



Measuring the Results of Skills Development Interventions

Experiences of Impact Evaluations by German, Swiss and Austrian Development Cooperation

Implemented by

giz Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

Imprint

Published by the

Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices

Bonn and Eschborn, Germany

Friedrich-Ebert-Allee 32 + 36

53113 Bonn, Germany

Phone +49 (0)00 123 456 789

Fax +49 (0)00 123 456 789

mario.eckardt@giz.de

www.giz.de

As at

March 2023

Design

GOPA Worldwide Consultants GmbH (GOPA)

Bad Homburg (Germany)

Photo credits

RichardoImagen (Getty Images): cover page

Text

Dr. Ursula Esser, GOPA, Bonn (Germany)

Franziska Holzaepfel, GOPA, Bad Homburg (Germany)

Patricia Zamalloa-Huegel, GOPA, Ladenburg (Germany)

Support

Janina Jasper, GOPA, Bad Homburg (Germany)

Carmen Kuntz, GOPA, Parksville (Canada)

Julia Schaefer, GOPA, Bad Homburg (Germany)

Karsten Weitzenegger, GOPA, Hamburg (Germany)

On behalf of the

German Federal Ministry for Economic Cooperation and Development (BMZ)

Alternatively: German Federal Foreign Office

\ EXECUTIVE SUMMARY

This study was conducted on behalf of the *Deutsche Gesellschaft für Internationale Zusammenarbeit* (GIZ) to support the Inter-American Development Bank (IDB) in the joint endeavor to improve **labor market oriented skills development interventions in Latin America and the Caribbean** (LAC). Skills development interventions refer to formal measures, like Technical and Vocational Education and Training (TVET), as well as non-formal, informal and on-the-job measures to increase the supply of qualified and skilled personnel for a sustainable and inclusive labor market participation and economic development in the LAC region.

The present study analyses **how results and impacts of skills development interventions are measured** in German, Austrian and Swiss (DACH) development cooperation (DC) and presents an overview of the range of impact evaluation designs and approaches used in DACH-DC. Impact evaluations (IEs) refer to evaluation designs that aim to measure the causal effect of an intervention (e.g., a skills development intervention) on an observed variable of interest, such as skills improvement, employability or income. IEs should ideally allow the formation of robust conclusions about the impact caused by an intervention and therefore causal attribution. Rigorous IE use a control group as the counterfactual situation to compare what happened due to the intervention and what would have happened without the intervention to establish causality and to assess if an intervention works.

Nine case studies present the range of impact evaluation designs – namely experimental, quasi-experimental and non-experimental designs– in a descriptive and exemplary manner. The case studies intend to identify transferable learnings from impact evaluation designs of skills development interventions for LAC and other regions. The research team was unable to identify any IEs performed by German DC in LAC. The only presented case study from the LAC region is *Case Study 5 from Brazil*, which was financed by IDB. The **methodological IE designs, approaches and methods are transferable to other contexts**, including LAC, whereas single technical IE findings are very context specific and cannot be transferred to other contexts. Systematic reviews of impact evaluations of skills development interventions are not available yet, which would be required to be able to generalize and transfer technical findings to other contexts. Thus, the study generates multiple methodological findings, derives general and specific conclusions on multiple focal topics and presents related recommendations.

The two general conclusions are summarized as follows:

1. **The understanding and use of the term rigorous IE differs in research and practice in DACH.** On the one hand, quantitative prone researchers understand experimental and quasi-experimental designs as the core of rigorous IE designs, which is in line with the publication requirements of peer-reviewed journals. On the other hand, DACH DC practitioners prefer a more comprehensive definition of rigorous IE and prefer the term robust IE, which also includes non-experimental designs and thus qualitative, theory-based approaches like contribution analyses. The DACH DC practitioners tend to select the impact evaluation design according to the utility and use of an appropriate method for a given context. DACH DC practitioners argue that there is no single right or best evaluation design or approach, but that it needs to be tailored to the evaluation purpose, objective and questions.
2. **In line with their broader definition of rigorous impact evaluations, non-experimental designs, especially contribution analyses, are the standard instruments DACH DC uses to measure results and impacts of skills development interventions.** Quasi-experimental designs were used in multiple pilot projects in the skills development sector and there is a growing trend and increasing (financial) support for these in Germany. The present study did not find any experimental designs (namely RCTs as the most rigorous method) to measure results and impacts of skills development interventions in DACH-DC practice. This is because these are perceived as rather unsuitable and not feasible in DC as they require randomization, large sample sizes, large volumes of high-quality data with high costs for data collection and ex-ante evaluation settings etc. However, international and German research institutes implemented their own skills development interventions and evaluations with the purpose to use randomized controlled trials (RCTs), but they deliberately refrained from collaborating with German DC practice to be able to implement these experimental designs. Experimental designs often contradict the adaptive, results-based management approach applied during project and program implementation in DC.

The conclusions on the eight focal topics are summarized as follows:

3. **Using Existing M&E Data:** M&E data is frequently used in non-experimental designs (*in Case Study 7 – Global and Case Study 8 – Egypt*). It is rarely used in quasi-experimental or experimental rigorous IE designs, due to: small sample sizes; incomplete and insufficient data in terms of extend and quality; and lack of suitable control group information. Most quasi-experimental and experimental rigorous IE designs collect their own data instead.
4. **Using Existing Administrative Data:** Administrative data has the potential to drastically reduce the costs of rigorous IE. However, these data sources are rarely available in an adequate data quality, and they are not easily accessible. Therefore, the rigorous IE examined have collected additional data or used purely new data. The report presents three examples for the use of administrative data in quasi-experimental designs (*Case Study 3 – Serbia, Case Study 5 – Brazil, and Case Study 6 – Philippines*). The insights from these case studies help to derive success factors and

challenges for the use of administrative data. Due to the good administrative data availability in many LAC countries, this data may pose a cost efficient alternative or – as shown in the presented case studies – rather a complement to newly collected data for rigorous IEs in LAC.

5. **Gender:** Some DC projects and studies focus on women (*like Case Study 2 – India*), so that targeted indicators were used to indicate progress for women only. Most DC-projects/programs and IEs target men and women and tended to use corresponding disaggregated indicators (*e.g. Case Studies 1, 2, 3, 4, 5, 6*). Despite existing progress in gender-responsive M&E, there is room for improvement in IE reporting. Gender-specific results and impacts are not always reported (*Case Studies 7, 8*), especially if no difference was found (*Case Studies 1, 6*) or the mechanisms causing heterogeneity are not always identified, even though gender-disaggregated indicators exist (*Case Study 5 – Brazil*).
6. **Measurement of ‘Employability’, ‘Entrepreneurship’ and ‘Non-cognitive skills’:** While no coherent definition of employability can be identified, most studies measure employability in terms of TVET graduates’ labor market outcomes such as current employment, time until employment was found, and wage earnings (*Case Studies 2, 3, 4, 5*). Employability is frequently measured using subjective self-assessment of TVET graduates only (*Case Studies 7, 8*), while employers could provide a more objective source of information. As observed in the present study, fewer studies include additional outcomes like non-cognitive skills and entrepreneurial skills of TVET graduates (*Case Studies 1, 5*). These concepts can be measured with specific instruments based on questionnaires and personal interviews about graduates’ behaviors in different scenarios.
7. **COVID-19, Green Transformation, and Technological Change:** As in many other areas, the COVID-19 pandemic has led to increased use of internet connectivity, digital tools and platforms, etc. to continue DC operations, including the conduction of remote evaluations. Video calls were used for coordination, (incl. kick-off and validation workshops), as well as for data collection, (such as virtual interviews and focus group discussions). Reduced travel and technological change are beneficial in terms of green transformation and climate change. However, purely remote data collections can lead to blind spots (*e.g., due to insufficient observations and limited access to confidential data*), which can only be overcome by on-site evaluation activities.
8. **Sustainability of Impacts:** The case studies present examples of sustained impact measurement at the beneficiary, institutional and/or systemic level (*Case Study 6 – Philippines*). This required regular tracing of participants, institutional or systemic actors and the control group since inception, over longer periods of time and beyond the project completion (*Case Studies 1, 2, 4*). However, evaluations rarely assess the sustained impact of skills development interventions beyond a period of two to three years after training completion.
9. **Efficiency and Cost-Effectiveness:** The case studies present examples for using the follow-the-money approach to measure efficiency (*Case Studies 7 – Global and 8 – Egypt*) and the value-for-money approach to measure the cost-effectiveness of skills development interventions (*Case Study 4 – Kenya*).
10. **Private sector:** The present study on measuring the results of skills development interventions provides some exemplary insights for the involvement of the private sector in the implementation and evaluation of skills development interventions (*Case Studies 4 – Kenya and 6 – Philippines*). However, these insights cannot be generalized for other contexts like the LAC region.

How to use this study? The study is structured as follows:

- Chapter 1: Presents the objective, scope and structure and the methodology of the study.
- Chapter 2: Provides an overview of evaluation designs and approaches to measure results and impacts of DC-interventions.
- Chapter 3: Summarizes evaluation policies and impact evaluation trends in DACH DC, focusing on German, Swiss and Austria’s bilateral DC actors.
- Chapter 4: Contains the condensed analysis of nine case studies of more or less rigorous IEs of skills development interventions in German DC and presents two experimental, four quasi-experimental and three non-experimental designs.
- Chapter 5: Provides a toolbox with reflection questions for practitioners who would like to evaluate skills development interventions.
- Chapter 6: Presents some strategies on how project planners, project implementers and policymakers can use the findings from impact evaluations in the skills development sector.
- Chapter 7: Summarizes the main findings, which lead to the general and focal topic-specific conclusions and key recommendations for the LAC and other contexts.

\ CONTENTS

\ EXECUTIVE SUMMARY	1
CHAPTER 1 STUDY CONCEPT	1
1.1 INTRODUCTION.....	1
1.2 METHODOLOGY.....	2
1.2.1 DESK RESEARCH.....	2
1.2.2 CASE STUDY SELECTION.....	2
1.2.3 ANALYSIS OF THE CASE STUDIES	3
1.2.4 COMPILATION	3
CHAPTER 2 EVALUATION DESIGNS AND APPROACHES TO MEASURE RESULTS AND IMPACTS OF DC-INTERVENTIONS	4
2.1 EXPERIMENTAL DESIGNS	5
2.2 QUASI-EXPERIMENTAL DESIGNS.....	5
2.3 NON-EXPERIMENTAL DESIGNS	8
2.4 COVID-19-RELATED CHALLENGES AND SOLUTIONS FOR IE	10
2.5 DIGITAL TOOLS FOR IE	11
CHAPTER 3 EVALUATION POLICIES AND TRENDS WITHIN DACH DC	12
3.1 GERMANY'S DC EVALUATION POLICY AND TRENDS	12
3.1.1 GERMANY'S GIZ EVALUATION POLICY	14
3.1.2 GERMANY'S KfW EVALUATION POLICY	15
3.2 SWITZERLAND'S DC EVALUATION POLICY.....	17
3.2.1 SWITZERLAND'S SDC EVALUATION POLICY.....	17
3.2.2 SWITZERLAND'S SECO EVALUATION POLICY	19
3.3 AUSTRIA'S DC EVALUATION POLICY.....	20
CHAPTER 4 CASE STUDIES - OVERVIEW OF EXISTING IE AND STUDIES OF SKILLS DEVELOPMENT INTERVENTIONS IN GERMAN DC	23
4.1 EXPERIMENTAL DESIGNS	23
4.1.1 CASE STUDY 1: UGANDA – RANDOMIZED EVALUATION OF STUDENT TRAINING FOR ENTREPRENEURIAL PROMOTION	23
4.1.2 CASE STUDY 2: INDIA – RANDOMIZED EVALUATION OF SUBSIDIZED VOCATIONAL TRAINING FOR WOMEN	26
4.2 QUASI-EXPERIMENTAL DESIGNS.....	29
4.2.1 CASE STUDY 3: SERBIA – EMPLOYMENT IMPACTS OF GERMAN DC INTERVENTIONS	29
4.2.2 CASE STUDY 4: KENYA – EMPLOYMENT & INCOME EFFECTS OF SKILLS DEVELOPMENT INTERVENTIONS	33
4.2.3 CASE STUDY 5: BRAZIL – NON-COGNITIVE SKILLS AND LABOR MARKET OUTCOMES.....	38
4.2.4 CASE STUDY 6: PHILIPPINES – DUAL VOCATIONAL TRAINING	41
4.3 NON-EXPERIMENTAL DESIGNS	45
4.3.1 CASE STUDY 7: GLOBAL – SKILLS FOR REINTEGRATION.....	45
4.3.2 CASE STUDY 8: EGYPT – EMPLOYMENT PROMOTION	49
4.3.3 CASE STUDY 9: GLOBAL – SUCCESS FACTORS FOR GENDER EQUALITY IN VOCATIONAL EDUCATION AND TRAINING	52
CHAPTER 5 REFLECTION QUESTIONS PRIOR TO IMPACT EVALUATIONS	54
CHAPTER 6 STRATEGIES ON HOW TO USE THE FINDINGS FROM IMPACT EVALUATIONS IN THE SKILLS DEVELOPMENT SECTOR	57
6.1 USING META-EVALUATIONS FOR PROJECT PLANNING.....	57
6.2 USING PARTICIPATORY APPROACHES FOR PROJECT IMPLEMENTATION	58
6.3 USING IE FINDINGS FOR POLICYMAKING.....	59
CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS	60

ANNEXES	70
ANNEX 1	EVALUATIONS SELECTED FOR ANALYSIS70
ANNEX 2	PERSONS INTERVIEWED71
ANNEX 3	SEMI-STRUCTURED QUESTIONNAIRES FOR INTERVIEWS.....72
ANNEX 4	DEFINITION OF KEY CONCEPTS75
ANNEX 5	FURTHER LITERATURE77

\ FIGURES

Figure 1: OECD/DAC evaluation criteria (OECD 2021).....	4
Figure 2: The OECD/DAC evaluation criteria and the theory of change (based on Sammeth et al. 2010)	4
Figure 3: Number of rigorous impact evaluations over time	13
Figure 4: The QUER app allows digital and interactive access >1,000 evaluation findings	16
Figure 5: Three phases and 15 steps of the evaluation process (ADA 2020)	21
Figure 6: The Austrian DC guidance on evaluability assessment.....	22

\ TABLES

Table 1: When to use different quasi-experimental designs.....	6
Table 2: Description of non-experimental designs.....	8
Table 3: Six steps with reflection questions prior to conducting an impact evaluation	54

GENERAL ABBREVIATIONS

ADA	Austrian Development Agency
BMZ	German Federal Ministry for Economic Cooperation and Development (<i>Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung</i>)
C4ED	Center for Evaluation and Development
CPE	Central Project Evaluation
DeGEval	German Evaluation Society (<i>Deutsche Gesellschaft für Evaluation</i>)
CEval	Center for Evaluation (<i>Centrum für Evaluation</i>)
CS	Case Study
CSO	Civil Society Organization
DAAD	German Academic Exchange Service (<i>Deutscher Akademischer Austauschdienst</i>)
DAC	Development Assistance Committee
DACH	Germany (D), Austria (A) and Switzerland (CH) or German (D), Austrian (A) and Swiss (CH)
DC	Development Cooperation
DEP	3ie's Development Evidence Portal
DEval	German Institute for Development Evaluation (<i>Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit</i>)
DiD	Difference-in-Difference
DTS	Dual Training System
EC	European Commission
E4D	Employment and Skills for Development in Africa initiative (Kenya)
EEP	Employment Promotion Project (Egypt)
EGM	Evidence Gap Map
FC	Financial Cooperation
FGD	Focus Group Discussions
GIO	German Implementing Organizations
GIZ	<i>Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH</i>
IDB	Inter-American Development Bank
IDOS	German Institute of Development and Sustainability
IDP	Internally Displaced People
IE	Impact Evaluation
IZA	Institute of Labor Economics (<i>Forschungsinstitut zur Zukunft der Arbeit</i>)
IV	Instrumental Variables
KAM	Kenya Association of Manufactures
KfW	German Development Bank (<i>Kreditanstalt für Wiederaufbau</i>)
KM	Kirkpatrick Model
LAC	Latin America and the Caribbean
MAPP	Method for Impact Assessment of Programs and Projects

M&E	Monitoring and Evaluation
MoESTD	Ministry of Education, Science and Technological Development (Serbia)
MSC	Most Significant Change
NES	Serbian National Employment Service
NGO	Non-Governmental Organization
ODA	Official Development Assistance
OECD	Organization for Economic Co-operation and Development
OeEB	Development Bank of Austria (<i>Österreichische Entwicklungsbank AG</i>)
PRONATEC	National Program for Access to Technical Education and Employment (Brazil) (<i>Bolsa-Formação</i>)
QUER	Quick Evidence Results
RCT	Randomized Controlled Trial
RED	Rigorous Evidence Database (by the German Development Cooperation)
ROSCA	Rotating Savings and Credit Association
RQ	Research Question
RWI	Leibniz Institute for Economic Research (<i>Leibniz-Institut für Wirtschaftsforschung</i>)
SDC	Swiss Agency for Development and Cooperation (<i>Schweizer Direktion für Entwicklung und Zusammenarbeit</i>)
SDG	Sustainable Development Goal
SECO	Swiss State Secretariat for Economic Affairs (<i>Schweizer Staatssekretariat für Wirtschaft</i>)
SR	Systematic Review
STEP	Student Training for Entrepreneurial Promotion (Uganda)
TC	Technical Cooperation
TESDA	Technical Education and Skills Development Authority (Philippines)
ToC	Theory of Change
TVET	Technical and Vocational Education and Training
VET	Vocational Education and Training
VET project	Reform of Vocational Education and Training project (Serbia)
VSD	Vocational Skills Development
YEP	Youth Employment Promotion (Serbia)

CHAPTER 1 STUDY CONCEPT

1.1 INTRODUCTION

The aim of increasing cooperation between the German Federal Ministry for Economic Cooperation and Development (BMZ) and the Inter-American Development Bank (IDB) is to improve labor market oriented **skills development interventions in Latin America and the Caribbean (LAC)**. The objective is to contribute to improving the employability of TVET graduates and to increasing the supply of qualified and skilled personnel for sustainable, inclusive and green economic development in the region.

To support this objective, a study was conducted on behalf of the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). The present study focuses on the **approaches and methodologies for measuring the results and impact of skills development interventions** in German, Austrian and Swiss (DACH) development cooperation (DC), referred to as DACH DC throughout this report. Skills development interventions refer to formal measures, like Technical and Vocational Education and Training (TVET), as well as non-formal, informal and on-the-job measures to acquire productive capabilities.

The **objective of the study** is to examine the impact evaluations (IE) of projects, programs and sectoral schemes carried out by bilateral actors in DACH-DC, with the intention of identifying suitable IE designs for skills development interventions in LAC countries. The study focuses mainly on **impact evaluation designs** and presents a range of methodological approaches to assess the results of skills development measures in DACH DC. In this context, the study also addresses cross-cutting and focal topics such as: using existing monitoring and evaluation (M&E) or administrative data; analyzing gender-specific effects; conceptualize and measure the concepts “employability” and “entrepreneurship”; addressing the challenges of impact evaluations during the COVID-19 pandemic; examining the green transformation, and technological change (digital tools); incorporating the sustainability of impact; reviewing the efficiency and cost-effectiveness (incl. follow-the-money analysis and value-for-money); and private sector involvement. Therefore, this report reflects on the transferability of evaluation results for improvements in skills development interventions as well as their relevance for institutional learning.

Impact evaluations (IEs) refer to evaluation designs that try to measure the causal effect of an intervention (e.g., a skills development intervention) on an observed variable of interest, such as skills improvement, employability or income (German Institute for Development Evaluation, [DEval 2021](#)). IEs should ideally allow the formation of robust conclusions about the impact caused by an intervention and therefore causal attribution. According to the Development Assistance Committee of the Organization for Economic Co-operation and Development (OECD-DAC 2010), **impact** can be defined as a “positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended”. In the following report, the **superordinate term “IE”** is used for the three main design options in IEs, namely experimental, quasi-experimental and non-experimental IE designs (see *Annex 4 Definition of Key Concepts*). We use the term **“rigorous IE”** to describe the first two categories, namely experimental and quasi-experimental designs. Rigorous IE use a control group as the counterfactual situation to compare what happened due to the intervention and what would have happened without the interventions. This helps to identify if an intervention works and to establish causality (see *Chapter 2 Evaluation Designs and Approaches to Measure Results and Impacts of DC-Interventions*).

The **study report** is structured as follows: Chapter 1 presents the objective, scope and structure of the study as well as its methodology. Chapter 2 provides an overview of evaluation designs and approaches to measure results and impacts of DC-interventions. Chapter 3 summarizes evaluation policies and impact evaluation trends in DACH DC, focusing on German, Swiss and Austria’s bilateral DC actors. Chapter 4 contains the condensed analysis of nine case

Definitions

Skills Development refers to the productive capabilities acquired through all levels of learning and training, occurring in formal, non-formal, informal and on-the-job settings. It enables individuals to become fully and productively engaged in livelihoods, and to have the opportunity to adapt these capabilities to meet the changing demands and opportunities of the economy and labour market. The types of skills required for employment can be divided into: (I) **Basic and foundation skills** (acquired through the primary and secondary formal school system or through non-formal or informal learning processes); (II) **Transferable skills** (incl. the abilities to learn and adapt, solve problems, communicate ideas effectively, think critically and creatively and the ability to manage self and others); (III) **Technical and vocational skills** (specialized skills, knowledge or know-how to perform specific duties or tasks, mainly in a professional environment); and (IV) **Professional and personal skills** (incl. individual attributes relevant to work such as honesty, integrity, reliability, work ethic and judgement) ([SIDA 2018](#)).

studies of more or less rigorous IEs of skills development interventions in German DC¹ and presents two experimental, four quasi-experimental and three non-experimental designs. Chapter 5 provides a toolbox with reflection questions for practitioners who might evaluate their skills development interventions. Chapter 6 presents some strategies on how project planners, project implementers and policymakers can use the findings from impact evaluations in the skills development sector. Lastly, Chapter 7 summarizes the main findings, which lead to the general and specific conclusions on the focal topics and key recommendations for the LAC and other contexts.

