extra services				ref.	ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services				0.072**	0.083*	0.083*
				(0.031)	(0.043)	(0.043)
Training in classroom and at workplace				ref.	ref.	ref.
Training only in classroom				-0.034	-0.026	-0.026
				(0.042)	(0.042)	(0.042)
Training only at workplace				-0.025	-0.011	-0.011
				(0.041)	(0.045)	(0.045)
Training duration: short (<400 h)				ref.	ref.	ref.
Training duration: medium (400-800 hours)				-0.061	-0.058	-0.058
				(0.041)	(0.040)	(0.040)
Training duration: long (>800 hours)				-0.085**	-0.108**	-0.108**
				(0.041)	(0.044)	(0.044)
Program does not provide business skills					ref.	ref.
Program provides business and entrp. skills training					0.056	0.056
					(0.033)	(0.033)
Program does not provide soft skills					ref.	ref.
Program provides soft skills training					-0.006	-0.006
					(0.032)	(0.032)
Program does not provide certification					ref.	ref.
Program provides certification					0.039	0.039
					(0.039)	(0.039)
No participation incentives					ref.	ref.
Participation incentives					0.037	0.037
					(0.040)	(0.040)
Non-experimental design (IV, RDD, ...)						ref.
Experimental design (RCT)						0.000
						(.)
Constant	0.125***	0.160***	0.192***	0.175**	0.138*	0.138*
	(0.036)	(0.059)	(0.048)	(0.075)	(0.074)	(0.074)
Estimates	1075	1075	1075	1075	1075	1075
Reports						
Interventions						
Adjusted R-squared	0.017	0.056	0.077	0.186	0.204	0.204

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: all, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 21: LMIC outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5
Employment construct	ref.	ref.	ref.	ref.	ref.
Income construct	0.026	0.031	0.034	0.034	0.034
	(0.023)	(0.024)	(0.024)	(0.024)	(0.024)
Non ITT	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.099**	-0.084**	-0.089**	-0.080**	-0.071**
	(0.037)	(0.036)	(0.035)	(0.032)	(0.032)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	0.041	0.041	0.044	0.043	0.039
	(0.043)	(0.049)	(0.047)	(0.043)	(0.041)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.000	0.010	0.004	0.003	-0.005
	(0.032)	(0.025)	(0.024)	(0.024)	(0.026)
Unconditional outcome		ref.	ref.	ref.	ref.
Conditional outcome		-0.059**	-0.059**	-0.076***	-0.082***
		(0.026)	(0.026)	(0.027)	(0.027)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		-0.026	-0.029	-0.044	-0.054*
		(0.030)	(0.031)	(0.027)	(0.028)
Long-term follow-up(>2 years)		-0.010	-0.011	-0.025	-0.023
		(0.029)	(0.030)	(0.034)	(0.031)
Both male and female participants		ref.	ref.	ref.	ref.
Measurement: Male participants		-0.031	-0.032	-0.042	-0.050
		(0.033)	(0.035)	(0.033)	(0.036)
Measurement: Female participants		-0.019	-0.031	-0.036	-0.035
		(0.035)	(0.036)	(0.034)	(0.033)
Older participants (>25y)		ref.	ref.	ref.	ref.
Younger Participants (<25y)		-0.017	-0.038	-0.045	-0.022
		(0.035)	(0.030)	(0.032)	(0.033)
Imple: Private/NGO not involved			ref.	ref.	ref.
Imple: Private sector or NGO			-0.006	-0.038	-0.054
			(0.034)	(0.041)	(0.046)
Design: Public/NGO not involved			ref.	ref.	ref.
Design: Private sector or NGO			0.071**	0.099**	0.077
			(0.031)	(0.043)	(0.049)
Does not provide extra services				ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services				0.011	0.003
				(0.032)	(0.037)
Training in classroom and at workplace				ref.	ref.
Training only in classroom				-0.089	-0.121*
				(0.055)	(0.068)
Training only at workplace				-0.056	-0.025
				(0.045)	(0.055)
Training duration: short (<400 h)				ref.	ref.
Training duration: medium (400-800 hours)				-0.081*	-0.080*
				(0.046)	(0.041)
Training duration: long (>800 hours)				-0.025	-0.039
				(0.049)	(0.049)
Program does not provide soft skills					ref.
Program provides soft skills training					0.037
					(0.041)
Program does not provide certification					ref.
Program provides certification					0.074
					(0.057)
No participation incentives					ref.
Participation incentives					-0.041
					(0.051)
Constant	0.141***	0.165***	0.133***	0.211***	0.218***
	(0.036)	(0.033)	(0.046)	(0.052)	(0.063)
Estimates	1002	1002	1002	1002	1002
Reports					
Interventions					
Adjusted R-squared	0.039	0.049	0.064	0.103	0.118