1.2 METHODOLOGY

The study was prepared by a team of four international evaluation and TVET experts using the subsequent four-step methodology.

1.2.1 DESK RESEARCH

The research team conducted a **literature review of evaluation policies and trends** in DACH DC, focusing on the main bilateral DC actors from Germany, Switzerland and Austria and the existing **evaluation designs and approaches** in order to measure results and impacts of DC interventions. The respective findings are summarized in Chapter 2 and Chapter 3. To avoid missing information on evaluation policies of DACH donor organizations, interviews with evaluation departments of bilateral DACH DC organizations were conducted, namely the Swiss Agency for Development and Cooperation (SDC) and the Austrian Development Agency (ADA). The findings for German bilateral DC organizations, GIZ and KfW Development Bank, were generated from desk research. Therefore, the evaluation policies and trends within DACH DC reflect the evaluation teams understanding after desk research of publically available policies and a few additional interviews.

In addition, the research team conducted an intensive online search of publicly available (rigorous) IE reports that fulfilled the following **selection criteria** to some extent:

- **Availability and accessibility** of evaluation reports and interview partners
- **Topic:** TVET evaluations² or studies
- **Type and scope of impact evaluation methods** including: (i) methodological variation (quantitative and qualitative); (ii) contextual variation (e.g., project type/size, evaluation timing/costs); (iii) good example for learning
- **Relevant focal topics:**
 - ◆ using existing M&E or administrative data;
 - ◆ measuring gender-specific effects;
 - ◆ defining and measuring the concepts “employability” and “entrepreneurship”;
 - ◆ dealing with challenges of impact evaluations during the COVID-19 pandemic, green transformation, and technological change (digital tools);
 - ◆ studying the sustainability of impact;
 - ◆ analyzing the efficiency and cost-effectiveness (incl. value-for-money and the follow-the-money analysis); and
 - ◆ measuring private sector involvement and the transferability of evaluation results for learning and improvements in skills development interventions³.
- **Actors:** Mainly DC evaluations of DACH countries
- **Region:** Preferably LAC, but due to the absence of sufficient data, also other regions

Online research using these selection criteria led to the identification and pre-selection of **62 evaluation reports and studies**.

1.2.2 CASE STUDY SELECTION

The selection of **case studies** does not constitute a representative sample. Instead, the case studies were **purposefully selected** to show the **range of evaluation designs and approaches used in DC of DACH countries**. The 62 pre-selected evaluation reports or studies were scored according to the following more specific selection criteria, using seven questions and scores between zero to maximum six points per question. The **12 studies** with the highest scores (between 26-34 out of 38 possible points) were selected for further analysis and interviews, using the below mentioned questions, to guide the scoring of the selection criteria. Of those 12, **nine case studies are presented in this report**. Three studies were excluded from this report, because the authors of two studies were not available for an interview, which led to insufficient information and because the author of one study rated the underlying data quality as very low.

¹ This chapter focuses on German DC only. All other chapters refer to DACH DC.

² All selected evaluations deal with interventions in the field of vocational training in the broadest sense (see definition of TVET in Chapter 1.1). This means that interventions dedicated to employability or entrepreneurship education have been considered as well. Evaluations of projects, programs and sector projects were selected.

³ These focal topics were partially taken from the Terms of Reference and partially added to the list during the reflection workshop between the IDB, GIZ and the expert team on 9th August 2022. These focal topics make the findings more useful and align these to the needs of IDB, while increasing the scope of the study.

Detailed questions and scoring of TVET result and impact evaluations:

1. Is the evaluation report available (incl. detailed methodological information)? (*N/A = 0 points, low = 2 points, medium = 4 points, high = 6 points*)
2. Does the evaluation analyze the results or impact(s) of TVET projects/programs? (*N/A = 0 points, low = 2 points, medium = 4 points, high = 6 points*)
3. Was the evaluation implemented or contracted by a DACH evaluator or institute? (*N/A = 0 points, low = 2 points, medium = 4 points, high = 6 points*)
4. Does the impact evaluation contain an innovative or rigorous methodological approach (e.g., quantitative, qualitative or mixed-method impact evaluation methodological approach)? (*N/A = 0 points, low = 2 points, medium = 4 points, high = 6 points*)
5. Does the evaluation address any focal topic of interest (*see Subchapter 1.2.1*)? (*N/A = 0 points, low = 2 points, medium = 4 points, high = 6 points*)
6. Does the evaluation assess projects/programs implemented in LAC? (*no = 0 points / yes = 4 points*)
7. Does the evaluation assess DC projects? (*no = 0 points / yes = 4 points*)

For this selection exercise, a **main challenge and limitation** was that many potentially interesting **evaluation reports were not publicly available online**. In other cases, the research team was able to find information about the main technical findings of studies, but **detailed methodological information was not published** (e.g., due to confidentiality or to expected lack of interest from the side of readers regarding the evaluation methods used). Therefore, many potentially relevant evaluations were excluded from further analysis. If publicly available reports named relevant IE methods, but did not specify details, we requested access to additional internal reports describing the methodological details. An **additional challenge was the copyright on evaluation reports**. Externally contracted evaluation consultants are usually not allowed to share reports, because their contractors own the copyright and they are often reluctant to share their evaluation reports. Therefore, the research team did not get access to all underlying internal methodological reports, (e.g., for Case Study 9). Furthermore, **most IEs address only a few focal topics** and most DACH DC IE reports in the area of TVET **do not cover projects/programs located in LAC**. Therefore, the team decided to include IEs from other continents, since the IE methods are universal and can be applied to LAC contexts as well. The desk research and the application of the selection criteria revealed that DACH DC actors have conducted only a few rigorous IE in the TVET sector so far. Most DACH DC actors use non-experimental designs with qualitative, theory-based IE approaches, (especially contribution analysis), much more frequently than experimental or quasi-experimental designs with quantitative approaches.

1.2.3 ANALYSIS OF THE CASE STUDIES

The research team contacted the authors of the 12 selected studies and requested a semi-structured interview to gain further insights from the authors regarding the IEs (full list of persons interviewed and the interview guideline in *Annexes 2 and 3*) and to get access to non-published reports (especially those presenting methods used). Ten interviews with authors of case studies took place in total. For the triangulation of data, the evaluation report, the interview insights and in some cases further methodological reports, were analyzed and summarized and make up the **nine short case studies** which follow the same structure (*see Chapter 4 and Annex 1*).

1.2.4 COMPILATION

The descriptions of the ten different case studies—which comprise various types of IEs—present the whole range of quantitative and qualitative IE methodological approaches used in the sample (including rigorous IE, but not limited to rigorous IE), while providing relevant insights into focal or cross-cutting topics. The case studies are presented in clusters according to their evaluation design. Two experimental designs are presented, three quasi-experimental designs and at least⁴ four non-experimental evaluation designs (*see Chapter 4*).

Each case study follows the same structure: a brief description of the project including key data, project context and results; the evaluation objectives (incl. research questions, RQs) and indicators as well as the evaluation approach and methods used; the limitations of the evaluation; focal and/or cross-cutting topics; and key evaluation findings, conclusions on methods used and contact data for enquires. The case studies do not entirely reflect the projects or programs with all their activities and do not summarize the complete evaluation reports, but rather set priorities that are relevant to the objective of this study. These examples should enable learning for future evaluation studies.

Based upon the main findings of the nine case studies, the following is presented: a toolbox with reflection questions for the design of future IEs are presented (*see Chapter 5*); main strategies for using findings from IEs are identified (*see Chapter 6*); conclusions and lessons learnt are summarized and recommendations are formulated with specific view on the LAC context (*see Chapter 7*).

⁴ Qualitative approaches of non-experimental designs may be applied in the experimental and quasi-experimental designs to gain additional insights.

CHAPTER 2 EVALUATION DESIGNS AND APPROACHES TO MEASURE RESULTS AND IMPACTS OF DC-INTERVENTIONS

In 1991, the OECD/DAC established five criteria for the evaluation of DC-interventions, specifically identified as: relevance, effectiveness, efficiency, impact and sustainability. Since then, the criteria has served as the core reference for evaluative judgements of DC interventions for all its member countries, especially in the context of Official Development Assistance (ODA). At the end of 2019, the OECD/DAC finished a review and consultation process, which led to the inclusion of an additional criteria, coherence, to the previous five (see Figure 1). These six criteria constitute the current internationally agreed standard for evaluations. Nevertheless, most evaluation reports analyzed for the present study refer to assessment assignments conducted before the introduction of the sixth criteria. Furthermore, the present study focuses on the **measurement of results and impacts of interventions** (more precisely, the impact of TVET interventions), therefore, the analysis will mainly focus on two evaluation criteria:

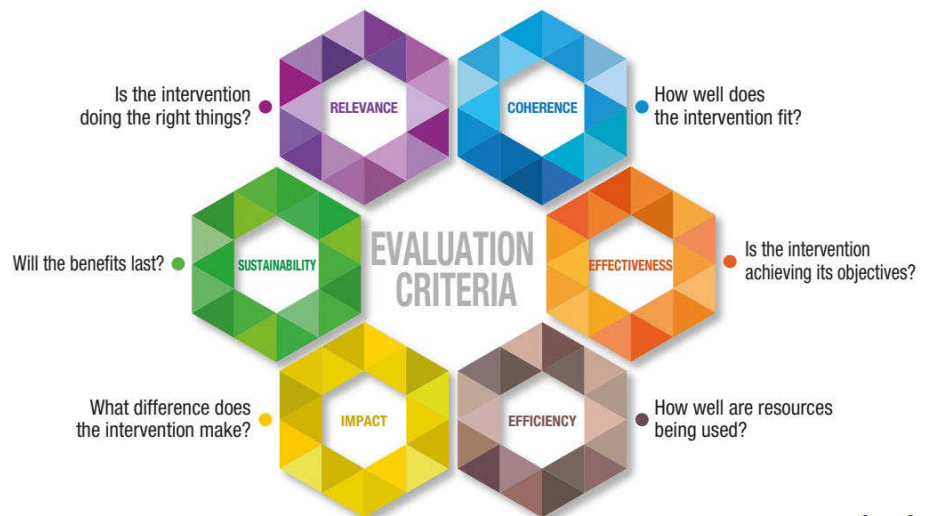


Figure 1: OECD/DAC evaluation criteria (OECD 2021)

- **Impact:** The extent to which the intervention has generated or is expected to generate significant higher-level or long-term effects (positive or negative, intended or unintended). Respective RQ: *What difference does the intervention make (in the long-term)?*
- **Effectiveness:** The extent to which the intervention achieved or is expected to achieve its objectives and its results. Respective RQ: *Is the intervention achieving its objectives?*

The overarching term **results** refers to outputs, outcomes and impacts. Following the theory of change (ToC) of DC projects, the intervention's medium- to long-term results (outcome and impact level) are of particularly interest. This study aims to understand the extent to which the intervention **outputs** (the immediate and concrete consequences of project activities), lead to planned **outcomes** (the direct to medium-term effects on beneficiaries), which is referred to as the **effectiveness criterion**. This study also examines the extent to which **outcomes lead to impacts**, which is referred to as the **impact criterion** (see Figure 2). Impacts refers to the consequences of outcomes, and therefore the long-term results and achievements, towards the overall objective.

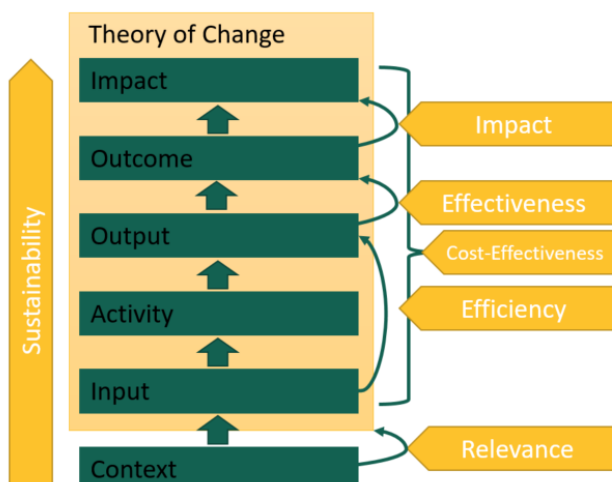


Figure 2: The OECD/DAC evaluation criteria and the theory of change (based on Sammeth et al. 2010)

Evaluation design is the overall strategy chosen for assessing, analyzing and estimating the causal results and impacts. Evaluation designs are classified as follows:

1. **Experimental designs:** Involves the random assignment of beneficiaries to the intervention (treatment) and control groups (non-intervention). These two groups are compared before and after the intervention (see Chapter 2.1).
2. **Quasi-experimental designs:** A comparison between the intervention (treatment) and control groups pre- and post-intervention, even though random assignment of beneficiaries was not possible (non-random assignment occurred) (see Chapter 2.2).
3. **Non-experimental designs:** Considers the extent to which changes have occurred only for those affected by the program or project, without using a comparison between treatment and control (non-treatment) groups (see Chapter 2.3).

The **evaluation approach** is the methodological approach which includes the method for data collection and analysis. Evaluation approaches may be **quantitative or qualitative**, or both. This offers a wide spectrum of methods for data collection and analysis (see Chapter 2.1-2.3) (ADA 2020).

Based on the literature review of evaluation policies (see Chapter 3) and on the interviews conducted, one **main finding** was identified: **The definition and understanding of rigorous IE differs in research and in practice.**

On the one hand, **(quantitative prone) researchers understand experimental and quasi-experimental designs as the core of rigorous IE designs** (DEval 2021). This research definition is in line with requirements for publishing research in peer-reviewed journals. DEval (2022c) concludes that systematic reviews (SRs) and evidence gap maps (EGMs) of existing rigorous IE are non-experimental designs, but are closely linked to rigorous IE since the underlying studies are rigorous IE.

On the other hand, **DACH DC practitioners prefer a more comprehensive definition of rigorous IE, which also includes various non-experimental or theory-based designs and, therefore, qualitative approaches.** For example, contribution analysis is a frequently used standard evaluation approach of DCs in Austria and Germany. Qualitative approaches have been more frequently used for the evaluation of DC interventions by DACH countries than quantitative approaches. The use of quantitative IE approaches require: a large number of observation units; longer observation periods (which start before the intervention itself); higher budgets for large-scale high-quality data collection and corresponding quantitative methodological knowledge in econometrics and statistics. Quantitative approaches are rarely used because of ethical concerns about randomization and because practitioners often prefer to be able to adjust the activities (treatment) as necessary during project implementation (in line with the adaptive, results-based management). From the point of view of DACH DC practitioners, rigorous is an approach that provides an appropriate methodology to answer the questions of interest.



Throughout the present chapter overview of methodologies used for experimental, quasi-experimental and non-experimental designs is presented.

2.1 EXPERIMENTAL DESIGNS

Experimental impact evaluation designs include different variations of randomized controlled trials (RCTs). In an RCT, units of observation from a population of interest (e.g., households) are randomly assigned to two groups: (1) the **intervention group**, which experiences a development intervention; (2) the **control group**, which does not experience the intervention. Random assignment of a **sufficiently large number of observational units** (e.g., households) ensures that the two groups are expected to be identical in terms of both their observable and non-observable characteristics prior to the intervention. The difference in the outcome of interest (e.g., employability and employment) between the two groups after the intervention thereby represents an undistorted estimate of the true effect of the intervention in this case, TVET participation (according to Duflo et al. 2008 and Gertler et al. 2016, cited in DEval 2022d). A prerequisite for this evaluation design is that there are sufficient units to be assigned to treatment and control groups and the intended effects have to be clear and stable enough for RCTs.

One of the **strengths** of RCT is that it provides the most powerful response to test causality. It is therefore referred to as the **“gold standard” of rigorous IE designs**, because it provides evaluators and program implementers with sufficient evidence to assert that the achievements observed are a result of the intervention and not of anything else. RCTs **eliminate the risk of selection bias** by randomly assigning subjects of observation (e.g., households, individuals, villages), to the treatment and control groups. Selection bias occurs when treatment and non-treatment (control) groups differ in characteristics (that may not be evident) and that may affect outcomes. Due to randomization, observable and unobservable characteristics of the population of interest are considered in the selection process.



Practical **challenges** could include: the large number of observable units (subjects) required; high costs (budget constraints); logistical difficulties; the **ex-ante timing of randomization**; ethical problems of ex-ante randomization (before an intervention takes place); political pressure; loss of follow-up; non-compliance and contaminations; and potential spillover effects of TVET programs. **RCTs are not suitable for ex-post evaluations** and the methodological hurdles mentioned above lead to the fact that there are few RCTs in DACH DC and in the TVET sector (White et al. 2014).

2.2 QUASI-EXPERIMENTAL DESIGNS


Similar to experimental designs, **quasi-experimental research designs** can test causal hypotheses. A quasi-experimental design identifies a control group that is as similar as possible to the intervention group in terms of baseline (pre-intervention) characteristics. By measuring the variable of interest in both the control and treatment groups, the control group states what the situation of the variable of interest (outcome, for the intervention group) would have been if the program or policy had not been implemented (i.e., counterfactual). The key difference between an experimental and quasi-experimental design is that the latter **lacks random assignment**, and the assignment often takes place **ex-post the intervention**. These quasi-experimental impact evaluation designs are less rigorous than experimental designs because quasi-experimental designs rely purely on observable characteristics and cannot take unobservable characteristics into account. Quasi-experimental designs include regression discontinuity designs, different matching techniques, difference-in-differences (DiD) estimation, interrupted time series, instrumental variable (IV) approaches and fixed effects models.

Table 1 presents an overview of different approaches to quasi-experimental evaluation designs with recommendations regarding usage as well as the advantages and disadvantages of each of them.

Table 1: When to use different quasi-experimental designs

QUASI-EXPERIMENTAL DESIGNS	USAGE	ADVANTAGE	DISADVANTAGE
Difference-in differences (DiD) See <i>Case Study 3, Serbia - Employment Impacts of German DC Interventions</i> and <i>Case Study 4, Kenya - Employment and Income Effects of Skills Development Interventions</i> 	<p>DiD compares the changes in outcomes over time between a population enrolled in a program (treatment group) and a population that is not (control group) taking two differences into account: Firstly, DiD takes the before-after difference in treatment group's outcomes. In comparing the same group to itself, the first difference controls for factors that are constant over time in this group. Secondly, to capture time-varying factors, DiD takes the before-after difference in the control group, which was exposed to the same environmental conditions as the treatment group. Finally, DiD "cleans" all time-varying factors from the first difference by subtracting the second difference from it, which enables an impact estimation (World Bank 2022a). DiD is used when two groups are growing at similar rates and when baseline and follow-up data is available.</p>	<ul style="list-style-type: none"> • Eliminates fixed differences not related to treatment 	<ul style="list-style-type: none"> • Can be biased if trends change; two pre-intervention periods of data are ideal • Relies on observable data • Parallel or equal trend assumption has to hold: external factors (unobservable characteristics) do not have different effects on the outcome variable of the treatment and control groups
Matching See <i>Case Study 6, Philippines - Dual Vocational Training</i> and <i>Case Study 3, Serbia - Employment Impacts of German DC Interventions</i> 	<p>Matching methods are statistical techniques to construct an artificial control group by matching each treated unit with a non-treated unit of similar characteristics. Matching methods aim to equate (or "balance") the distribution of co-variables (independent variable that can influence the outcome of a given statistical trial, which are not of direct interest) in the treatment and control groups. Matching methods are used when other methods are not possible (World Bank 2022b), and sometimes combined with other designs (e.g., DiD).</p>	<ul style="list-style-type: none"> • Overcomes observed differences between treatment and control units • Does not require parallel trends 	<ul style="list-style-type: none"> • Assumes no unobserved differences between the treatment units and the matched control units (often implausible) • Requires all confounders to be balanced between the two groups • Matching requires extensive datasets with information on treated and non-treated units' characteristics before the treatment

QUASI-EXPERIMENTAL DESIGNS	USAGE	ADVANTAGE	DISADVANTAGE
Randomized promotion design/ randomized offering/ encouragement design	Randomized promotion designs are used when an intervention is universally implemented (nobody can be excluded from receiving the treatment) and one cannot choose who gets the program (e.g. national projects open to all) (CEGA/University of Berkeley 2017).	<ul style="list-style-type: none"> Provides exogenous variation for a subset of beneficiaries 	<ul style="list-style-type: none"> Only looks at sub-group of samples Power of encouragement design only known ex-post Selection bias will influence (reduce) the internal validity, because the decision to enroll is correlated with observable and unobservable characteristics
Regression discontinuity design	Regression discontinuity designs are used when potential treatments are designed around an essentially arbitrary cutoff (e.g., clear, sharp assignment rule), where those above the threshold receive the treatment and those below the threshold do not receive the treatment. The differences between the two groups near this threshold are often very minimal or nearly non-existent, so that the individuals just below the threshold are used as a control group and those just above as a treatment group (World Bank 2022c).	<ul style="list-style-type: none"> Project beneficiaries often must qualify through established criteria (clear, sharp and arbitrary assignment rules/cutoffs) 	<ul style="list-style-type: none"> Only look at sub-group of samples, assignment rule in practice often not implemented strictly
Interrupted time series / quasi-experimental time series analysis	Interrupted time series involves collecting data at multiple, equally spaced time points over a long-term period before and after a point of intervention to assess the intervention's effects. The time series refers to the data over the period, while the interruption is the intervention, which is one or multiple controlled external influences. Effects of the intervention are evaluated by changes in the level and slope of the time series and statistical significance of the intervention parameters (Hudson, J., Fielding, S., Ramsay, C. 2019).	<ul style="list-style-type: none"> Detects changes that are delayed or intermittent Determines if the change is permanent or temporary Allows evaluation of variables that are changing before the intervention (e.g., comparing slopes of trend lines before and after the intervention) Historical data can be used Makes it easier to control for confounding variables and regression to the mean Can be conducted with a small sample size 	<ul style="list-style-type: none"> Determining whether a change noted is due to the intervention or to other factors, such as another event occurring at a similar time to the intervention Cyclical changes (e.g. seasons) may be overlooked if the number of observations or interval between observations is not great enough. These challenges can be overcome by some variations, such as a non-intervention control group, removal of the intervention, or applying the intervention to two or more groups at different times. These steps increase the cost and time required for making large number of observations, which may be problematic

QUASI-EXPERIMENTAL DESIGNS	USAGE	ADVANTAGE	DISADVANTAGE
Instrumental variables (IV)	Instrumental variables are a valid instrument to overcome endogeneity (i.e., omitted variables, measurement error, or simultaneity) when measuring causal impact (World Bank 2022d).	<ul style="list-style-type: none"> Valid instrument to overcome endogeneity when RCTs (which ensure exogeneity) are not logistically or ethically feasible or ordinary least squares regression leads to inconsistent estimates (due to endogeneity). 	<ul style="list-style-type: none"> Use an IV that is strongly correlated with the exposure, because the IV estimator will be imprecise (large standard error) and biased when the sample size is small, and biased in large samples when one of the assumptions is violated slightly
Natural experiment / natural randomized experiment <i>See Case Study 5, Brazil-Non-cognitive Skills and Labor Market Outcomes</i> 	In natural experiments, an external event or situation (“nature”) leads to a random or random-like assignment of people to the treatment group. Researchers have no control over the independent variable, but they can study the effect of the treatment. Natural experiments are not considered to be true experiments (even though some use random assignment), because these are observed in nature and not in laboratory or field experiments.	<ul style="list-style-type: none"> Allow research that is otherwise unethical (external events cause treatment and control group) Very little bias from sampling or demand characteristics High ecological validity - resulting in many real-world applications 	<ul style="list-style-type: none"> Difficult to infer cause and effect due to lack of control and no direct manipulation of the intervention Extremely difficult to replicate - difficult to test for reliability Many extraneous variables may threaten the validity (no control over confounding variables) Participants may be aware of being studies causing participant effects, investigator effects and demand characteristics

2.3 NON-EXPERIMENTAL DESIGNS

There are multiple non-experimental impact evaluation designs using **qualitative impact evaluation approaches**. These **non-experimental IE designs are less rigorous** than experimental or quasi-experimental designs using quantitative IE approaches. Most of these qualitative approaches are based on a theory, which states the results hypothesis or how project activities determine the outcomes and impacts of an intervention. In a theory-based impact evaluation, all steps and underlying assumptions of the causal chain, which link activities and outcomes, are spelled out and tested. Table 2 below contains the description of the five most commonly used non-experimental designs as well as **other qualitative data collection and analysis methods and techniques used in the TVET sector**.

Table 2: Description of non-experimental designs

NON-EXPERIMENTAL DESIGNS	DESCRIPTION
Systematic reviews (SR) of rigorous IEs	Systematic reviews summarize and synthesize the existing evidence of many available rigorous IE to answer specific RQs. The existing evidence from rigorous IE studies has to fulfil minimum standards of scientific rigor. SRs apply clear criteria for the study inclusion and exclusion, explicit and transparent search strategies, systematic procedures for data extraction and offer a critical analysis. A meta-analysis is conducted identifying average effects of interventions from a large number of quantitative studies (DEval 2022e). In cases of a sufficiently large evidence base, SRs of rigorous IE give the best possible generalizable statements about what is known about interventions and can reliably inform decision-makers which interventions work and why (Waddington et al. 2012).
Evidence gap maps (EGM)	Evidence gap maps are visual compilations of (rigorous) evidence on the impact of policies and programs in a specific sector or thematic area (using SRs or single scientific studies, mainly rigorous IE). EGMs help decision-makers obtain an overview of where sufficient evidence is available to be used and where evidence needs to be generated because of evidence gaps (few or no existing studies) (DEval 2021).

NON-EXPERIMENTAL DESIGNS	DESCRIPTION
Contribution analysis  See <i>Case Study 7, Global - Skills for Reintegration and Case Study 8, Egypt - Employment Promotion</i>	<p>Contribution analysis is a theory-based approach used for assessing causal questions and inferring causality in program evaluations. Based on a ToC, a step-by-step approach⁵ is used to arrive at plausible conclusions (with some level of confidence but no definitive proof), about the contribution of programs to particular outcomes in the past or present. It helps to reduce uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not) and the roles played by the intervention and other internal and external factors. Contribution analysis is used in non-experimental context when a relatively clear ToC exists (or can be created or revised) and in case there is little or no scope for varying how the program is implemented (Better Evaluation 2013).</p>
Realist evaluation	<p>Realist evaluation is a form of theory-driven evaluation with a philosophical underpinning. Realist evaluations answer the question: “<i>What works, for whom, in what respects, to what extent, in what contexts, and how?</i>” In order to answer these questions, realist evaluators aim to identify the underlying mechanisms that explain ‘how’ the outcomes were caused and the influence of context (Better Evaluation 2016a).</p>
Process-tracing	<p>Process tracing is a case-based approach to causal inference which focuses on the use of clues within a case (causal-process observations, CPOs) to adjudicate among alternative possible explanations. Process tracing can be used to see if results are consistent with the program’s ToC and, to see if alternative explanations can be ruled out. Process tracing involves four types of causal tests⁶, namely: ‘straw in the wind’, ‘hoop’, ‘smoking gun’, and ‘doubly definitive’ (Better Evaluation 2016b).</p>
Other qualitative data collection and analysis approaches and methods in the TVET sector:	
Kirkpatrick Model (KM) or Kirkpatrick’s Four Levels of Training Evaluation  See <i>Case Study 7, Global - Skills for Reintegration</i>	<p>The Kirkpatrick Model is a tool for evaluating and analyzing the results of educational, training and learning programs. It consists of the following four levels: 1. Understand reaction of learners to the training measure (e.g., acceptance, satisfaction, use or usefulness). 2. Analyze learning success of participants (e.g., improving knowledge and skills of the learners through the teaching activity/learning materials). 3. Studying behavior of the learners in their everyday life to determine the transfer performance. 4. Analyze results to evaluate the effectiveness of the measure at the (individual and) organizational level. The KM assumes that each successive evaluation level is based on the information provided by the lower levels, so that evaluators analyze each level one after the other (Kirkpatrick Partners 2022).</p>
Outcome mapping	<p>Outcome Mapping is a qualitative and participatory approach used to develop a system to record the (qualitative) effects of projects/programs by focusing on changes in people’s behavior. Therefore, “outcomes” are behavioral changes in direct partners with whom the project is working (called “boundary partners”). Thus, outcome mapping applies an alternative understanding of outcomes and does not focus on linked project outputs and their effects on the target groups like conventional impact assessment methods. Outcome mapping proposes practical instruments for project planning and recording project/program progress. It is used for learning and self-evaluation (Outcome Mapping Learning Community 2022).</p>
Most significant change (MSC)  See <i>Case Study 7, Global Skills for Reintegration and Case Study 8, Egypt - Employment Promotion</i>	<p>Most significant changes is a qualitative and participatory technique for recording the effects of projects/programs. Participants are asked to describe the most important change that has happened, from their perspective, as a result of the project/program. These individual experiences are systematically analyzed for evaluating the impact of a project/program (or for ongoing monitoring). MSC can be used when no baseline data or indicators are available, because “data” about projects/programs outcomes and impacts (incl. unexpected effects) is generated (Davies and Dart 2005).</p>

⁵ The six steps approach for credible contribution analysis: 1. Set out the attribution problem to be addressed; 2. Develop a ToC and risks to it; 3. Gather the existing evidence on the ToC; 4. Assemble and assess the contribution story, or performance story, and challenges to it; 5. Seek out additional evidence; 6. Revise and (ideally) strengthen the contribution story.

⁶ The four types of causal tests consist of: 1. ‘Straw in the wind’ which lends support for an explanation without definitively ruling it in or out. 2. ‘Hoop’ tests which fail when the examination of a case shows the presence of a necessary causal condition, when the outcome of interest is not present. Common hoop conditions are more persuasive than uncommon ones. 3. A ‘Smoking gun’ test is passed when the examination of a case shows the presence of a sufficient causal condition. Uncommon smoking gun conditions are more persuasive than common ones. 4. A ‘Doubly definitive’ test is passed when examination of a case shows that a condition is both necessary and sufficient support for the explanation. These tend to be rare.

NON-EXPERIMENTAL DESIGNS	DESCRIPTION
Method for impact assessment of programs and projects (MAPP)	<p>Method for impact assessment of programs and projects is also a qualitative and participatory impact analysis method for recording the effects of projects/programs (incl. multi-dimensional development schemes). Groups discuss and analyze the effects and developments of a program retrospectively. MAPP uses a fixed sequence of six to eight interrelated instruments for the group discussion to be able to come up with a robust assessment of changes and to assign impacts to measures (incl. intended and unintended impacts). Firstly, the group analyses the effect of the project in general and then in detail, by means of various self-defined indicators. Next, the relevant project measures and activities (and additional actors) are listed and prioritized. Finally, the group assesses the contribution made by the individual development measures to the observed developments (Neubert 2010).</p>

2.4 COVID-19-RELATED CHALLENGES AND SOLUTIONS FOR IE

Many evaluation tools rely on the on-site work of evaluators. Participative approaches are crucial to better understand projects and their stakeholders. However, the **COVID-19 pandemic** has forced a shift to **remote work**, including evaluations of DC-interventions. After more than two years of virtual work due to the pandemic, things are slowly getting back to normal, it is safe to assume that the COVID-19 pandemic will not remain the only challenge for evaluations. For example, risks and hazards arising from **climate change** impacts like floods, storms, landslides, that make intervention regions inaccessible or, in case of extreme fragility and/or violence, too risky. These factors may make travel and on-site evaluation of DC projects and programs more difficult in the future. In addition, contributions to a **green transformation** (a primary goal of Agenda 2030) includes a commitment to reduce CO²-emissions, and thus mobility, which may lead to more evaluations being performed remotely. Effective **expectation management** is important when planning such remote evaluations. Both, the project and the evaluation team need to be very clear in advance about the limitations as well as the opportunities and advantages of remote evaluations. Close and transparent **cooperation and communication** between international and national/regional evaluators and with the project staff becomes even more important when international evaluators are unable to visit the project area and must conduct virtual interviews or rely entirely on their national colleagues. For a successful evaluation, the communication processes, channels, formats and frequency must be agreed in advance; values and standards must be stated and transparent to all, and the roles and responsibilities of the team members must be clear and supportive. All the data collection can take place remotely using **digital tools**, if corresponding surveys are sent out as a web link and are accessed online on various devices (tablets, mobile phones and computers). To maintain a sound evidence basis for decision-makers, some institutions increased their efforts in **making better use of existing secondary sources**. This involves the integration of multiple data sources (big data), which are processed through Artificial Intelligence systems (European Commission, [EC 2021](#)).

For interviews and field studies, stronger adaptations are necessary. Here are some recommendations:

- Individual interviews can be conducted via digital communication tools (Zoom, Teams, Skype, WebEx...), but adequate internet access and connectivity is a prerequisite.
- Focus group discussions (FGD) can be implemented remotely using the above-mentioned tools as well. If any individuals have poor internet access, the project facilities should be available to them.
- In the context of interviews and FGD, national and regional consultants have become more pivotal. They have easier access to interview partners and should be effectively involved in the evaluation.
- Some staff might be less experienced in the usage of digital tools (e.g., due to their age, education or limited availability and usage of certain software in some countries or regions), so it could be necessary to provide specific training at an early stage.
- The active involvement of regional staff can obviate the need for translation. Their regional and cultural knowledge should be used in a sensitive manner.
- On-site surveys/field visits can be replaced by video calls. The local evaluator can make a tour with simple means, like walking around with a cell phone, and provide a visual impression of the situation on-site.
- Face-to-face interviews with large groups and control groups must be conducted with field staff. Interviews should be recorded and analyzed by the evaluators.
- In a situation of crisis where it is not always possible to have face-to-face consultations, data can be gathered remotely by monitoring social media platforms (e.g. Twitter, Facebook, Instagram) or by conducting a robust Social Network Analysis ([EC 2021](#)).

2.5 DIGITAL TOOLS FOR IE

Intensified by the pandemic and climate change challenges, **digital tools and technologies** are gaining importance in the context of evaluations. For example, GIZ has developed a new support service called the **data service center**⁷, available to programs and projects as well as for evaluation teams to support them in the design and conduction of evaluations using data sources (e.g. geo-data). One aspect that needs to be considered when using digital tools for evaluations is **data protection**. Further, national and regional (e.g. EU) data protection regulations are not homogeneous and must be taken into account. With this in mind, GIZ has established a support service and a checklist for evaluators to ensure that the **do-no-harm principle** and the organization's reputation are safeguarded. Another aspect that needs to be taken into account is **cost-efficiency**. It was initially assumed that the use of digital tools would reduce the cost of evaluations (e.g., less costs for airplane tickets, accommodation of consultants, etc.) but the reality shows that the **replacement of on-site visits with online conducted evaluation activities** (e.g., interviews or FGD) **leads to other costs** (e.g., increased consultants' work time, specific expertise, etc.) that needs to be considered to ensure the quality of the evaluation results.

The following points list some of the advantages and good reasons for IE to use various digital tools.

- **Virtual meetings** may be helpful and efficient in the organization phase of an evaluation, as they allow for short and frequent meetings between the evaluation team and the contractor (incl. the evaluation kick-off). They can also facilitate: stakeholder discussions (e.g., meetings with project staff and stakeholders located in different project areas, jour-fixed meetings between evaluation team and the project or contractor, consultation with other donors in the case of co-financing, or subsequent dissemination of findings to a large audience); interviewer trainings and data collections (e.g., surveys, interviews and FGD); and desk research (if access to sufficient and adequate data sources is provided). These virtual formats reduce the need for travel and related costs.
- **Data collection and storage** being moved away from paper-based formats and towards computer or web-based formats using adequate software on computers, tablets or smartphones. Questionnaires and surveys can be made available online. This may reduce data collection costs but requires adequate advertisement and expertise.
- **Data analysis and visualization** can be conducted jointly by members of the evaluation team based in different locations using various quantitative and qualitative tools. This requires the corresponding technical and methodological skills and may increase associated costs.
- **Publications** of evaluation findings and studies can be made available online via donor-specific websites like the [KfW](#) or [GIZ](#) evaluation databases or more academic databases like the [international 3ie's Development Evidence Portal \(DEP\)](#) and [German rigorous evidence database \(RED\)](#) for rigorous IE.

⁷ Source: GIZ Consultants Day / KerngutachterInnen Tag 14.11.2022

CHAPTER 3 EVALUATION POLICIES AND TRENDS WITHIN DACH DC

This chapter examines evaluation trends in DACH DC and summarizes the research teams understanding of publically available evaluation policies of the following actors:

- The German Federal Ministry for Economic Cooperation and Development (Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, BMZ), Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) and German Development Bank (Kreditanstalt für Wiederaufbau, KfW) as well as research trends regarding RIE in German DC by the German Institute for Development Evaluation (Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, DEval) (see Chapter 3.1)
- The Swiss Agency for Development and Cooperation (SDC) and the Swiss State Secretariat for Economic Affairs (SECO) (see Chapter 3.2)
- The Austrian Development Agency (ADA) and the Development Bank of Austria (*Österreichische Entwicklungsbank AG, OeEB*) (see Chapter 3.3).

3.1 GERMANY'S DC EVALUATION POLICY AND TRENDS

In brief, in order to assess the lasting developmental results and impacts of DC measures and derive lessons for support activities, the German implementing organizations (GIZ and KfW), DEval and the German ministries (especially BMZ) carry out evaluations. The evaluation policies, systems and evaluation types, and designs and approaches of GIZ and KfW are summarized in Chapter 3.1.1 and 3.2.2. All German DC evaluations follow the **criteria and standards for independent evaluations** set by the OECD/DAC. Cooperation partners, relevant stakeholders and the target group in partner countries should be involved in the evaluation process. **Evaluation reports should be shared with partner countries and the public** (if there are no overriding reasons not to do so). DEval, BMZ and other German ministries evaluate cross-cutting issues and large-scale programs ([BMZ 2021a](#)).

The importance and number of rigorous experimental and quasi-experimental impact evaluations are increasing in German DC and this trend is expected to continue in the following years due to multiple ongoing and unpublished rigorous IE. A new funding program and a new German rigorous evidence database (RED) support the implementation of rigorous IE and the sharing of rigorous IE findings. Globally, there have been only few quantitative, experimental or quasi-experimental rigorous IE designs in LAC. To date, German DC lacks rigorous IE experience in the TVET sector in LAC.

The German DC Evaluation Policy and System:

The BMZ guidelines for bilateral financial cooperation (FC) and technical cooperation (TC) include information on the BMZ evaluation strategy ([BMZ 2021a](#)). BMZ conducts its own evaluations as well as evaluations via DEval and the German implementing organizations – GIZ and KfW. These evaluations and studies shape the BMZ's continuous policy dialogue with other donor governments and the corresponding partner country. All evaluations commissioned by BMZ follow the **OECD/DAC criteria and standards for independent evaluations**. The cooperation partners, the target group and other relevant stakeholders in the partner countries should be involved in the evaluation process. Ideally, these **evaluation reports should be shared with partner countries and the public**, if there are no overriding reasons not to do so.

The **German implementing organizations supervise and forward all final and ex-post evaluation reports** to the German government (mainly BMZ). In addition, the **German government carries out its own evaluations**, especially of cross-cutting issues and larger-scale programs. These evaluations by the German government support development activities and aim at gaining insights for the design of future DC support. **DEval carries out its own independent analyses and evaluations** of Germany's DC work ([BMZ 2021a](#)) and BMZ responses to each DEval evaluation. These responses reflect the importance that the BMZ attaches to accountability and enhancing aid effectiveness. A short version of those responses is published on BMZ's website ([BMZ 2022](#)).

German DC Evaluation Trends:

The importance of (rigorous) IEs is increasing in Germany's DC ([GIZ 2018](#)). A recent research project on rigorous IE implemented by DEval for the period of 2014-2020 showed that **GIZ, KfW and German civil society organizations (CSOs) have all been involved in rigorous IEs** (either as project implementer or by financing the rigorous IE). With a total of 58 rigorous IEs, GIZ reported the highest number of rigorous IEs, followed by CSOs with 23 rigorous IEs, and KfW with 12 rigorous IEs. DEval has conducted four rigorous IEs in the relevant period (see Figure 3, [DEval 2021, p. 17](#), and see Chapter 2 for information on rigorous IE designs and approaches). The DEval research study expects that the **number of rigorous IEs will increase** in the period after the study, because multiple rigorous IEs were in process,

planned or approved (but not yet finished) at the time of the study (see Figure 3)⁸. DEval studied the **regional distribution of rigorous IEs worldwide** by searching the 3ie's DEP, which is a global platform and not limited only to German-DC. They found that Sub-Saharan Africa is by far the region with most rigorous IEs (71%). Particularly, many rigorous IE were found in Eastern Africa in countries like Uganda, Ethiopia, Malawi and Zambia. They found significantly fewer rigorous IE in the Middle Eastern and North African region (9%) and in Southern and Eastern Asia and the Pacific (11%), in Europe and Central Asia (5%), in LAC (3%) and in North America (1%) (DEval 2021, p.19).

As part of the above-mentioned research project, DEval created the RED⁹, which contains a collection of rigorous IEs conducted by German DC and is publicly accessible on the DEval webpage. However, the selection on this database reveals a **lack of German TVET rigorous IE experience in the LAC region so far** (DEval 2022a).

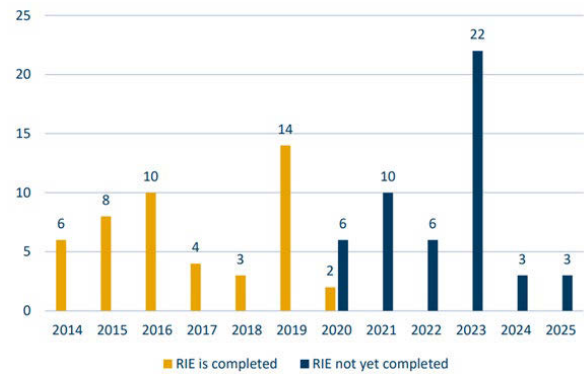


Figure 3: Number of rigorous impact evaluations over time

Further Relevant Initiatives/Partnerships, Projects and Trends Regarding Rigorous IE:

New DEval Funding Program

In 2022, a **new funding program for rigorous IE** (financed by BMZ), was publicly launched, which aims help to anchor rigorous IE more firmly in Germany's DC and thus increase the effectiveness of Germany's DC. DEval is responsible for offering this funding program and identifies projects suitable for rigorous IE. The new program promotes knowledge exchange between scientific research and DC practitioners via **matchmaking between German DC projects and interested German scientific institutions**. The funding program will **financially support approximately nine rigorous IEs with up to 363,000 USD¹⁰ for each rigorous IE from 2023 to 2025**. Rigorous IEs that evaluate DC interventions funded by BMZ, including CSO-led projects, are eligible for funding (DEval 2022b).