Variables not shown (<4 interventions):

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: all, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01

Table 22: HIC outcomes, WLS, controlling for publication bias

	Spec1	Spec2	Spec3	Spec4	Spec5
Employment construct	ref.	ref.	ref.	ref.	ref.
Income construct	-0.036	-0.027	-0.014	-0.030	-0.025
	(0.045)	(0.038)	(0.030)	(0.029)	(0.029)
Non ITT	ref.	ref.	ref.	ref.	ref.
Intention-to-treat effect(ITT)	-0.058	-0.075*	-0.104**	-0.073	-0.086*
	(0.045)	(0.043)	(0.044)	(0.047)	(0.049)
Non-experimental design (IV, RDD, ...)	ref.	ref.	ref.	ref.	ref.
Experimental design (RCT)	-0.045	-0.056	-0.112**	-0.147***	-0.132**
	(0.056)	(0.049)	(0.043)	(0.053)	(0.049)
Publication not peer reviewed	ref.	ref.	ref.	ref.	ref.
Peer-reviewed publication	-0.091*	-0.112*	-0.098**	-0.056	-0.031
	(0.049)	(0.058)	(0.040)	(0.042)	(0.051)
Unconditional outcome		ref.	ref.	ref.	ref.
Conditional outcome		-0.123**	-0.107	-0.095**	-0.080
		(0.060)	(0.063)	(0.047)	(0.057)
Measurement: <1 year from program end		ref.	ref.	ref.	ref.
Medium-term follow-up (1-2 years)		0.073	0.045	0.048	0.066
		(0.058)	(0.050)	(0.052)	(0.054)
Long-term follow-up(>2 years)		0.065	0.058	0.067	0.086*
		(0.046)	(0.048)	(0.046)	(0.051)
Both male and female participants		ref.	ref.	ref.	ref.
Measurement: Male participants		-0.053	-0.063*	-0.042	-0.043
		(0.031)	(0.035)	(0.033)	(0.029)
Measurement: Female participants		0.023	-0.014	0.008	0.009
		(0.041)	(0.036)	(0.038)	(0.036)
Older participants (>25y)		ref.	ref.	ref.	ref.
Younger Participants (<25y)		0.076	-0.014	-0.085	-0.094
		(0.073)	(0.057)	(0.052)	(0.057)
Imple: Private/NGO not involved			ref.	ref.	ref.
Imple: Private sector or NGO			0.078**	0.044	0.027
			(0.038)	(0.037)	(0.058)
Design: Public/NGO not involved			ref.	ref.	ref.
Design: Private sector or NGO			0.164***	0.108*	0.124**
			(0.050)	(0.062)	(0.058)
Does not provide extra services				ref.	ref.
Provides extra services (empl. services, subsidized empl. or entr. services				0.110*	0.063
				(0.065)	(0.069)
Training in classroom and at workplace				ref.	ref.
Training only in classroom				-0.049	-0.058
				(0.048)	(0.063)
Training only at workplace				-0.109**	-0.063
				(0.047)	(0.067)
Training duration: short (<400 h)				ref.	ref.
Training duration: medium (400-800 hours)				0.094**	0.085**
				(0.037)	(0.038)
Training duration: long (>800 hours)				0.115***	0.104**
				(0.036)	(0.043)
Program does not provide business skills					ref.
Program provides business and entrp. skills training					-0.060
					(0.093)
Program does not provide soft skills					ref.
Program provides soft skills training					0.039
					(0.044)
Program does not provide certification					ref.
Program provides certification					0.041
					(0.060)
No participation incentives					ref.
Participation incentives					0.049
					(0.039)
Constant	0.170***	0.096	0.097*	0.066	0.026
	(0.034)	(0.071)	(0.053)	(0.060)	(0.077)
Estimates	649	649	649	649	649
Reports					
Interventions					
Adjusted R-squared	0.040	0.073	0.180	0.211	0.217

Variables not shown (<4 interventions): Business skills

Censoring SMD at -1 and 1, Censoring Inverse SE at 1 and 100, Setting Rho at 0.8

Outcome group: all, Outcome: SMD, Estimation: ols, pub. bias covar: , cluster: st_id

* p<0.10, ** p<0.05, *** p<0.01