Focelac+ for Evaluation and Learning in LAC

The aim of the DEval implemented project Focelac+ (*Fomento de una cultura de evaluación y de aprendizaje en América Latina con proyección global/Strengthening a Culture of Evaluation and Learning in Latin America with a Global Outlook*), states that "evaluations in selected countries (especially in Latin America), make a greater contribution to accountability, transparency and learning. (...) The **project partner is the Costa Rican Ministry of Planning, Mideplan**. The long-standing advisory services and capacity building of its evaluation unit in the framework of the predecessor projects Foceval [2011-2014] and Focelac [2019-2020] has made it possible for **Mideplan to engage as a multiplier for strengthening national evaluation capacities in the region**. (...) In this regard, the project promotes not only the competencies of the various stakeholders involved in the evaluation system, but also the exchange and cooperation with one another. (...) To promote systemic change, Focelac+ works together with various international early childhood development (ECD) stakeholders that join forces, for example, in the Global Evaluation Initiative (GEI)" (DEval 2022c).

Switzerland, Austria and Germany support the **Global Partnership for Effective Development Co-operation (GPEDC)** as the primary multi-stakeholder vehicle for driving development effectiveness, to "maximize the effectiveness of all forms of co-operation for development for the shared benefits of people, planet, prosperity and peace." The implementation of the four effectiveness principles¹¹ (namely, **country ownership, focus on results, inclusive partnerships, transparency and mutual accountability**), shall re-build partnerships on a more equitable basis and for more sustainable results (UN 2022).

In the following two subchapters, the focus is specifically on the evaluation policies and systems applied by Germany's technical DC, namely GIZ (see Chapter 3.1.1) and Germany's financial DC, namely KfW (see Chapter 3.1.2), who are committed to the principles and standards of evaluation of the German government, especially BMZ.

⁸ Source: stocktaking survey; questions include the expected start of endline data collection and status of rigorous IE; N=97.

⁹ This effort is closely linked to the 3ie's DEP that includes more than 10,000 rigorous studies on what works in international DC. RED does not aim to replace the DEP.

¹⁰ This is equivalent to 365,000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

¹¹ As agreed in the Nairobi Document by more than 161 countries and 56 organizations in 2011.

3.1.1 GERMANY'S GIZ EVALUATION POLICY

In brief, *Deutsche Gesellschaft für Internationale Zusammenarbeit* (GIZ) has its own evaluation system and policy ([GIZ 2018](#)), which states the aim, scope and approaches of evaluations. GIZ's guiding evaluation principle is **results-orientation**, which means that GIZ measure "changes that can be attributed to a project or object of an evaluation" (results) and identify a clear and plausible causal link from the project activities to actual project results. GIZ highlights the **utilization focus** of their evaluations, so that these are "done for and with specific intended primary users for specific, intended use." GIZ conducts **different types of evaluations** of their own projects, which are externally assessed on behalf of its contractors, especially BMZ.

According to GIZ, **rigorous IE designs** and approaches are not limited to experimental IE design. Rigorous IE designs and approaches include the whole range of "experimental, quasi-experimental, statistical, theory-based and participatory approaches." This means that GIZ considers experimental, quasi-experimental and non-experimental designs as rigorous IEs. On the one hand side, GIZ acknowledges a **rising interest and increasing use of quasi-experimental and experimental IE designs** within the organization. GIZ applies quasi-experimental designs in collaboration with research institutes and consulting companies (see *Case Study 2 and 4*). However, GIZ's **main external (ex-post) evaluation tool is central project evaluations** (CPE), which use non-experimental designs and qualitative, theory-based approaches (like contribution analysis, realist evaluations or process tracing). **Contribution analysis** is the standard and most frequently used IE approach (see *Case Study 7 and 8*). An explanation for this is the utilization focus and results-oriented guiding principles of GIZ.

GIZ is one of the main implementing organizations of Germany's ODA. This organization is a publicly owned company that operates internationally in the field of TC¹² on behalf of various German ministries, but mainly BMZ. It implements TC interventions in around 120 partner countries around the world, focusing on different sectors and areas of interest, according to the priorities of the German government. GIZ has a long experience supporting its partner countries in developing and implementing strategies and policies for TVET. Furthermore, TVET is a central pillar for sustainable and viable economic development and for achieving the Sustainable Development Goals (SDGs). Thus, it also constitutes a central area of action for GIZ ([GIZ 2022a](#), [GIZ 2022b](#)).

GIZ's Evaluation Policy

Results-orientation is the guiding principle of any DC-intervention implemented by GIZ since the 2005 Paris Declaration on Aid Effectiveness. According to GIZ, "results are understood as the changes that can be attributed to a project or object of an evaluation," meaning that a causal link must be clearly or plausibly identified to deem observed changes as actual results. Any evaluation of GIZ is based on a results model of the intervention, which provides a graphical representation of the ToC ([GIZ 2018](#)). GIZ typically implements interventions at all levels (employing a holistic approach), which makes the attribution of results more complex and challenging. A bilateral TVET program normally addresses all three levels: the macro-level (e.g., policy dialogue, labor laws, economic development); the meso-level (e.g., institutional strengthening, knowledge-management, networking, curricula, and standards); and the micro-level (teacher training, pilot workshops, etc.). Very few interventions are limited to the micro level. Therefore, measuring the impact of a TVET intervention needs to address all three levels of intervention.

GIZ applies the evaluation criteria for German DC based on the **six OECD/DAC criteria** and adds the additional criteria **complementarity and coordination** to these. CPEs also examine the **quality of implementation**. The evaluation policy of GIZ describes the **utilization focus as a key feature of evaluations**, which means that evaluations are "done for and with specific intended primary users for specific, intended uses. (...) Use concerns how real people in the real world apply evaluation findings and experience the evaluation process" (M. Patron on utilization-focused evaluations as cited in [GIZ 2018](#), p. 6).

GIZ defines three key functions of evaluations ([GIZ 2018](#), p. 6):

- **Support for evidence-based decisions**
- **Transparency and accountability**
- **Organizational learning, including its contribution to knowledge management**

GIZ distinguishes between different types of evaluations ([GIZ 2018](#)). **Central evaluations** are steered by the GIZ evaluation unit and **decentral evaluations** are steered by the respective project-managing organizational units:

Central evaluations steered by the evaluation unit include:

- **CPEs for BMZ**
- **Corporate strategic evaluations** on behalf of the managing board on service delivery or corporate development
- **Cross-section evaluations** cover meta-evaluations of CPEs, contracting evaluations and decentralized evaluations
- **Contracting evaluations** for German public sector clients, international services and internal parties

¹² TC consists primarily of advice and support to capacity development in partner countries' institutions and organizations. It provides, development services, supports the establishment and promotion of project sponsors, the provision of equipment and material and the preparation of studies and reports. Financial contributions are relatively small, as this is the field of the German FC.

Decentralized evaluations steered by the commissioning project offers and advised by the evaluation unit:

- **Decentralized evaluations of projects and measures** for BMZ and others, (e.g., German public sector clients or international services)

GIZ's Evaluation System

The **evaluation unit** is responsible for evaluation system and methods, the organization and contracting of evaluation assignments and the quality assurance. It reports to the management board to safeguard its independence and supports institutional evaluation policy development. Cooperation with **external evaluation experts** increases and improves the transparency as a foundation for **evidence-based decision-making** processes. IE findings are used for early dialogue with BMZ and learning processes, which should contribute to effective knowledge management as to better plan new intervention strategies ([GIZ 2018](#)).

For GIZ, the **evaluation system** contributes to measuring impacts better and by doing so improving the projects' overall quality. Due to the incorporation of additional evaluation questions, the impacts related to the SDGs of the 2030 Agenda are identified ([GIZ 2018](#)). It should be noted that **GIZ is constantly reforming its implementation system** and new instruments do not always link well to past practice. For example, measurement tools for efficiency require a result-based accounting tool, which was recently introduced, meaning most of the evaluated projects did not follow the new system yet.

Types of Evaluations Conducted by GIZ (Especially IE Designs, Approaches and Methods)

GIZ conducts multiple types of evaluations (see Subchapter 3.1.1 GIZ's Evaluation Policy above). Concerning the **evaluation design and methodological approaches for data**, GIZ aims to use an appropriate mix of quantitative and qualitative empirical social research methods. GIZ acknowledges the **increasing importance of IEs** which do not only capture results but provide clear evidence (attribution) or plausible evidence (association) of a causal relationship between measures and results and that preclude the effects of other external influencing factors. This **requires a theoretically sound and verifiable (rigorous) methodological approach**. According to GIZ, "a rigorous approach includes not just experimental evaluation designs but also any methodological approach that systematically deals with the attribution of results to measures. These include experimental, quasi-experimental, statistical, theory-based and participatory approaches" ([GIZ 2018](#), p. 10). The GIZ evaluation policy mentions **quasi-experimental and counterfactual experimental** (i.e., RCT designs) for measuring results. They acknowledge the **need and rising interest for these designs to measure results**. RCTs are used to examine the impact of innovative interventions (e.g., pilot projects to decide about the effectiveness for up scaling), or large-scale projects of political relevance. Both, the organizational units managing the project and the evaluation unit, must agree on the use of experimental IE designs. GIZ stipulates that **non-experimental IE designs using qualitative, theory-based approaches (namely, realist evaluation, process tracing and contribution analysis)**, are used as standardized IE design and methods to ensure the robust verification of results in CPEs ([GIZ 2018](#), p. 10-11).

3.1.2 GERMANY'S KfW EVALUATION POLICY

In brief, the **primary purpose of evaluations** according to the German Development Bank KfW's is to verify its interventions contributions to **sustainability**. KfW puts emphasis on conducting independent **ex-post evaluations** of their projects and programs following the **OECD/DAC evaluation criteria**. They evaluate about 50% of their projects/programs in individual standard evaluations and these reports are publicly available for (institutional) learning in the **Quick Evaluation Results (QUER) online application**. Furthermore, KfW also implements more comprehensive thematic evaluations. These may make use of **rigorous IE methods**, including quasi-experimental designs. The rigorous IE design should "follow the function" and it is adapted for the respective project context.

KfW is one of the main implementing organizations of Germany's ODA. It is a German state-owned investment and development bank involved in FC on behalf of German ministries, but mainly the BMZ. In the TVET sector, KfW is mainly financing infrastructure (e.g., school buildings) and equipment (e.g., IT-workshops to enhance digital capabilities of teachers and students or IT platforms), as well as scholarships to enhance the access to quality TVET to vulnerable groups of youth ([KfW 2022d](#)). In 2018, KfW reported to be financing a total of 69 TVET projects with a volume of 1.19 billion USD¹³, most of them in Asia (75%), and Africa (20%), with few projects in LAC ([KfW 2018](#)).

KfW's Evaluation Policy

A publicly available evaluation policy of KfW was not accessible, but the following insights were extracted from the KfW website and their 16th Evaluation Report ([KfW 2022c](#), [KfW 2021](#)). Evaluations of KfW comply with the **OECD/DAC evaluation criteria** as well as with the German Evaluation Society (DeGEval) standards of evaluation. **Independence is a particularly important principle for KfW**. Independent external experts or employees from the operating units of KfW, who were not involved in the implementation of the project/program, conduct the respective project/program

¹³ This is equivalent to 1.2 billion EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

evaluations. KfW conducts multiple **types of ex-post evaluation** of the impact of its projects and programs. These evaluations include:

- **Standard Evaluations:** This type of ex-post evaluation of individual projects/programs is conducted on behalf of BMZ and the evaluation reports are published online. The ex-post evaluations take place about three to five years after the FC intervention is completed. Until 2006, all completed projects were evaluated ex-post and the results were published in a summarized analysis report every two years. Since 2007, KfW draws a stratified random sample (random and representative) from its projects from all sectors ready for evaluation in a particular year. In total, 50% of the more than 100 projects and programs each year are selected for standard ex-post evaluations to reliably estimate the effectiveness of all projects. The evaluation assesses the impact achieved throughout the entire project cycle. Available documents and reports are analyzed, interviews take place on-site, and data and statistics are analyzed to create a final rating of the project based on a numerical scale of 1 to 6 ([KfW 2021](#), p. 42-43, [KfW 2022b](#)).
- **Thematic evaluations:** The evaluation unit prepares more comprehensive analysis on select topics. This enables KfW to identify interdependencies, sector-specific issues or the suitability of certain promotion concepts. These **thematic evaluations can also use rigorous (i.e., empirical/statistical) IE methods** and customize the design from the rigorous IE toolbox as appropriate for the respective project (“form follows function”). A database containing the results of around 3,700 ex-post evaluations of individual FC projects also allows cross-referenced analyses that can be global or focus on particular countries, regions and/or sectors ([KfW 2021](#), [KfW 2022b](#)).

KfW’s Evaluation System

KfW’s sustainability guideline ([KfW 2022b](#)) emphasizes that the **main purpose of evaluations is to verify the contribution of KfW’s interventions to sustainability**, including economic, environmental and social dimensions. KfW established an **impact management system** to measure the economic, environmental and social impacts of its promotional activities based on international standards and the 2030 Agenda. The ToC constitutes the basis for impact management and describes the correlations between KfW’s promotional activities and their specific impacts on the **three dimensions of sustainability**. KfW has defined strategic impact categories and indicators, which are continuously improved. These indicators are integrated in a KfW-wide “**impact balance sheet**” and KfW-wide, “**automatic, efficient and consolidated data management system**” as the centerpiece of KfW-wide impact management. KfW’s impact management system strengthens the sustainability dialogue with customers, stakeholders and the public. It delivers specific findings for planning and further improvement of promotional instruments ([KfW 2016](#)).

To ensure **independence and impartiality** in the assessment of the impact of its projects, KfW’s FC Evaluation Unit is not part of KfW Development Bank’s organizational structure. It reports directly to the executive board. The unit works independently of the regional divisions and is managed by an external person from academia. It has implemented **ex-post evaluations** of the projects and programs for more than 20 years ([KfW 2022b](#)). Since 2019, the evaluation unit defines its mission as, “EVALUATE – MEASURE – LEARN”. KfW denotes the core products of the evaluation unit as: **IEs for ongoing projects, measurement of program success for completed projects and institutional learning** ([KfW 2021](#)).



Institutional learning (“evidence to practice”) is embedded in the use of ex-post evaluation results, because the evaluation unit is systematically channeling its evaluation knowledge back into KfW. One particularly **innovative feature is the new digital knowledge tool, the QUER app**. This interactive app enables project managers to access more than 1000 evaluation results from 2007 onwards and find the evaluations and lessons learned they need to plan new projects ([KfW 2021](#)).

Figure 4: The QUER app allows digital and interactive access >1,000 evaluation findings

Types of Evaluations Conducted by KfW (Especially IE Designs, Approaches and Methods)

The KfW conducts rigorous IEs as part of its **thematic evaluations** (see Subchapter 3.1.2, KfW’s Evaluation Policy above) and defines rigorous IEs as follows: “Rigorous IEs describe a toolbox of experimental and semi-experimental methods that measure the causal effects of a project. The emphasis is on causality. In other words, on identifying those effects that can be attributed exclusively to the project and isolating them from concurrent developments or other connections between projects and target indicators (...)” ([KfW 2021](#)). The evaluation unit at KfW increasingly provides institutional and methodological knowledge to support implementation of rigorous IEs. The evaluation unit adapts the use of rigorous IEs – considering the methodological possibilities and limits, the relevant context, the needs and capacities of its partners. This is consistent with the “**form follows function**” principle and the related question “what do I want to know, and what method is best suited to answer that question?” Depending on needs, households are surveyed, or analysis are conducted with satellite or other secondary data. Ideally, rigorous IEs are implemented in cooperation with other development banks (such as the World Bank or the French Development Agency, *Agence Française de Développement*) as well as local or academic partners. This allows synergies in learning, both between development banks and between partners ([KfW 2021](#)).

3.2 SWITZERLAND'S DC EVALUATION POLICY

In brief, the Swiss actor's—**Swiss Agency for Development and Cooperation (SDC)**, and **Swiss State Secretariat for Economic Affairs (SECO)**—main purposes of evaluations are learning, evidence-based decision-making and improved accountability. SDC/SECO consider the OECD/DAC evaluation criteria (relevance, coherence, effectiveness, efficiency, impact, and sustainability) as useful evaluation criteria and encourage evaluations which are focused on a selection of these for specific uses and users. The Swiss DC actors describe various evaluation methods in their evaluation policies, including impact evaluations to establish the causal effect of a project, program or policy on one or several outcome(s) to generate greater evidence before up-scaling innovations. **They do not determine specific evaluation systems or impact evaluations designs, approaches and methods.** SDC has a lot of experience using non-experimental designs and very limited (only pilot) evaluation experience using quasi-experimental designs. SDC considers randomized designs as rather inappropriate for practice. SDC and SECO aim at achieving full transparency and put **special emphasis on participation, communication and making evaluation reports with management responses publicly available.**

3.2.1 SWITZERLAND'S SDC EVALUATION POLICY

The SDC is one of two agencies within the Swiss government engaged in overseas development. The SDC is Switzerland's international cooperation agency within the Swiss Federal Department of Foreign Affairs. Its responsibilities comprise, among others, the overall coordination of development activities in cooperation with the federal offices in partner countries. SDC focuses on primary education as well as vocational training up to the secondary level II. Vocational Skills Development (VSD) was identified as a key theme, alongside basic education in Switzerland's International Cooperation Strategy 2017-2020. **VSD and private sector engagement remain a thematic priority for the creation of jobs in Switzerland's Strategy 2021-2024 (SDC 2020a).** In 2016, the SDC was implementing **54 core VSD projects** and private sector development projects with significant VSD components in 35 partner countries and regions. As of 2016, 77.6 million USD¹⁴ was invested in VSD projects overall, which was equivalent to 8.5 % of SDC's total expenditures for project interventions in non-EU countries (SDC 2016).

SDC's Evaluation Policy

Similar to the other bilateral DC organizations of DACH countries, SDC measures the **impact of its projects through continuous monitoring of ongoing projects and through evaluations.** This allows the continuous review and improvement of programs or projects and strategies.

Evaluations can address issues that are not readily apparent in the monitoring of ongoing projects, and enable the assessment of complex questions from an external perspective. SDC evaluation principles and quality standards for evaluations are **usefulness, feasibility, correctness, quality and reliability, participation, impartiality and independence, transparency and partnership.** Further, SDC uses (selected) **OECD/DAC evaluations criteria** for the assessments of Swiss DC-projects/programs depending on the specific use and users¹⁵. According to its policy, evaluations at the SDC serve **three interrelated purposes (SDC 2018):**

- **Learning** and gathering knowledge about what works and why, in order to improve the quality and DC results.
- **Evidence-based/results-based decision-making, steering and management** of programs, projects, initiatives, co-operation strategies, networks and policy dialogue as a core of SDC's organizational culture, used to achieve improvements in accordance with the OECD/DAC criteria.
- **Accountability** through reporting and communicating DC results to public stakeholders (e.g., the Federal Department of Foreign Affairs and Parliament) and the wider Swiss public and abroad, including beneficiaries.

SDC conducts multiple **types of evaluation** (including IE) according to their evaluation policy (SDC 2018), including:

- **Cooperation strategy evaluation:** The mid-term or end-term evaluation of a country or regional cooperation strategies are often of a summative or formative nature.
- **Impact evaluation:** According to the SDC: "Impact evaluations establish the causal effect of a project, program or policy on one or several outcome(s). These are usually conducted for greater evidence before up-scaling innovations."
- **Institutional evaluation:** These are sector-wide, organizational or partner evaluations based on overarching institutional objectives which frequently uses developmental approaches.
- **Joint evaluation:** An evaluation in which different donors and/or partners participate. Potential benefits include mutual capacity development, joint learning, harmonization, reduced transaction costs, and broader scope.
- **Meta evaluation:** Meta evaluations are designed to aggregate findings from a series of evaluations. They can also be used to denote the rating of an evaluation to judge its quality and/or assess the performance of the evaluators.

¹⁴ This is equivalent to 77.4 million CHF in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

¹⁵ and they apply the evaluation criteria correctness, coverage and coordination to humanitarian aid evaluations.

- **Program/ project evaluation:** Mid-term or end-term evaluation of a program/project or a set of program/projects within the SDC's operational departments. External evaluations are the norm; sometimes also conducted with the participation of peers.
- **Self-evaluation:** Self-evaluation is an evaluation by those who are entrusted with the design and delivery of a development intervention. Self- or internal evaluations are mainly used for learning purposes.
- **Thematic evaluation:** Evaluation of a selection of development interventions, all of which address a specific development priority that cuts across countries, regions and sectors.

SDC's Evaluation System

Evaluations may be conducted and commissioned by the **evaluation and controlling unit** (according to an annual thematic or systematic evaluation schedule agreed upon with the SDC's senior management), or the **operational units at the head office or in the Swiss cooperation offices**. The operational units conduct evaluations of strategic interest to them. The **quality assurance and internal digitalization unit** promotes the use of evaluations as an integral part of a comprehensive monitoring and evaluation (M&E) concept contributing to strengthening the results-based focus and to promoting a learning-oriented operational management ([SDC 2018](#), [SDC 2022](#)).

SDC has a **decentralized structure**. The Swiss cooperation offices conduct, or rather commission, most evaluations because **program/ project mid-term reviews** are the most frequent type of evaluations (*see Subchapter 3.2.1 Types of Evaluations Conducted by SDC for further information below*). These aim at learning, at supporting the further development of the projects/programs and identify scaling-up potentials. The evaluation unit provides some tools and guidance documents for the country offices. Despite this guidance, there is a need for training in evaluation designs, approaches, methods and systems for country office staff at the project-level¹⁶.

The SDC usually plans a **budget for evaluations** during the program approval process, which can be accessed during project implementation. A **mid-term review** usually costs between 40.000-80.000 USD¹⁷ and is likely to be approved during project implementation. About 10 % of SDC's projects are evaluated at country level. **Thematic or systematic evaluations** are commissioned centrally by the evaluation unit based in Bern and may cost about 200.000 USD¹⁸.

Types of Evaluations Conducted by SDC (Especially IE Designs, Approaches and Methods)

Impact evaluation is defined as one of many types of evaluations in the evaluation policy of the SDC (*see Subchapter 3.2.1 SDC's Evaluation Policy above*). The evaluation policy ([SDC 2018](#)) defines IE as follows: "Impact evaluations establish the causal effect of a project, program or policy on one or several outcome(s). These are usually conducted for greater evidence before up-scaling innovations." This definition is based on the standard definition of the OECD. It does **not specify the evaluation approaches and methods to be used to measure results and impact**. SDC is currently in the process of revising its evaluation policy, because it is not very specific to SDC's internal practices and does not provide concrete guidance for practitioners on: the type of evaluation approaches and methods; timing of evaluation; and level of evaluations. In this context, the SDC commissioned the Center for Evaluation (CEval) to conduct a meta-evaluation of the existing evaluation reports to learn and improve their quality. The CEval evaluation found that one of the **main weaknesses of SDC evaluation reports is the frequent absence of a methodological chapter**, which makes the evaluation findings less informative and makes it difficult to prove their validity. This shortcoming has a structural origin in the SDC's evaluation system, as the SDC usually commissions evaluations to external evaluators, who make methodological suggestions to the SDC in their tenders, because the ToR rarely contain methodological specifications. However, the SDC is currently discussing if this should change in the future¹⁶.

The **frequently conducted program/project mid-term reviews or evaluations** use mixed methods that usually comprise desk studies, quantitative and qualitative data analysis derived from FGD and interviews as well as triangulated data. These mid-term project evaluations are perceived as very useful and helpful for project development, even though rigorous IE (experimental or quasi-experimental designs) are not used. These mid-term reviews do not assess the impact level, but rather the outcome level (effectiveness).

According to the SDC evaluation unit, rigorous IEs, namely experimental and quasi-experimental IE designs, are not conducted at project level. The head of the evaluation unit stated multiple reasons for this: Firstly, experimental designs (RCTs) have to be considered from the beginning, because these cannot be applied to ex-post evaluation settings, which are the most frequent evaluation settings in DC contexts. Secondly, RCTs should be accompanying evaluations and have to be implemented over several (3-4) years. Thirdly, adequate data must be availability, which is often not the case. Fourthly, professional scientific support is required, because the technical expertise for experimental or quasi-experimental designs is usually not available in the projects. Fifthly, the DC project context and needs are important and even non-rigorous evaluations meet these needs. Most non-rigorous mid-term evaluations are perceived as very helpful and are highly appreciated by the projects.

¹⁶ According to the evaluation unit of SDC.

¹⁷ According to the evaluation unit of SDC. This is equivalent to approximately 40.000-80.000 CHF in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

¹⁸ According to the evaluation unit of SDC. This is equivalent to approximately 200.000 CHF in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

The SDC evaluation unit has **piloted only one quantitative rigorous IE** so far titled *Impact Evaluation of the Support Program for Education and Training of Children Excluded from the Education System in Benin*. It was a rigorous IE trial conducted by the Center for Evaluation and Development in Mannheim (C4ED), Germany in 2020. This impact evaluation used both a **quasi-experimental difference-in-difference (DiD) design and a matching design**. The SDC published the impact evaluation report, a short presentation of the impact evaluation findings and a one-page fact sheet with many infographics ([SDC 2020b](#)). The SDC evaluation department derived **two general learnings about rigorous IEs** from this experience: Firstly, this IE cost the SDC a few hundred thousand USD and it was therefore **very expensive**. Secondly, it was centrally coordinated and commissioned. In the future, IEs should not be centrally managed but rather commissioned by the country offices, as they are in a better position to provide local insights. Also, they know best about their results and impact knowledge-generation needs and can better assess if it is worthwhile conducting such studies. Therefore, a **decentralized needs assessment will take place** in the future. It is rather unlikely that future rigorous IEs will be coordinated by the evaluation unit at the headquarter again, but country offices may contract these at decentralized level.

Use of existing M&E data: According to the SDC evaluation unit, SDC projects have made progress in becoming quite disciplined at capturing baseline data in the first six months. Furthermore, implementation reports are used as a basis for mid-term evaluations and every project has a log frame, which is updated at mid-term to assess the progress and verify the status during field visits. However, **it is questionable if these data (including baseline data) fulfills the quality standards for a rigorous IE**.

3.2.2 SWITZERLAND'S SECO EVALUATION POLICY

SECO is the other agency within the Swiss government which is engaged in overseas development. SECO is the federal government's center for all core issues relating to economic and labor market policy. Its portfolio covers market-oriented skills with the aim to create income opportunities. SECO focuses on TVET and promotes the incorporation of market-oriented expertise in higher vocational training (post-secondary and tertiary levels) and therefore complements the work of the SDC. In addition, the private sector is at the heart of SECO's activities to create jobs.

SECO's Evaluation Policy

SECO's evaluation policy ([SECO 2021](#)) describes **purposes of evaluations, types (categories) of evaluations** and reviews as well as **different actors and their roles** in the evaluation process and how transparency is maintained.

The evaluation purposes are similar to those of SDC: **learning, accountability and steering**. SECO evaluations are also based on the OECD/DAC evaluation criteria. Evaluations are expected to provide credible and useful information, which enables the incorporation of lessons learned into the decision-making processes of recipients, implementing partners, and SECO itself. The **credibility and effectiveness of SECO evaluations** involve the following principles: "(1.) Clear governance (roles and responsibilities) incl. an independent oversight of the evaluation activities. (2.) A sound system of independent evaluations. (3.) Integration of evaluation results in knowledge management processes. (4.) Transparency with regards to the results of evaluations."

SECO defines three types (categories) of evaluations and reviews. They carry out **internal reviews, external evaluations and independent evaluations** at the project and program level.

SECO's Evaluation System

Within SECO, many different actors may be involved in an evaluation, including sections with operational activities, the evaluation unit, the senior management of the division, an external evaluation committee and external evaluators. SECO project and program managers may evaluate their projects internally, which is called an **internal review**. These internal reviews are self-assessments, which serve the purpose of institutional learning and improving planning and implementation of projects. In addition, **external and independent specialists** evaluate SECO projects, programs or thematic areas at the portfolio level. This guarantees impartiality and accounts for the money used in international cooperation ([SECO 2022](#)). The evaluation policy requires SECO to report to an external evaluation committee and the senior management ([SECO 2021](#)).

Types of Evaluations Conducted by SECO (Especially IE Designs, Approaches and Methods)

The evaluation guidelines of SECO ([2017](#)) specify a rather **unusual definition of impact evaluation** as an in-depth assessment, which looks at positive and negative, intended and unintended, short-term and long-term effects of a development intervention. SECO specifies in their guidance as "an evaluation focused on the expected results achieved at impact level is not IE." The SECO **evaluation policy (2021)** does not provide any further information about **impact evaluations** besides the fact that **the SECO evaluation unit may commission IEs as a type of independent evaluation**. SECO expects that only a few IE will be conducted due to the high costs of IEs. They define, that impact evaluations are adequate when there are questions about the adequacy of the instrument used or when there is strong pressure to assess all positive and negative effects of an intervention or to enable the assessment of specific types of interventions. They highlight that evaluation officers should be involved in the design and decision on conducting IE, because of potential methodological issues ([SECO 2017](#)).

Use of existing M&E data: The SECO evaluation policy states that M&E are mutually interdependent. “With careful monitoring, important data on project or program progress can be collected, and the availability of quality monitoring data is necessary for good evaluation. In turn, evaluations provide lessons for improving the design and implementation of the monitoring systems” ([SECO 2021](#)).

3.3 AUSTRIA’S DC EVALUATION POLICY

In brief, Austria’s DC actors—Austrian Development Agency (ADA) and the Development Bank of Austria (OeEB)—have a joint evaluation policy and they view learning, steering, accountability and communication as the key functions of evaluations. According to ADA/OeEB, impact evaluations have the intention to causally attribute impacts to specific development measures. Therefore, they examine and assess the causal links and effects of development interventions at different levels. According to their definition, **IE does not pre-determine the use of a specific evaluation design**, which should remain amenable to different notions of causality. They **systematically apply non-experimental evaluation designs and theory-based, qualitative approaches (especially contribution analysis) to measure the results and impacts of projects/ programs**. The ADA evaluation guidance states very clearly, “**there is no single right or best evaluation design or approach**, which needs to be tailored to the specific evaluation purpose, objectives and questions” ([ADA/OeEB 2020](#)).

The Austrian Development Agency (ADA) is the operational unit in charge of implementing all bilateral projects and programs of the Austrian DC. The Development Bank of Austria (OeEB) finances private investment projects in developing countries and emerging markets. ADA engages in TVET and the labor market under its theme “education.” TVET facilitates access to adequately paid work and productive employment. ADA strengthens modern educational services and effective national TVET systems ([ADA 2022a](#)).

Austria’s Evaluation Policy for DC

The Austrian’s DC actors have a joint evaluation policy ([ADA/OeEB 2019](#)), which sets the quality standards and benchmarks for the Austrian DC, calling for **robust findings on impacts achieved**. It also defines the institutional requirements that ensure useful and credible evaluations as well as the transparent communication of findings to partners and the public at large. Evaluations make a major contribution to generate robust findings on achieved impacts and are essential for fostering a **learning, evidence-based and strategically oriented** Austrian DC. ADA’s main standards and principles for good evaluations are **independence, impartiality, credibility, transparency, utility, feasibility, fairness, accuracy, participation and partnership**. Their evaluation criteria adhere to the OECD/DAC evaluation criteria.

Austrian DC evaluations perform three **interconnected functions**:

- **Learning:** Evaluations support institutional learning and contribute to the ongoing improvement and optimization of the quality and effectiveness of Austrian DC.
- **Steering:** Evaluations provide reliable findings that contribute to the evidence-based planning of development-policy objectives and underpin strategic and operational decision-making processes.
- **Accountability and communication:** Evaluation findings give account of the use of public funds and report and communicate impacts achieved to partners, donors and the public.

Different **types of evaluations** are conducted according to suitability of the following four evaluation functions:

- **Object of the evaluation (evaluand):** SDC distinguishes: institutional evaluations; cooperation strategy evaluations (e.g., ADA/OeEB’s engagement in TVET in a partner country or region); project/program evaluation (e.g., specific TVET projects/programs); and thematic, sectoral or instrumental evaluations (e.g., TVET evaluations).
- **Methodology:**
 - ♦ **Evaluability assessments** consider to which extend the object of an evaluation (a measure, project, program, instrument, strategy or organization) can be evaluated in a reliable and plausible way, which requires an ex-ante appraisal.
 - ♦ **Impact evaluation:** According to the Austrian DC actors “Evaluation that examines and assesses the causal links and effects of development interventions at different levels. The term is based on the intention of causally attributing impacts to specific development measures. It does not pre-determine the use of a specific evaluation design, but is amenable to different notions of causality.”
 - ♦ **Systematic review (SR)/Meta-evaluation:** Synthesizes the findings of various evaluations or assess the quality of evaluations or the performance of evaluators.
- **Timing:** Real-time evaluation deliver direct feedback on an ongoing intervention in order to identify and address policy, organizational and operational constraints for steering and learning and **ex-post evaluation, which take place after the completion of a development measure**.
- **Mode of implementation:** Joint evaluation, which involves multiple donors and/or partners.

Austria's Evaluation System for DC

Austria's evaluation system is described in the evaluation policy, specifying its key aspects related to **planning, implementation and application/utilization** of evaluations ([ADA/OeEB 2019](#)).

Austrian DC **usually outsources evaluations** (which they call “reviews”) due to the **independence of external evaluators** (in line with the first quality standard) and limited internal evaluation capacities within ADA and OeEB. Usually, there is sufficient budget available for evaluations/reviews. If no budget is available for external evaluations or reviews, they conduct an internal reflection instead. The evaluation department is responsible for the strategic evaluations, which usually cost between 80.000 and 120.000 USD¹⁹ per evaluation. There have been a few more costly strategic evaluations of about 299.000 USD²⁰. Program/project evaluations typically cost between 25.000 and 90.000 USD²¹. The programs/projects are responsible for the implementation and contract external consultants. These strategic and project/program evaluations **apply non-experimental evaluation methods** and focus on effectiveness rather than on impact. For the commissioning of (quantitative) experimental or quasi-experimental rigorous IE, ADA would expect higher budgetary requirements of at least 249.000 USD^{22,23}.

In the past, each Austrian DC project had to carry out at least one evaluation, but this led to very low budgets per evaluation. Therefore, ADA has limited the number of evaluations to at least 30% to 50% of all approximately 600 (currently ongoing) projects and assigns sufficient budget to these, which must be evaluated at least once per project cycle, in line with their *Guidelines for Programme and Project Evaluations* ([ADA 2020](#)). This guideline contains very practical recommendations for the three phases and fifteen steps of the evaluation process, which can help to take initial steps before an adequate **evaluation design** (incl. experimental, quasi-experimental and non-experimental designs) and **evaluation approach** (quantitative and qualitative) are defined (see *Figure 5, Step 4*). Figure 5 shows the three phases of the ADA evaluation process with their corresponding steps:

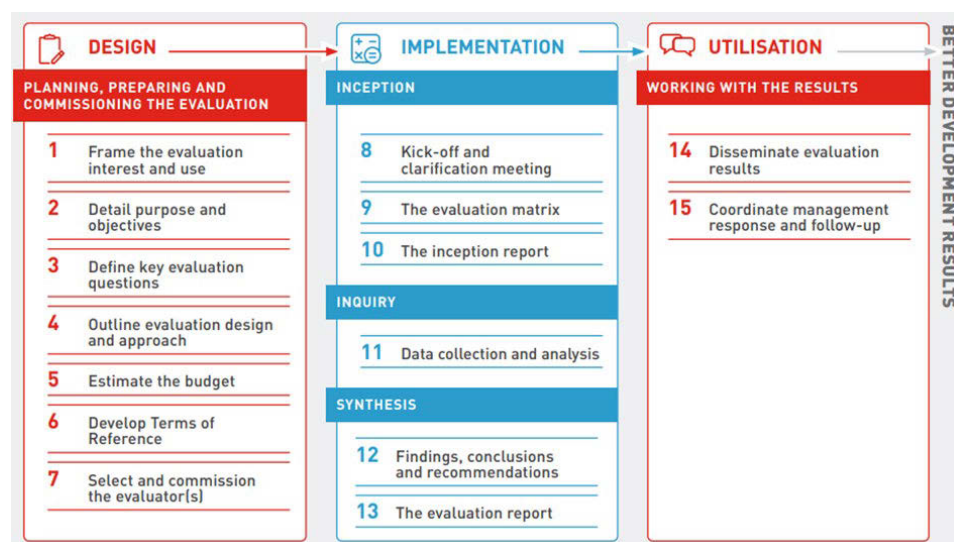


Figure 5: Three phases and 15 steps of the evaluation process ([ADA 2020](#))

Types of Evaluations Conducted (Especially IE Designs, Approaches and Methods)

Impact evaluations are one of three types of evaluation methodologies mentioned in the Austrian evaluation policy. **ADA does not use experimental rigorous IE designs.** They decided against conducting experimental evaluations (RCT), because randomization is difficult or not feasible in the DC context²³. However, ADA has **conducted one quasi-experimental IE** with propensity score matching that of a recent rural cooperative project in Armenia and Georgia, which was not related to the TVET sector ([ADA 2022b](#)). **In the TVET sector, ADA has no experience with experimental or quasi-experimental designs.**

Instead, ADA tends to use **theory-based methods for program and project evaluations as well as strategic evaluations.** Contribution analysis is the standard and systematically applied tool within ADA and is most frequently used to test theories (incl. results and impacts of interventions). ADA has also conducted some systematic reviews, evidence synthesis reports and case studies. **Austrian DC tends to include non-experimental designs with qualitative approaches among the rigorous IE, even though the strict academic definitions would only include quantitative approaches with experimental or quasi-experimental designs.**

¹⁹ This is equivalent to about 80.000 and 120.000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

²⁰ This is equivalent to about 300.000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

²¹ This is equivalent to about 25.000 EUR and 90.000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

²² This is equivalent to about 250.000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

²³ According to the evaluation unit of ADA and the source [ADA 2020](#).

Existing M&E data must be improved in many cases, because monitoring data is often not sufficiently available or of insufficient quality for evaluations. Therefore, they have started to **assess the evaluability of projects** in depth (see *Evaluability Assessments in Austrian DC* textbox below) ([ADA 2022c](#)). The evaluation unit of ADA is not aware of good practice examples of **administrative data use** for (impact) evaluations within Austrian DC²³.

Evaluability Assessments in Austrian DC:

The **guidance document on evaluability assessments** summarizes four facets of evaluability and presents associated checklists. “Evaluability” stands for “the extent to which an activity or project can be evaluated in a reliable and credible fashion” (OECD/DAC definition). The Austrian DC aims at making interventions more evaluable and strengthen the quality of subsequent evaluations by making these more feasible, meaningful and cost effective:

- Evaluability “in principle” concerns the **quality of the intervention design**, incl. ToC.
- Evaluability “in practice” stands for the **availability and accessibility of data** (M&E).
- The utility of an evaluation reflects the **needs of different stakeholders**
- The practicality of an evaluation in the boundaries of the **institutional and physical context**.

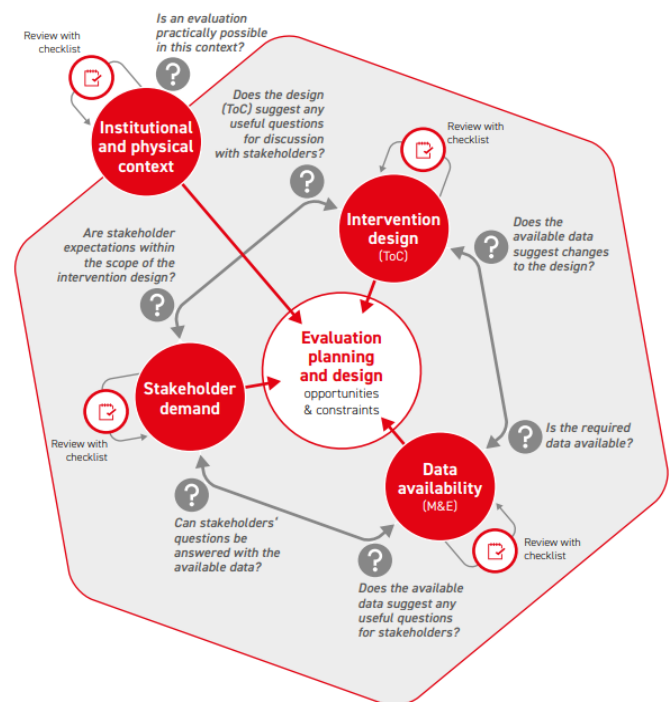


Figure 6: The Austrian DC guidance on evaluability assessment

CHAPTER 4 CASE STUDIES - OVERVIEW OF EXISTING IE AND STUDIES OF SKILLS DEVELOPMENT INTERVENTIONS IN GERMAN DC

The following nine case studies present an overview of existing impact evaluations and studies of German DC TVET interventions. In line with Chapter 2, these are clustered according to: experimental designs (see two Case Studies in Subchapter 4.1); quasi-experimental designs (see four Case Studies in Subchapter 4.2); and non-experimental designs (see three Case Studies in subchapter 4.3).

4.1 EXPERIMENTAL DESIGNS

4.1.1 CASE STUDY 1: UGANDA – RANDOMIZED EVALUATION OF STUDENT TRAINING FOR ENTREPRENEURIAL PROMOTION

Project Description

Title	Action and Action-Regulation in Entrepreneurship: Evaluating a Student Training for Promoting Entrepreneurship
Commissioned by	Leuphana University Lüneburg, Germany
Implementing organization	Center of Evidence Based Entrepreneurship Development at Leuphana University
Implementing partners	Ugandan universities in Kampala and Mukono
Research institute	Leuphana University (Gielnik, Michael M.; Frese, Michael; Kahara-Kawuki, Audrey et al.)
Project area	Uganda (similar trainings implemented in various countries)
Target groups	Undergraduate students in their last year (all disciplines except business administration)
Project term	Since 2006 until present
Project cost	Up to 100.000 USD annually (university is able to rely on academic staff and students) Main third-party donors: German Academic Exchange Service (DAAD), World Bank, foundations (Baden Aniline and Soda Factory, BASF), German Commission for UNESCO and BMZ
Evaluation term	Since 2009
Evaluation cost	Included in project cost
Evaluation design (evaluation approach)	Experimental design (quantitative, RCT approach)
Publication date	2015

Project Context and Results

In 2006, the Leuphana University Lüneburg in Germany has developed—in cooperation with several African universities—two training programs in the field of entrepreneurship. The training covered by this case study, **Student Training for Entrepreneurial Promotion (STEP)**, is targeted at students and youth. The trainings have been implemented in cooperation with different partners (e.g., DAAD, World Bank, German Commission for UNESCO, etc.) in several countries. Various countries in Africa, Asia, and Latin America (including Mexico) have participated in STEP since the first training in 2009. All trainings are evaluated through a **randomized controlled trial (RCT)**.

The evaluation is based on a **randomized controlled field experiment**, which was conducted at two Ugandan universities in Kampala in 2009. This was the first cohort that attended a STEP training. The researchers developed, together with the lecturers from four universities, an **action-based entrepreneurship training for undergraduate students** in their last year. The training comprised of twelve different modules (such as identifying business opportunities, marketing, leadership, and financial management), and was taught on a weekly basis over a period of twelve weeks. The training was voluntarily and independent of the regular university program. Although the participants received a certificate, the training was not graded. The 12-month evaluation study showed that the **training had a significant impact on business creation** meaning students in the training group were significantly more likely to start a new business than students in the control group.

Evaluation Objectives and Indicators

The goal of the evaluation was to investigate the **long-term effects of the action-based training on business creation** (i.e., the probability of new start-ups). Specifically, the authors of the evaluation wanted to develop and investigate a theoretical model, that explains why and how the action-based entrepreneurship program has a positive effect on entrepreneurial action and business creation. Therefore, the authors of the evaluation hypothesized that the **training**

positively influenced four action-regulatory factors: students' entrepreneurial goal intentions, action planning, entrepreneurial self-efficacy, and action knowledge. These factors are short-term outcomes of the training that transmit the effect of the training on the long-term outcome of entrepreneurial action.

Evaluation Approach and Methods

To evaluate the impact of the training on business creation through the four regulatory factors, the team conducted a **RCT** comparing a treatment group with a control group. The treatment was **the action-based entrepreneurship training**, and the researchers **randomly assigned** those students that had applied for the training to the treatment and control group.

Randomization and control group: Undergraduate students from two universities were invited to voluntarily participate in entrepreneurship training. The deans of the universities handed out application forms and in total 651 students applied for the training (424 from University A and 227 from University B). Since the training capacity was limited, 203 **applicants were randomly selected for participation** while another 203 applicants from the list were randomly allocated to the control group, which did not receive the training. The control group was a "waiting" group, which means that they received the training after the evaluation. To take part in the training and to create a certain degree of commitment to participate throughout the training, the students had to pay a deposit of approximately 10 USD, which was refunded at the end of the training if all modules were attended.

Data collection: To collect the data, the researchers **employed a pretest-posttest design** and conducted three measurements (T1, T2, and T3). The first measurement (T1) took place in **the month before the training**. Since some of the control group students did not participate in the data collection and some students of the treatment group failed to complete the training, so the treatment group was reduced to 194 and the control group to 190 students. The second measurement (T2) took place in the **month directly after the training** had ended. The authors of the evaluation were able to trace 184 former training participants and 153 control group students. The third measurement (T3) took place **12 months after T1** and included 162 students from the training group and 142 from the control group. All data was collected with **personal interviews and questionnaires**. The interviewers received a comprehensive interviewer training including: sessions on interview techniques to probe participants' answers; the use of prompts to clarify abstract statements; note taking; and typical interviewer errors (e.g., non-verbal signs of agreement). To test whether non-response biased the data, the researchers analyzed whether the non-respondents of the training group differed significantly from the non-respondents of the control group. The reasons for non-response were either lack of time to conduct the interview or lack of motivation to further participate in the study but did not differ between the two groups.

Measurement: As mentioned above, the authors aimed at investigating how the training would affect business creation, based on the four regulatory factors. To measure **action knowledge**, a situational interview was conducted and the students were asked to identify actions based on a scenario. Based on twelve questionnaire items, students were asked to indicate how confident they are of performing certain entrepreneurial tasks to measure **entrepreneurial self-efficacy**. To identify **entrepreneurial goal intentions**, a 5-point Likert scale was used to ask students if they intended to pursue specific start-up activities within the next six months. Considering **action planning**, the students were asked whether they were in the process of starting a business or planning to start one within the next 12 months and which concrete plans they had. The answers were rated by two independent researchers, based on how detailed their business plans were.

Focal and Cross-cutting Topics

- **Sustainability of impacts:** The implementation of STEP over the course of three years within the partner institution and with the accompanying data collection, allows the researchers to study the long-term effects of the training over more than 32 months. Moreover, many partners continue with the program after the third year, which means that the trainings are fully locally implemented (ownership), which enables sustainability.
- **Gender-sensitive impact assessment:** The participants' gender was recorded, however, the impact effects did not differ by gender.
- **Measurement of entrepreneurial skills:** The authors developed certain methods to measure different entrepreneurial skills such as, action knowledge, entrepreneurial goal intentions, action planning and entrepreneurial self-efficacy through personal interviews and questionnaires as described in the section above.
- **Impact of COVID-19 pandemic:** The training implementation was interrupted for more than one year due to the pandemic. They tried to set up online training courses, however the implementation was not successful. Online training required different methods compared to physical training and they were not able to adapt the methodology. However, STEP trainers have continued to receive online training.

Key Evaluation Findings

The 12-month evaluation study showed that the **training had a significant impact on business creation**: students in the training group were significantly more likely to start a new business than students in the control group. In line with the described hypotheses, the training had significant effects on entrepreneurial goal intentions, action planning, action knowledge, and entrepreneurial self-efficacy. Action knowledge and the interaction between entrepreneurial goal intentions and action planning were significant **predictors of entrepreneurial action**. The action-regulatory factors fully mediated the effect of the training on entrepreneurial action. Furthermore, the training had positive effects on **business opportunity identification**. The study showed that **action-regulatory mechanisms are of central importance** in entrepreneurship, and they help to explain how action-based entrepreneurship trainings have a positive impact on entrepreneurship. Promoting entrepreneurship is possible if during training, trainers take into consideration action-regulatory mechanisms important for entrepreneurial action.

The study has also practical implications for **future studies**, evaluating the effectiveness of entrepreneurship trainings. Entrepreneurial goal intentions alone may have a positive effect on action; however, this effect is not so strong. The researchers found that entrepreneurial goal intentions are necessary but not sufficient predictors of action; entrepreneurial goal intentions instigate actions only when **entrepreneurs specify what they will do and how they will do it**. This finding implies that intervention programs focusing only on increasing the strength of entrepreneurial goal intentions without increasing the level of action planning do not have a positive impact on entrepreneurship. In addition, entrepreneurial goal intentions and action planning must be part of the evaluation to assess the effectiveness of training interventions.

Conclusion on Methods Used

This **academic approach** is different from DC approaches. The authors developed and rolled out two different trainings in various countries in order to study the impact **applying RCTs to constantly improve the methodology**. The primary interest of the project under study lies in research and in producing rigorous evidence. Therefore, RCTs are used to study the impact of every training. The evaluation through an RCT allows the researchers to **publish the results** in high-ranked journals. Interestingly, the authors developed an approach to study not only the effect of the training on business creation, but also the regulatory factors through which effects are transmitted. This enabled the team to **understand the mechanisms and determine successful factors of entrepreneurial trainings**. Moreover, the project setting allows the authors to constantly study long-term training effects.

Initially, the researchers had **small case numbers** (quality rather than quantity), but since have been able to scale up the trainings with additional funds. Even though the funds were limited in the beginning, the authors were able to **implement a pilot study in Uganda and then scale up the training** guided by empirical evidence. Usually, project implementation and up-scaling follows three steps. In the first year, STEP is introduced, implemented, and evaluated to demonstrate its beneficial effects on students' entrepreneurial behavior. In the second year, the partner institution assumes responsibility for organizing and implementing STEP to include STEP in the regular curriculum of one or more programs of study. In the third year, the partner institution independently organizes the training and decides about the institutionalization of STEP in their academic program.

The University of Lueneburg **continues to set up its projects independently** in order to answer relevant research questions. Evaluating GIZ projects would restrict their scientific scope and their publication success.

If these training measures are introduced in other country or regional contexts, these must be **adapted to the partner situation**. In **Mexico** for example, the partner institutions were better organized and had more specific ideas on training content. In consequence, the training must be **adapted to the local needs and wishes**. The Mexican partners preferred to have a different focus and a different strategy (e.g., focus on accelerators), which challenges the training's long-term implementation.

The **partner selection is crucial** for a successful and sustainable implementation. So far, the authors used a bottom-up approach: they initiated small projects with individual universities. To date, the universities (e.g., in Uganda) have introduced entrepreneurship training to the national curriculum and not the other way around. They are now **considering a top-down approach** via national ministries, which could positively affect the project success.

Reference

Action and Action-Regulation in Entrepreneurship: Evaluating a Student Training for Promoting Entrepreneurship

MICHAEL M. GIENNIK
Leibniz University of Lüneburg, Germany
MICHAEL FRESE
National University of Management and Leadership, Lüneburg, Germany
AUDREY KAWUKI
Makerere University Business School and Makerere Ltd, Uganda
ISAAC KAHARA-KAWUKI
Uganda Christian University, Kampala, Uganda
SAMUEL KIZIYAMA
Makerere University Business School, Kampala, Uganda
JACOB OYITE
Kampala University, Kampala, Uganda
SAMUEL KIZIYAMA
Makerere University Business School, Kampala, Uganda
THOMAS WALTER
GIZ Berlin, Germany
KIM KASCH
University of Lüneburg, Germany

Gielnik, Michael M.; Frese, Michael; Kahara-Kawuki, Audrey et al. (2015): Action and Action-Regulation in Entrepreneurship: Evaluating a Student Training for Promoting Entrepreneurship. Academy of Management Learning and Education. <https://doi.org/10.5465/amle.2012.0107> [accessed online on 07.09.2022]

4.1.2 CASE STUDY 2: INDIA – RANDOMIZED EVALUATION OF SUBSIDIZED VOCATIONAL TRAINING FOR WOMEN

Project Description

Title	Learning and Earning: Evidence from a Randomized Evaluation in India
Commissioned by	N/A due to independent research (Forschungsinstitut zur Zukunft der Arbeit / Institute of Labor Economics (IZA) website published the article due to one of the authors affiliation with the IZA)
Implementing organization	Social Awakening Through Youth Action and Pratham Education Foundation
Implementing partners	Social Awakening Through Youth Action and Pratham Education Foundation
Evaluation institute	Pushkar Maitra (Monash University, Australia), Subha Mani (Fordham University, USA)
Project area	New Delhi, North and South Shahdara
Target groups	Women (18-39 years, with at least five completed grades of schooling)
Project term	August 2010 – January 2011 (duration of the training)
Project cost	About 30,000 USD
Evaluation term	Baseline survey: July-August 2010, midline survey: July-August 2011, endline survey: July-August 2012
Evaluation cost	About 15,000 USD
Evaluation design (evaluation approach)	Experimental design (quantitative, RCT approach)
Publication date	2017

Project Context and Results

The evaluation focused on a **vocational training program targeted at women in low-income households** in India. The program included a 6-months of classroom training in sewing and tailoring. It was designed and implemented by two NGOs; the Pratham Education Foundation and the Social Awakening Through Youth Action. The program was offered in two low-income areas of New Delhi, North and South Shahdara. The evaluators were involved before the implementation started, which allowed them to implement a randomized field experiment. The research was led by researchers at Monash University and Fordham University. The training program had a positive effect on participating women's income and employment probability.

Evaluation Objectives and Indicators

The study evaluated the **treatment effects** from participating in the subsidized **vocational training program** targeted at women residing in low-income households in India. Specifically, the evaluation team studied the impact of the program on:

1. **Labor market outcomes** such as casual/full-time employment (binary), self-employment (binary), hours worked, and monthly earnings.
2. **Entrepreneurship and empowerment** measured by ownership of a sewing machine, rotating savings and credit association (ROSCA) membership, and happiness at home.

In addition to evaluating the benefits from receiving the training program, the researchers also studied the **factors that hindered women's ability to participate in/or complete the training**.

Evaluation Approach and Methods

A quantitative evaluation approach and experimental evaluation design was used to measure the causal impact of the training. The **researchers conducted a RCT** wherein approx. 10,000+ women (aged 18 to 39 years, with at least five completed grades of schooling) from the target areas in India were invited to apply to the vocational training program. They were informed about the program through an extensive advertising campaign. This campaign was not targeted at any specific sub-group in the population and was distributed to every household in the target area. In order to avoid attracting only women with specific characteristics (e.g., women with a higher education), the description of the program was kept general enough to encourage all eligible women to apply.

Randomization and control group: Randomization allows for causal interpretation, hence, of the 658 women who applied to participate in the program 442 were randomly assigned (using public lottery) to the treatment group and the remaining 216 were assigned to the control group. This process ensured that the **control and treatment group are comparable in pre-existing socio-economic characteristics** such as age and education. However, women who decided to apply for the program might differ from those who did not.

Data collection: Before the program started and the applicants were informed about whether they had been accepted for the training (the applicants were aware of the random allocation), a **baseline survey** was conducted to collect data on the treatment and control group **before the intervention** (594 women interviewed). The treatment and control groups were then surveyed again **6 months after the training** (midterm survey, 504 women interviewed) and **18 months after the training** had ended (endline survey, 491 women interviewed). The surveys included socio-economic characteristics of the women (e.g., age, education, marital status) as well as the labor market outcomes and entrepreneurship and empowerment outcomes mentioned above. Local (mostly female) enumerators conducted the surveys and filled in the questionnaires with the women.

Collaboration: The researchers and the two implementing NGOs collaborated before the intervention had been introduced. This **created several benefits**, which ultimately allowed the evaluation team to set up a randomized field experiment. Firstly, the team was able to create “clean” treatment and control groups ensuring randomized participant selection. Secondly, the team was able to collect the specific data that met their needs before the intervention started. Therefore, they were able to compare the outcomes between the two groups before and after the training program.

RCT costs: The conducted RCT was relatively inexpensive (~30,000 USD).

Limitations

Attrition: Some participants of the training program **dropped out** before the training had been completed. Moreover, some women dropped out during the course of the three data collection rounds. This **attrition needs to be examined** and therefore, the researchers **carefully investigated which women dropped out** and whether they had specific characteristics that are different from the rest of the group, especially if the attrition rates were distinct between women assigned to the treatment and control groups. Moreover, the whole team tried to **anticipate participation barriers** before the implementation and tried to address those. For example, they conducted a survey before the curricula development, to find out which skills the targeted women would like to attain (they chose sewing and tailoring) which would ensure somewhat high program take-up. Moreover, they allocated women to the training centers that are nearest to their home to reduce commuting times and security issues.

External validity: The external validity and applicability of the study findings to other contexts is limited (a common problem of RCTs). The training program was implemented in two specific low-income areas of India. Those locations exhibit characteristics (e.g., average income, infrastructure, religion, ethnicities) that are not representative for India as a whole, because populations are heterogeneous. Thus, the study results are not necessarily easily transferable to other contexts.

Focal and Cross-cutting Topics

- **Gender-sensitive impact assessment:** The evaluated training program specifically addressed women between the ages of 18 and 39 and thus only includes the training impact on females. In the evaluation, several gender-sensitive issues were considered. The evaluators ensured that most enumerators, who conducted the survey together with the women were female. Moreover, the researchers studied barriers to training participation and identified certain constraints that specifically apply to women; lack of adequate childcare and security concerns when commuting to the training centers.
- **Usage of available M&E/administrative data:** The evaluators was not able to use existing administrative data. For example, census data from the targeted areas were either not available, of low quality or unsuitable for the research purpose. The team thus had to conduct a census of all women in the target areas, measure the distance from the training center for all participants, conduct a pre-intervention survey to identify preferences for training content and conduct three survey rounds (baseline, midterm, end line).
- **Sustainability of impacts:** The researchers investigated sustainability in terms of sustained impact after the program had ended. Therefore, they conducted a survey 6 months after the training and again 18 months after the training. This allowed them to analyze whether positive training effects were sustained over a longer period.
- **Measurement of employability:** The evaluators included labor market-related outcome in order to study the training program’s impact on the employment situation of the women. To measure the impact, the women were asked about their employment status, whether they were self-employed, their income as well as their average working hours.

Key Evaluation Findings

The evaluators compared post-intervention outcomes between treatment and control while controlling for pre-existing differences between the two groups to quantify the 6- and 18-month treatment effects of the training program. Looking at the **labor market outcomes** introduced above, the 6-month effects of the program indicated that women who were offered the training program were 6 percentage points more likely to be employed, 4 percentage points more likely to be self-employed, work 2.5 additional hours per week, and earn 150% more per month than women in the control group. Using a second round of follow-up data collected 18 months after the intervention; they found

that the 6-month treatment effects are all sustained over this period. Concerning the **outcomes in entrepreneurship and empowerment**, the 18 months effects indicate that women from the treatment group are 13 percentage points more likely to own a sewing machine (proxy for entrepreneurship) but had no effect on a ROSCA membership (proxy for empowerment) or happiness at home.

As mentioned above, the evaluators also analyzed the **barriers to program completion**:

- Women who completed secondary schooling were more likely to complete the training in tailoring and sewing.
- The distance to the training centre was identified as a barrier. Women who lived further away from the training centres were less likely to complete the program.
- The lack of proper childcare support was an additional barrier. Married women with childcare support (e.g., mother-in-law present in households) were more likely to complete the program.

Conclusion on Methods Used

In close collaboration, the **evaluators and the implementing NGOs have developed and implemented an RCT** to study the effect of a training program in tailoring and sewing for women from low-income areas in India. RCTs are commonly considered the gold standard in measuring causal impact. RCTs compare relevant labor outcomes of a control group and a participant group, who have been randomly assigned, thus allowing the identification of the causal impact of the training, meaning the impact evaluation can be considered **rigorous**.

Lessons Learnt

Two important **policy implications** emerge from the results of the evaluation conducted in India. First, investing in vocational training programs **can result in significant economic benefits for women** from low-income households in developing countries. Second, **constraints** on accessibility, credit/financial resources, and the availability of in-home childcare support are critical and can discourage women from entering and completing any educational program. For the **design of a training program**, the following factors might be considered:

- **Needs-oriented training:** Training programs should be tailored to the needs and preferences of their target group. This evaluation concludes that program participation and completion is higher if the training content is tailored to the interests of participants, in this case, sewing and tailoring.
- **Identifying constraints:** To ensure a high participation and completion rate, it is critical to identify barriers. The researchers found several barriers that hindered women from participating in the training. In conclusion, training programs should be carefully designed to address those barriers (e.g., proximity to training center, providing childcare, ensuring safe and affordable commuting options to the training institutions, timing that does not interfere with their household chores and responsibilities).
- **Collaboration:** As mentioned above, this RCT was designed and implemented in close cooperation between researchers and NGOs, which was highly beneficial for the quality of the evaluation.
- **Evaluate different training designs:** In this RCT, all trainings offered were identical in content and method. It might thus be interesting to compare different training designs in an RCT to identify which designs (such as, varying the length and type of curriculum) are most effective in achieving the objectives.

Reference



Pushkar Maitra, Subha Mani, Learning and Earning: Evidence from a Randomized Evaluation in India. Published in: Labour Economics (Special Issue on Field Experiments in Labor Economics and Social Policies), 2017, 45: 116-130, <https://www.sciencedirect.com/science/article/abs/pii/S0927537116303384> or <https://docs.iza.org/dp8552.pdf> [accessed online on 01.03.2023]

4.2 QUASI-EXPERIMENTAL DESIGNS

4.2.1 CASE STUDY 3: SERBIA – EMPLOYMENT IMPACTS OF GERMAN DC INTERVENTIONS

Project Description

Title	Employment Impacts of German Development Cooperation Interventions: A Collaborative Study in Three Pilot Countries
Commissioned by	Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH (GIZ)
Implementing organization	Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH (GIZ)
Implementing partners	Institute for Improvement of Education and Upbringing, Ministry of Education, Science and Technological Development of the Republic of Serbia (MoESTD), Serbian National Employment Service (NES)
Research institute	Leibniz Institute for Economic Research (RWI), Essen Germany
Project area	Serbia (the report also references Jordan and Rwanda)
Target groups	Youth aged 15 to 35 years
Project term	Reform of Vocational Education and Training project in Serbia (TVET project): 01/2016-12/2019 Youth Employment Promotion (YEP) project: 07/2015 – 12/2019
Project cost	TVET project: 4 million USD YEP project: a total budget of almost 10 million USD
Evaluation term	3 years (accompanying evaluation; the evaluators were not involved during project planning, but rather during project implementation)
Evaluation cost	More than 124.000 USD GIZ-funds ²⁴ for three IEs in Serbia, Jordan and Rwanda and additional research funds
Evaluation design (evaluation approach)	Quasi-experimental design (quantitative difference-in-difference approach for the TVET project and statistics matching approach for the YEP project)
Publication date	08/2019

Project Context and Results

GIZ implemented several projects (on behalf of BMZ) under the overarching **Sustainable Growth and Employment in Serbia** program. The GIZ program is aimed at supporting companies to be more competitive and to supply or create jobs, so that job seekers can benefit from these measures to find employment. Within this program, two projects were selected to implement a rigorous IE; the **TVET project** and the **YEP project**. The objective of both projects was to integrate young people into the labor market. The **TVET project aimed to improve the offer of demand-oriented cooperative education in technical professions** in the formal Serbian TVET system, by introducing elements of dual training in 3-year TVET profiles. The **YEP project developed local employment initiatives** (like additional skills trainings, employment in hubs and rural areas, internships, career guidance and counselling for vulnerable groups), and supported 21 social enterprises in order to improve the labor market integration of disadvantaged groups.

Evaluation Objectives and Indicators

The evaluation took place in close cooperation between the GIZ team in Serbia, the GIZ Sector Project Employment Promotion and the RWI evaluators. The overall evaluation objective was to **test rigorous but practical and cost-efficient solutions**, which can be **replicated or up-scaled** in similar programs. For this purpose, the evaluation analyzed the **employment impact** of two separate projects of the **Sustainable Growth and Employment in Serbia program**.

The **TVET project** aimed to improve the employment prospects of graduates from the Serbian vocational education and training system. To achieve this, the project modernized six occupational profiles by adding elements of dual training in 52 vocational schools across Serbia. These schools cooperate with 200 companies and jointly offer a dual training program to students. Approximately 2,700 students were trained in these occupations.

The **YEP project** supported Serbian unemployed youth in improving their labor market outcomes (quality of educational profiles, employment status and job characteristics) by implementing active labor market measures. The evaluation estimated the impact of two types of short-term skills trainings: (1.) matching youth to employer-based trainings offered by cooperating firms; and (2.) trainings in simulated workplace environments conducted by vocational training institutes.

²⁴ This is equivalent to 125.000 EUR in November 2022. The exchange rate of the EC was used for the conversion ([EC 2022](#)).

Evaluation Approach and Methods

For the IE of the **TVET project**, a **difference-in-difference (DiD) design** was used to assess the causal effect of graduating from a school with a modernized TVET profile. The DiD methodology compares the outcomes of students enrolled in modernized profiles to comparable students enrolled in non-modernized profiles within and across schools. Therefore, the treatment group was students attending an intervention profile in an intervention school. The treatment group was compared with three control groups: (1.) students attending a non-intervention profile in an intervention school; (2.) student's attending a profile similar to the modernized profile in control schools; and (3.) students attending a non-intervention profile in a control school. Using three control groups enabled comparison within and between schools.

The impact evaluation of the **YEP project** measured the project's effects on participants' labor market outcomes (employment status, formal employment). For this purpose, two datasets were combined: (1.) large-scale administrative data provided by the NES; and (2.) primary phone survey data was collected among training participants. This enabled an estimation of causal effects of participation in the YEP project on the labor market outcomes of 916 beneficiaries (treatment group). The impact evaluation applied **statistical matching** procedures to identify similar unemployed individuals among 1.5 million registered unemployed that did not participate in the training (for the control group).

Focal and Cross-cutting Topics (including limitations of the evaluation)

- **Use of existing M&E data:** The evaluators tried to **incorporate existing M&E systems and data** in close collaboration with the GIZ M&E team and local researchers. The local researchers were contracted specifically to handle and collect additional data, because the existing M&E data required some augmentation via surveys. Existing M&E systems are usually not geared to fulfil the requirements of tailor-made rigorous IE designs (see *conclusions on methods used below*).
- **Use of existing administrative data:** The **use of existing administrative data** required close collaboration with national stakeholders as well as their support and interest in the research, a critical quality assessment of the data as well as knowhow in understanding and analyzing the administrative data (with the respective tools for accessing specific dataset formats and quantitative data analysis). In case of the TVET project, the **coordination with national stakeholders** was key for the successful impact evaluation and implementation of the DiD design (quasi-experimental method). The Institute for Improvement of Education and Upbringing helped to identify comparison profiles and comparison schools as suitable control groups, while the MoESTD provided **additional administrative data** on enrolment scores and established contact with these schools. In case of the YEP project, access to large-scale administrative data from the NES enabled the used of **statistical matching methods** (quasi-experimental methods) and creating a control group. During implementation, many privacy and data protection concerns had to be solved before administrative data could be used (e.g., generated large, anonymized datasets).
- **Gender-sensitive impact assessment:** Due to the small sample size, the impact of a modernized TVET on the employment outcomes of underrepresented groups such as women or the Roma population was not analyzed. The IE of the YEP used a dichotomous variable with the categories "female" and "male", but the findings were not discussed.

Key Evaluation Findings

The modernized TVET profile includes a dual training component. In total, 52 TVET schools cooperate with 200 companies where students were able to receive cooperative employment-oriented TVET in Serbia. The key evaluation findings of the **TVET project** showed that **graduating from a modernized, employment-oriented TVET profile had a positive impact on perceived education quality and characteristics of employment** (e.g., hours worked weekly, type of contract and income). Graduates from modernized profiles were more satisfied with the quality of education, reported better school conditions, perceived themselves to be more prepared for working, and were more likely to claim that they would choose the same TVET again. However, there was no measurable impact on the overall probability to be employed six months after graduation from the modernized TVET profile. Students in modernized profiles were more likely to obtain their first job in the training companies. They had a higher likelihood to use their TVET skills and knowledge in their current job, and to earn higher wages. Especially the last finding indicates an important effect of the intervention towards improved long-term labor market success induced by the TVET reform.

The **YEP project** implemented **two different and separate types of short-term skills training**: (1.) "Training at employer's request" which matched youth to firm-based trainings at private-sector employers (employer-based training); and (2.) "Training for labor market needs" subsidized training set in simulated workplaces of accredited vocational trainings institutions (vocational training institute-based trainings). The key evaluation findings of the YEP project showed, on the one hand, that **employer-based training had a sizeable and sustained impact on registered formal employment**. One reason for this was that participants were largely hired and retained by the training firm. Even though an increasing share of the control group found jobs within 8 months after training end, the impact evaluation suggested that participants of the YEP training had a 45 % higher probability of employment. This is a large impact quantitatively. On the other hand, **vocational training institute-based trainings had also a positive impact on formal employment**,

which took longer to emerge. After 8 months, the probability to be registered as employed was 16 % higher than in the absence of the YEP project. Furthermore, medium-run trends showed that the gap to the control group widened over time. Sub-sample analysis for early training cohorts suggested that the impact increased to more than 22 % after 16 months, which indicated a sustained gain in human capital. Additionally, the survey data showed that a large share of the non-registered employment participants were likely informally employed. According to the survey data analysis, the majority of employed participants in both types of training were very satisfied with their employment, were working in the field of the GIZ YEP training and reported earnings around the national median wage.

Conclusion on Methods Used

The following main conclusions and lessons learned are drawn from the rigorous impact evaluation in Serbia (as well as from two further case studies from Jordan and Rwanda, which are also included in the same research report):

- This research shows that it is possible to rigorously assess employment effects of DC interventions in **close collaboration between DC practitioners and academics** over a three-year evaluation period. The collaboration succeeded in devising **tailor-made rigorous IE designs** and **collecting corresponding data** to measure employment impacts of **relevant and evaluable DC measures** at a country, module and intervention level.
- The empirical findings show that German DC interventions have **significant positive, and to some extent large, employment impacts**. The Serbia TVET results show that graduating from a modernized TVET profile with dual elements has a positive impact on perceived education quality and characteristics of employment (incl. hours worked per week, type of contract, spell termination reason), while the YEP impact evaluation found that employer-based training has a large and sustained impact on registered formal employment. It also found that vocational training and institute-based training effects are equally large and materialized especially in the long-term.
- Differential impacts across the range of interventions give **important feedback for steering and future program design**. GIZ learned from the impact evaluation that modernizing TVET is a promising approach and that disadvantaged youth can be supported effectively through on-the-job training. The modernized TVET profile added elements of dual education in secondary schools. Specifically, a three- or four-year TVET program was offered that prepared students to work in a specific occupation. By partly attending classes at school and partly attending training with the company, elements of dual education were successfully added to secondary school education.
- It is worth assessing the possibility to **cost-efficiently collect data** for IE about the employment effects **from existing M&E and administrative data sources**.
- For any rigorous IE, it would be ideal to already **start the collaboration** and exchange between intervention practitioners and researchers **when designing the intervention or when starting it**. The collaboration for these rigorous IEs started already at the outset of program implementation and had a three-year period for rigorous IE implementation, which enabled: (1.) the creation of rigorous and practicable rigorous IE designs; (2.) the collection of the required data; and (3.) thus producing meaningful and informative impact results.
- Even though these rigorous IEs constitute a good practice for integrating existing M&E systems and practice, the report highlights the importance of **bringing together “project thinking” and “research thinking”**, i.e., the practitioners’ perspective on the implementation of the respective intervention and the researchers’ perspective on what constitutes an appropriate rigorous IE design. This requires efforts to understand each other’s objectives, constraints and modus operandi. Researchers have to learn how interventions work and how they are to be evaluated using existing M&E data. Practitioners have to understand why researchers require a control group, stress the importance of the issues of selectivity, randomization of treatment, large sample sizes and comprehensive data for solid empirical evidence. Practitioners would have benefitted from a training session about different rigorous IE approaches. The researchers would have preferred to start the collaboration even earlier, when the intervention’s main results logic was set up, as this would have helped to develop more detailed pathways to achieve outcomes, which can be tested empirically as part of the IE.
- To be able to engage in program-accompanying rigorous IE, DC programs need **additional resources on top of their regular M&E staff** (even if they are collaborating with external researchers). Adequate budget supplements should be earmarked during project design phase. For example, in the Serbia rigorous IE, a local research institute was contracted to handle and collect data and was therefore able to provide the link between program operators and external researchers from the RWI team.



The complete report which this case study is based on, contains an additional **rigorous IE case study of the GIZ Employment Promotion Program (EPP) in Jordan**, which presents the results of a DiD design implementing a homogenous impact assessment approach across a broad range of smaller-scale labor market interventions that were implemented. Evidence from the EPP Jordan shows that **labor market matching interventions** had the largest and most consistently positive employment effects in Jordan (*see reference of the complete report for more information below*).



The full IE report also contains a third **case study on the Eco-Emploi Program in Rwanda**. A RCT **was designed**, but failed during project implementation, because the uptake of this **coding training for women** (WeCode intervention) turned out too low to enable an experimental IE design. Since few people registered during the short sign-up period, the sample size turned out low. Even though a control group was planned, these individuals were trained and moved to the treatment group during implementation (*see reference of the complete report for more information below*).

Reference



Bachmann, Ronald; Kluve, Jochen; Martinez Flores, Fernanda; Stöterau, Jonathan (2019): Employment impacts of German development cooperation interventions: A collaborative study in three pilot countries, RWI Projektberichte, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen. Project report commissioned by "Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH". Final report. August 2019. <http://hdl.handle.net/10419/215904> [accessed online on 07.09.2022]

4.2.2 CASE STUDY 4: KENYA – EMPLOYMENT & INCOME EFFECTS OF SKILLS DEVELOPMENT INTERVENTIONS

Project Description

Title	Employment and Income Effects of Skills Development Interventions: An Impact Evaluation of Three Employment Promotion Measures in Eastern Africa within GIZ's Employment and Skills for Development Program
Commissioned by	BMZ
Implementing organization	GIZ
Implementing partners	Kenya Association of Manufactures (KAM), national training providers, member companies
Research institute	Leibniz Institute for Economic Research (RWI), Essen in Germany; Innovations for Poverty Action in Kenya
Project area	Kenya (the report contains additional research covering Uganda)
Target groups	Youth up to 24 years old, (specifically TVET certified or diploma graduates of selected trades who graduated in the past 5 years)
Project term	10/2017 – 10/2019
Project cost	Unknown (large project)
Evaluation term	2017 – 2021 (5 years)
Evaluation cost	Not available
Evaluation design (evaluation approach)	Quasi-experimental design (The quantitative evaluation of the KAM program used a linear multivariable regression model similar to a Difference-in-Difference approach, which controlled for participant's background characteristics and baseline employment outcomes)
Publication date	06/2021

Project Context and Results

The program, Promoting Youth Employment Through Technical Human Capital Development, was an employment promotion program within the BMZ's Employment and Skills for Development in Africa (E4D) program implemented by GIZ and KAM in collaboration with national training providers and member companies. The KAM program consisted of a **2-3-day work readiness training** (incl. mentorship workshops) and a subset of trained beneficiaries received an **internship placement** at KAM member companies for three to six months. These two KAM program components – if implemented jointly – improved access to jobs and economic opportunities for youth. The authors studied if the youths are employed or self-employed and whether the employment is decent, formal or fulltime. The 2-3 day work readiness training alone did not have any impact.

Evaluation Objectives and Indicators

The **evaluation objective** of the study was to assess the impact of the KAM program on employment and labor market outcomes among vocational training graduates in Kenya. This IE focused on the effectiveness of receiving the two program components (work readiness training and internship placement), which aimed at overcoming a skills gap related to youth employment. The skills gap refers to the difference between the skills of young graduates and the actual skills demanded by employers and is considered a major challenge for youth in Eastern Africa. The two program components tackle the lack of practical experience of vocational training graduates and related difficulties in the school-to-work transition. Two RQs were answered by the quantitative evaluation of the KAM program:

- **RQ 1:** What are the gains in employment and earnings for vocational training graduates from participating in the work readiness training and internship placement of the KAM program? This means that the participation in both program components was compared to not participating in the KAM program at all. It did not study the effect of the participation in one single component.
- **RQ 2:** What are the gains in employment and earnings for vocational training graduates from an internship placement organized by KAM in addition to participation in the work readiness training organized by KAM? This means that the additional impact of participating in the internship placement on top of the benefits of the work readiness training was studied. Survey participants who took part in only the work readiness training were compared to those who participated in the work readiness training and were also placed in an internship.

Since the KAM program aimed at improving labor market prospects and outcomes for young Kenyans, the impact evaluation assessed **direct employment benefits** as primary outcome variables. These included multiple dichotomous **primary outcome indicators** like employment status, self-employment, formal employment, full-time employment)

and decent employment²⁵ and further income indicators measured in Kenyan Shillings. Multiple **secondary outcome indicators** were assessed as well, like aspiration for further education, employment aspiration, family structure, and banking and saving behavior (*see reference, evaluation report pages 32-34 for more information below*).

Evaluation Approach and Methods

KAM **beneficiaries** were students who graduated from TVET institutes during the past 5 years and **control group participants** were those students who already completed or were about to graduate from TVET facilities at the time of the **baseline survey**. The treatment group was comprised of two groups, one of which received the training and the internship placement and the other that received only the training. The treatment group participants were interviewed four times, while the control group participants, which had not received any KAM benefits, were interviewed three times both over a period of two years (longitudinal data). These surveys collected information about the participants' **current and retrospective employment status**, earnings, and socio-demographic background characteristics to be able to answer the RQs.

As the most rigorous IE applicable to this context, a **linear multivariable regression model** was used (in a similar way to a **DiD** design) to measure the impact of the KAM program on job search, employment and income indicators (primary outcomes). Furthermore, the program's effects on employment aspirations, family structure, as well as banking and savings behavior (secondary outcomes) were also analyzed. The impact evaluation took account of beneficiaries' heterogeneity with respect to gender, age, and prior sustained work experience.

Limitations of the evaluation: The main methodological challenges were: (1.) That not all participants of the work-readiness training received an internship and others rejected the offer. This was a challenge for the impact evaluation because of their selection, but it was not a challenge for the project implementation. (2.) The allocation of internships did not occur randomly, which prevented the use of an experimental RCT design. The allocation of internships was based on participants' merit and skills, so that participants with better labor market prospects were more likely to receive an internship placement than those with lower prospects. (3.) The length of the internship and time between completion and interviews differed across participants. The second and third methodological challenges led to considerable heterogeneity in the intensity of the program and time of program completion. Due to considerable heterogeneity, the treatment and control groups differed in their characteristics. Comparable groups were relatively small, which reduced the significance and made the samples less representative.

Focal and Cross-cutting Topics

- **Value-for-Money:** Depending on the assumptions made, the cost-effectiveness ratio of the KAM program was 0.00021-0.00036 jobs per EUR invested or 2,778-4,762 EUR²⁶ per job. A common tool to evaluate (not to implement) "value-for-money" analysis is the cost-effectiveness analysis, which summarizes a complex intervention in a ratio of total impact to total costs and allows comparisons of interventions easily. The report applied a step-by-step guide for **cost-effectiveness analyses** to the KAM program. The evaluation report conceptualized value-for-money following the four E's framework – economy, efficiency, effectiveness and equity by the Foreign Commonwealth & Development Office²⁷ (*see reference, evaluation report pages 243-275 for more information*).
- **Sustainability of impacts:** Longitudinal data was collected, including a baseline survey and four interviews with the treatment group (graduates from intervention schools who choose to participate in the program). The control group (graduates from non-intervention school and graduates at intervention schools who did not choose to participate in the program) was interviewed three times. The tracer study collected interview data at different points in time (9, 12 and 24 months after treatment), which enables the assessment of the long-term impacts of the training and therefore the sustained project impacts up to two years after program ends.
- **Gender-sensitive impact assessment:** The evaluation assessed the program effect heterogeneity by respondents' gender, age and prior work experience and the key findings are described below (*see reference, evaluation report pages 71-73 for more information*).
- **Private sector involvement:** The E4D initiative was implemented by KAM in collaboration with national training providers and member companies, so that the private sector was strongly involved in the implementation of the TVET interventions covering both work readiness training and internship placement. The quantitative impact evaluation (described in this case study) collected data from graduates only. However, there was an additional qualitative evaluation, which gathered data from the companies' involved in the implementation using semi-structured interviews with company representatives. The qualitative study focused on investigating how a change in the internship stipend funding from E4D to companies affected those companies' ownership and sustainability of internship placements of the KAM program (*see reference, evaluation report pages 80-111 for more information*).

²⁵ Decent employment combines having paid work for at least 20 hours per week and a minimum income from that work of at least 6,209.93 KES per month.

²⁶ One EUR is equivalent to 0.9951 USD in November 2022. The exchange rate of the EC was used for the conversion ([EC 2022](#)).

²⁷ Formerly Department for International Development (UK).

Key Evaluation Findings

Answer for RQ 1: Vocational training graduates who participated in the work readiness training and internship placement of the KAM program experienced a significant improvement in their labor market outcomes.

The large and persistent effects on decent and formal employment, as well as on income, were striking. The results show an improved job search performance, a reduced financial dependency, and an increased probability of having a bank account. In detail, this means that the effect sizes on **full-time employment** were positive and large, but this effect was not significant anymore 24 months after the baseline survey, while the effects on decent and formal employment continued to be significant after these 24 months. Additionally, participants of the treatment group significantly **improved their incomes** by 52.8 % and these improvements were sustained until 24 months after the baseline survey. **Working hours** increased only by a few hours per months and not significantly. The increased income was due to **higher pay per hour worked**. The **pay per hour worked** rose significantly and sustainably by 37.2 % after 24 months, which is an **extraordinarily high effect size**. Due to the KAM training and placement program, the **probability for an formal job interview** increased significantly by 19.0 percentage points and for a full-time job interview by 18.0 percentage points until 24 months after the baseline survey, so that this effect was perceived as sustainable. The program significantly as well as sustainably lowered **participants' financial dependency** on the household head by 18.3 percentage points and it increased participants' **probability of having a bank account** by 9.9 percentage points. The effect sizes on **employment and income** tended to be larger for **participants without prior work experience**, although the coefficients of the two subsamples with and without experience did not differ from each other significantly. Concerning **gender**, the increase in decent, formal, and full-time employment were particularly large for women 16 to 24 months after the baseline survey, although the effect sizes for women were smaller than those for men at earlier follow-up periods. Verification of the results using the control group showed that **not all effect estimations were robust** in excluding National Competence Based Education and Training program (CBET) graduates from the control group observations. Overall, the effects remained significant and of similar magnitude for the time 10 to 15 months after the baseline survey, however the previously mentioned effects in the longer term (up to 24 months after the baseline survey) turned insignificant. When additionally restricting the sample to treatment and control group observations that attended the same TVET institutions **most effects ceased to exist**. The only results that **remained significant are short-term effects (3-9 months) on participants' job search outcomes and their probability of having a full-time employment** as well as having a **formal employment** for the time 10 to 15 after the baseline survey.

Answer for RQ 2: The positive effects on labor market outcomes were mainly driven by the work readiness training rather than the internship placement.

This means, that the estimation results suggested that the **impact** of receiving an **internship placement in addition to participating in the KAM training** was **small and insignificant**. Although, there were significant positive effects of receiving the KAM internship placement (in addition to the training), which increased decent employment on average by 12.7 percentage points and formal employment by 10.3 percentage points in the short term (i.e. 3 to 9 months after the baseline survey). **These positive effects did not sustain in the longer term** (until 24 months after the baseline survey). The **heterogeneity analysis** suggested that the **short-term effects on decent and formal employment were driven by participants without prior work experience**. Except for these short-term impacts on decent and formal employment, there were **no significant effects** of the KAM internship placement in addition to the training on other employment and income outcomes. Similarly, there were no effects on job search outcomes, such as the probability or number of job interviews. However, 16 to 24 months after the baseline survey, KAM beneficiaries were significantly more likely to have dependents and a bank account by 9.6 and 5.9 percentage points, respectively. The **heterogeneity analysis** of RQ2 enabled interesting findings: The KAM's placement component had significant **negative short-term impacts on women's general employment status** (i.e. not decent or formal employment), **fulltime employment** and total **monthly income** 3 to 9 months after the baseline survey. These **negative effects stopped** at later periods.

Conclusion on Methods Used

While the results of the KAM program suggest that the skills training rather than the internship placement was effective in improving employment outcomes, a further case study in Uganda (*covered in the same report, see reference below*) showed reversed effects. The evaluators therefore concluded (in line with other academic literature) that the **effectiveness of employment promotion programs or specific components is highly dependent on the local context, program design aspects and the target group**.

The authors derived the following recommendations from the IE for future program designs:

- A careful assessment of the context and the needs of the target groups should be conducted prior to the planning and implementation of employment promoting programs.
- Project partners must have sufficient resources for program implementation and monitoring.
- Quality assurance can be incorporated in the program design as project outputs (e.g. refresher trainings for trainers when interventions are delivered through a training of trainers).
- The private sector should be involved in program design to ensure context suitability and intervention quality.

The main lessons learned which the research team derives from this author's evaluation and the related interview for future impact evaluations in DC:

- Rigorous IEs of development projects require **close collaboration between researchers and practitioners**, because both parties follow different objectives and have different perspectives. Close collaboration and continual communication during project implementation are key (especially between local project managers and researchers). This may create an additional **workload for local project managers**.
- **The impact evaluation should be integrated in the project planning and implementation from the beginning** because IEs and the involvement of researchers should start at the conception phase of the interventions. It is difficult to implement rigorous IEs in many DC contexts, because DC IEs are often carried out ex-post, which makes it difficult to close the gap between practice and research. From a research point of view, **the specification of RQs** should take place before interventions are planned and implemented. This would be beneficial for the setup of a rigorous IE and can manage expectations.
- **The timing of intervention implementation and data collection is important**. From a research perspective, as many aspects as possible should be held constant during the intervention implementation, to be able to understand to which extent treatment and control groups are comparable and to be able to study where treatment effects come from. This may constitute a challenge for practitioners, who apply a “managing for results principle” and consider flexible planning and implementation as key for achieving results in a dynamic and constantly changing environments (as is the case in many DC project contexts). **Sufficient time has to pass** between the intervention completion and the follow-up data collection for treatment effects to unfold, potentially beyond the project phase. A **standard implementation period** of approximately three years in DC creates a challenge for rigorous IE because research usually needs more time.
- **Reliable results of rigorous IEs depend on large sample sizes**, so that **rigorous IE cannot be applied to all DC project contexts**. Many researchers would like to randomize the treatment for RCTs (gold standard for rigorous IE).
- In most cases there is a need to **collect their own data, because existing data (like M&E or administrative data) might not be of sufficient quality**. Existing data sources often use too small sample sizes for rigorous evaluations and do not provide information about a suitable control group. Household survey data exists in many countries but cannot be used for most rigorous IEs, because attributing DC project measures to beneficiaries is usually not possible. Most rigorous IEs require their own data collection, which also increased the costs of these evaluations.
- **Results from individual IEs are difficult to be transferred to another context** (such as the LAC region). Instead, conclusions can be drawn from the totality of existing rigorous IEs, specifically systematic reviews (SRs) of rigorous IEs. There are already some SRs of rigorous IEs and a **very diverse academic literature on DC TVET measures** that exist, which should be considered when selecting and planning new project measures. Researchers often observe a more idiosyncratic than results-oriented selection of project measures. For example, specific TVET training measures might be offered because the responsible project manager was trained in the same subject).
- **However, the learnings from the use of rigorous IE methods can easily be transferred to other contexts (such as LAC)**. The following are some examples of what that can include:
 - ♦ For ex-post evaluations, which are often requested in DC contexts, ideally the **selection mechanism of the beneficiaries should be precisely documented**. This is an important measure to construct adequate control groups for quasi-experimental IEs, but practitioners are often not aware of this.
 - ♦ For rigorous IE and **detecting impacts during research**, it is helpful to **clearly specify implementation periods and exact evaluation dates**. In DC, the rollout of measures is often spread over a long period of time, which constitutes a problem for rigorous IE. If the rollout of DC TVET measures takes place over 2 years, this can mean that some people are just newly trained at the end of the project, which makes it very unlikely to be able to detect impacts already.
 - ♦ **Rigorous IE require high quality data**, which makes data collections generally very important. Existing M&E and administrative data can be used in rather few cases and only if multiple challenges are overcome (see *Case Study 3*).
 - ♦ It is important to invest time at the **beginning of a rigorous IE** to bring together research and practice thinking and to understand each other's priorities.
 - ♦ From the researcher's perspective, the **provision of local resources is critical to the success of a rigorous IE**. For an evaluation to be succeed, it can be beneficial if the need for the rigorous IE comes from the project staff (before or during project implementation), as this means there is already interest in the research and resources are more likely to be provided for experimental or quasi-experimental research designs by local project leaders.

The evaluation report contains further case studies covering quantitative evaluations of the “ReadyToWork” program in Uganda and a “Skills for Construction” program in Uganda and additional qualitative RQs on the KAM program in Kenya, which are not described in this case study (see *reference, evaluation report for more information below*).

Reference



Ebert, Cara; Flörchinger, Daniela; Frohnweiler, Sarah; Ihring, Stephanie;

Rosadio Cayllahua, Karen Micaela (2021): Employment and income effects of skills development interventions: An impact evaluation of three employment promotion measures in Eastern Africa within GIZ's employment and skills for development program, RWI Projektberichte, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen. <http://hdl.handle.net/10419/251877> [accessed online on 07.09.2022].

4.2.3 CASE STUDY 5: BRAZIL – NON-COGNITIVE SKILLS AND LABOR MARKET OUTCOMES

Project Description

Title	Technical Education, Non-cognitive Skills and Labor Market Outcomes: Experimental Evidence from Brazil
Commissioned by	Coordination of Improvement of Higher Education Personnel (CAPES); Inter-American Development Bank (IDB)
Implementing organization	National Program for Access to Technical Education and Employment (<i>Bolsa-Formação PRONATEC</i>)
Implementing partners	<i>Serviço Nacional de Aprendizagem Industrial (SENAI SC)</i> ; <i>Serviço Nacional de Aprendizagem Comercial (SENAC SC)</i>
Evaluation institute	<i>Fundação Getúlio Vargas (FGV EESP)</i>
Project area	Santa Catarina State, Brazil
Target groups	Current and former public high school students
Project term	Since 2011
Project cost	Depends on application to national program
Evaluation term	June to August of 2016
Evaluation cost	About 20,000 USD
Evaluation design (evaluation approach)	Quasi-experimental design (quantitative, natural randomized experiment)
Publication date	2021

Project Context and Results

The impact evaluation in Brazil focuses on a student training scholarship, a policy under a large public program — the National Program for Access to Technical Education and Employment (*Bolsa-Formação PRONATEC*) — that offers **scholarships so that eligible youth can attend TVET courses** free of charge. The PRONATEC was created by the Federal Government of Brazil in 2011 to expand educational opportunities and qualified professional training for young people, workers and beneficiaries of income transfer programs. The scholarships allowed interested eligible individuals enrolled in high school or high school graduates to **attend two years of in-classroom occupational training** (e.g., in mechanics, workplace safety, computer networks, electro-technology, food technology, and informatics). As in most TVET programs, the courses focused on providing trainees with **occupational skills**. Moreover, they explicitly aimed to develop **soft skills** such as communication, creativity and the ability to work autonomously. The authors of the evaluation found that the **TVET courses had a positive effect on female labor market participation**, employment of females as well as women's wages. Interestingly, no impact was found for male graduates of the TVET courses.

Evaluation Objectives and Indicators

The experimental evaluation aims at identifying the causal effect on **labor market outcomes as well as non-cognitive skills** from being offered the opportunity to participate in the TVET courses through the scholarship. Specifically, the authors of the evaluation study the impact on the following outcomes:

1. **Labor market outcomes** such as employment, labor market participation (i.e., if the individual either had a job or was looking for a job), formal employment (signed contract), days and hours worked, duration of employment, work earnings and area of work (i.e., if the person found a job in the field of training).
2. **Non-cognitive skills** such as the agreeableness (i.e., the tendency to act cooperatively), conscientiousness (i.e., the tendency to be organized and responsible), extraversion (i.e., orientation towards external world), neuroticism (i.e., predictability and consistence of emotional reactions), openness to experiences and locus of control (i.e., how much individual attribute experiences to past decisions).

The authors theorize that **workplace-based programs also teach non-cognitive skills**, such as teamwork, discipline and responsibility, leadership, and flexibility either by explicitly teaching and exercising these skills or by providing an environment where they can be developed through the interaction with peers and teachers. Moreover, these acquired non-cognitive skills could also impact labor market outcomes.

Evaluation Approach and Methods

The evaluation of the TVET program is based on a **natural experiment with a random-like assignment** of an intervention and control group. The classes were offered by two of the main TVET providers in Brazil, which faced excess demand and chose to **admit applicants through randomization** in 2012, 2013 and 2014 in four mid-sized municipalities in Santa Catarina State. Since the authors of the evaluation were not involved in the program planning and had **no control over intervention implementation and randomization of beneficiaries**, this study can be classified as a natural

experiment in which the TVET providers decided to implement a random course admission due to limited training spots. The authors call the approach a “waiting list RCT”.

Randomization: In the evaluation period, the two TVET providers offered 29 classes where **the number of applicants exceeded the available training capacities**. Each class had 35 available spots for stipends and around 70 applicants (note that in some cases, more than 70 students had applied but the waiting list only included the first 70 individuals). Therefore, the TVET providers **randomized the applicant list** for each class and followed the randomized order of the list to fill all open slots. If applicants rejected the offer, the slot was given to the next applicant on the list. Based on the randomized admission, the evaluation team was able to retrospectively construct a treatment group (i.e. applicants who received an offer to participate in the training) and a control group (i.e. applicants who did not receive an offer). Moreover, students without a scholarship were in the same classes.

Data collection: In cooperation with the TVET providers, the authors of the evaluation received the **administrative contact data** (name and phone numbers) from all 70 applicants per class in the four municipalities for the years of 2012, 2013 and 2014. They collected the **survey data in 2016 through in-person and telephone interviews**. The survey contained questions on demographic and socio-economic characteristics and labor market outcomes and non-cognitive skills. Moreover, they included extensive questionnaires about behaviors to cover non-cognitive skills, which relied on self-reporting instead of interviews through enumerators. The authors of the evaluation were able to track and interview 735 individuals (237 women and 498 men) of which 126 were in the control group (i.e. no training offer received) and 609 were in the treatment group (i.e. training offer received). Regarding the self-reporting instrument on non-cognitive skills, 376 individuals responded (111 women and 258 men).

Limitations: Since the training intervention was not initially planned as an RCT but rather constitutes a natural experiment, several methodological challenges arose. Firstly, the **sample size is relatively small** with a control group of 126 individuals and a treatment group of 609 participants. When studying the female and male subsample, the sample size becomes smaller. The sample on non-cognitive skills is even smaller. Secondly, the **randomization is imperfect** to the extent that selecting the first 70 individuals can be regarded as random sampling. In some cases, more students had applied to the training but only the contact details of the first 70 randomized applicants were kept by the training providers. Moreover, there were issues of leakage, meaning that some individuals have participated in the training although they did not receive an offer. Thirdly, due to the constraints, the study results are **limited in external validity** meaning that the findings are probably not representative for students outside of the studies municipalities.

Focal and Cross-cutting Topics

- **Use of existing (M&E or administrative) data:** The evaluators were able to rely extensively on administrative data from the two TVET providers. Through the training institutions, they received the names, phone numbers, and information on gender and education from the applications on the waiting list as well as training participants. Based on this data, the researchers were able to track beneficiaries and conduct further surveys.
- **Gender-sensitive impact assessment:** The researchers specifically studied the impact on female and male students to investigate whether it differed. The information on the gender of the training graduates allowed them to split the sample into a male and female sub-sample. Interestingly, the training effects differed significantly for men and women.
- **Measurement of non-cognitive skills:** The evaluation team measured several non-cognitive skills through self-reporting on questions about individuals' behavior. The measurement of non-cognitive skills is based on Brazil's Social and Emotional Nationwide Assessment inventory, which was developed by the Ayrton Senna Institute.
- **Sustainability of impacts:** The authors of the evaluation evaluated the impact on labor market outcomes and non-cognitive skills around two years after training completion. This allowed them to assess the effect in the longer run. However, few studies investigate intervention effects long-term to identify whether the impact sustains.

Key Evaluation Findings

Considering the labor market outcomes, **women experienced large gains** while no significant effect was found for male beneficiaries. Compared to the female applicants that did not receive a training offer, women of the treatment group were more likely to be employed and to participate in the labor market and they earned more. In contrast, the training offer did not affect labor market outcomes of male participants.

Looking at the non-cognitive skills that were measured, the gender heterogeneity persisted. Women who received the scholarship offer **received a higher score on the extraversion and conscientiousness indicators** compared to women in the control group: there was no effect on males' non-cognitive skills. Potentially, the gender difference in labor market outcomes could have emerged through the acquisition of non-cognitive skills by women.

Conclusion on Methods Used

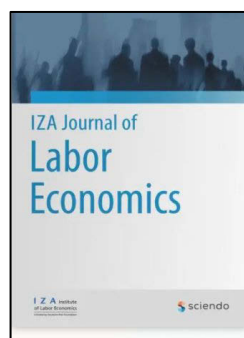
This evaluation is an **example for a natural experimental study that was conducted after the program had been implemented**. Typically, RCTs require methodological planning before an intervention takes place to ensure randomized treatment allocation, adequate documentation of the treatment and control group as well as the availability of suitable baseline data. In this case, the training allocation had been randomized by the training institutions due to high demand and capacity constraints. This **randomization and the availability of administrative graduate data** allowed the authors of the evaluation to evaluate causal effects after the training had taken place.

Content wise, the evaluation delivers novel information on **heterogeneous effects for male and female TVET beneficiaries**. The findings suggest that gender differences on labor market outcomes could have emerged through the enhancement of non-cognitive skills and that transmission channels need to be studied further.

Lessons learnt:

- The evaluation pointed out that the **impact of TVET programs may differ for male and female participants**. This emphasizes the importance of studying gender-specific training effects. Future evaluation should thus include a gender-sensitive impact assessment. Moreover, the potential channels/mechanisms should be studied in detail.
- Evaluations of TVET programs should **not only focus on labor market outcomes but also on non-cognitive skills**. Training courses can also affect this set of skills, which in turn are relevant for the labor market. For example, in Santa Catarina the evaluation contributed to a new technical education program explicitly designed for non-cognitive skills.

Reference



Camargo, Juliana, Lima, Lycia, Riva, Flavio and Souza, André Portela. "Technical Education, Non-cognitive Skills and Labor Market Outcomes: Experimental Evidence from Brazil" IZA Journal of Labor Economics, vol.10, no.1, 2021, pp.-. <https://doi.org/10.2478/izajole-2021-0002>

Kautz, T., et al. (2014), "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success", OECD Education Working Papers, No. 110, OECD Publishing, Paris, <https://doi.org/10.1787/5jxsr7vr78f7-en>.

Program links:

- <https://www.sed.sc.gov.br/programas-e-projetos/27461-pronatec>
- https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/br_9018.pdf
- https://siteal.iiep.unesco.org/eje/educacion_y_formacion_tecnica_y_profesional

4.2.4 CASE STUDY 6: PHILIPPINES – DUAL VOCATIONAL TRAINING

Project Description	
Title	Dual Vocational Training, Philippines
Commissioned by	BMZ
Implementing organization	GTZ, KfW, InWEnt, DED, CIM
Implementing partners	Technical Education and Skills Development Authority (TESDA), incl. 98 vocational training institutions
Evaluation institute	Centrum für Evaluation (CEval), Saarbrücken 2010 (Report authors: Silvestrini, Stefan; Garcia, Melody)
Project area	The Philippines
Target groups	Young women and men taking part in TVET; the trainers and institution staff
Project term	1996 – 2007
Project cost	30.83 million Euro (program costs)
Evaluation term	08/2009 - 02/2010
Evaluation cost	About 300,000 Euro, 286 work days for international experts
Evaluation design (evaluation approach)	Quasi-experimental and non-experimental design (the multi-method approach, including a quantitative, quasi-experimental approach based on propensity-score matching combined with a qualitative approach using explorative interviews)
Publication date	2010

Project Context and Results

The subject of the impact evaluation was the German contribution to the Philippine TVET system. The **evaluation comprised ten different contributions** in terms of programs, projects and measures implemented between 1996 and 2007 by *Deutsche Gesellschaft für Technische Zusammenarbeit* (GTZ), Capacity Building International (*Internationale Weiterbildung und Entwicklung*, InWEnt), the German Development Service (*Deutscher Entwicklungsdienst*, DED) (the three predecessors organizations of GIZ), as well as KfW and CIM. The German Implementing Organizations (GIO) focused on the **introduction and establishment of dual training programs²⁸ in the Philippine TVET system** to address the imbalance between the high numbers of school graduates that do not have labor-market relevant skills, and the high demand for properly trained, skilled workers. The different contributions improved graduates' satisfaction with their salaries and opportunities as well as their qualifications. However, positive effects on partner enterprises and training institutions were limited.

Evaluation Objectives and Indicators

The **aim of the evaluation** was to identify whether the German cooperation achieved its **programs' objectives** for the beneficiary (objectives 1 and 2 below) and institutional level (objectives 3 and 4 below). Therefore, the following overall objectives were evaluated and translated into concrete indicators:

1. Improved **employment situation of TVET graduates**, e.g., employment rates, usefulness of training, satisfaction with current salary, etc.
2. **Enterprises** increasingly participate in **dual trainings** and/or **employ graduates** from dual trainings, e.g., takeover of graduates, enterprises' ability to comply with legislative framework of the trainings, qualification of graduates, etc.
3. Improved **management and training capacities of public and private vocational training institutions**, e.g., training of trainers succeeded in reaching the target group, training institutions increasingly implement dual trainings, management efficiency improved).
4. Improved steering **capacities** of the Technical Education and Skills Development Authority (TESDA), e.g., management capacities improved; resources allocated to dual trainings increased; and monitoring the implementation of dual training systems (DTS).

Evaluation Approach and Methods

A **joint thematic country evaluation** among GIOs is rare, which makes this evaluation an interesting case. Since the evaluation is comprised of several programs and projects implemented by different organizations, they were not part of a joint concept. Thus, the evaluation team had to **develop a common results chain and corresponding indicators** to assess the joint contribution of all ten projects (see *list of 4 objectives and indicators above*).

²⁸ The evaluation also comprised so called "dualized" training programs/approaches, which refer to dual training approaches that do not comply with the regulatory framework (e.g. minimum wage for trainees, minimum duration).

The assessment of the German contribution to the TVET system in the Philippines was based on a **multi-method approach** that is comprised of both qualitative and quantitative data collection instruments (such as field surveys, online surveys, guideline-based interviews, structured group discussion), and analysis techniques to address the outcomes at beneficiary and institutional level. The evaluation began with a **qualitative exploration** to develop hypotheses and guide the data collection process. Therefore, the evaluation team conducted: a framework analysis based on national statistics (e.g., employment statistics); guideline-based interviews with TESDA staff; and structured group discussions with representatives from business associations. Hypotheses were developed that guided the qualitative and quantitative data collection process during the main evaluation mission.

To assess the impact of all ten projects on the identified outcomes, a comparative research design was applied. The IE was based on a **quasi-experimental method**, which included a comparative analysis of outcomes between the following groups:

- 197 graduates of supported (treatment group) and 112 graduates of unsupported training institutions (control group) to study objective 1.
- 26 representatives from partner enterprises of supported institutions (treatment group) and 11 representatives of unsupported institutions (control group) to study objective 2.
- 43 staff members from supported training institutions (treatment group) and 30 staff members from unsupported training institutions (control group) to study objective 3.

Quantitative data on the graduates were collected through field surveys using semi-standardized questionnaires. **Time and data constraints** prevented the use of econometric methods, such as RCTs and DiD approaches, to measure the impact of the DTS programs. RCTs require individuals to be randomly assigned to the treatment and control group at the start of the program or intervention, which was not the case for the TVET programs under study. Therefore, the evaluation team chose a quasi-experimental research design based on **propensity score matching**. Before calculating the difference in outcomes between the treatment group (graduates from training institutions supported by GIOs) and the control group (graduates from other training institutions that were not supported), the authors matched the graduates ex-post using propensity scores. Therefore, personal and socio-economic variables that could potentially influence treatment assignment were used (e.g., age, education, number of household members). This procedure ensures the **comparability of the treatment group and the control group** in the absence of random assignment before the intervention. By the inclusion of a comparison group and the comparison of baseline data (reconstructed by a retrospective interview and questionnaire design), and ex-post data, the evaluators were able to attribute the changes identified in the impact field of the program to the implemented measures.

Qualitative data to evaluate the other objectives were collected through guideline-based interviews with representatives from institutions' partner enterprises (objective 2) and with staff members of training institutions (objective 3). Moreover, an online-survey based on a semi-standardized questionnaire was conducted with 61 former participants of training of the trainers (objective 3). To study the effect on the institutional level (objective 4), structured group discussions with representatives from business associations and TESDA staff were held.

Limitations: The practical implementation of the survey was subject to a **number of constraints** that formed the data collection and evaluation process:

- Limited **time and budget** for the evaluation (e.g., a randomized selection of all 98 supported training institutions that were spread all over the country was not feasible, as it would have required much greater resources than those available).
- Difficult **traceability** of the graduates, particularly those of the comparison group. It was very challenging and hence required a manageable design.
- **Distortion effects** in the control group. One partner enterprise had an employment guarantee for graduates of the control group. Since this employment commitment influenced the employment situation of control group graduates (indicator under objective 1), these vocational schools were unsuitable as a control group and were subsequently excluded.
- Two institutions **refused to participate** and, since the Philippines is a country of emigrants, many respondents had to travel back to participate in the study. The necessity to travel could have influenced **the response behaviour** of graduates (e.g. employed graduates might be less likely to take the time to participate in the survey).
- Small **sample size** and **external validity**. The small sample size of graduates (197 out of 3.000 participants) limits the external validity of the study results and thus restricts the possibility of drawing general conclusions beyond the context in the Philippines.

Those challenges and constraints impeded a strictly rigorous IE survey design and accordingly led the evaluation team to make certain decisions that affected the data quality and particularly the external validity of the results (i.e., the representativeness of the survey). The evaluation was thus categorized a **"rigorized" approach**.

Focal and Cross-cutting Topics

- **Use of existing (M&E or administrative) data:** National statistics on vocational education and labor market data (i.e., from the Philippine National Statistics Office, the National Statistical Coordination Board, the Bureau of Labor and Employment Statistics, the National Wages and Productivity Commission and the International Labor Organization) were used for a preliminary framework analysis. For the examination of the joint indicators, monitoring data was of limited use since it was not suitable to detect intervention-induced effects on a macro-level. Monitoring data from the different projects was unavailable or incomplete. Firstly, the monitoring systems of the particular programs/projects followed their own logic and were not designed to measure overall achievements of the interventions. Secondly, the monitoring systems were not established before the implementation phase of the interventions or had changed meanwhile. The attempt to monitor program results has never been continued due to the inability of the TESDA to follow up with the data collection.
- **Gender-sensitive impact assessment:** Some of the interventions studied within this joint evaluation specifically focused on the improved employment of young women and the evaluation included “gender” as a cross-cutting topic. Specifically, one indicator under objective 1 (employability of graduates) addressed the gender dimension and found that female trainees benefit likewise from the support measures. In order to measure this, the treatment effect (on salary) was studied by gender of the graduates. The results show that the salary of female and male graduates did not differentiate except for the first salary. The interviews with representatives from training institutions and enterprises confirm the gender equity (for projects where this was intended).
- **Sustainability of impacts:** The sustainability dimension was addressed in terms of sustained project impact after the support has been completed. The evaluation team therefore assessed the sustainability of effects at the beneficiary, institutional and systematic level. The study concluded that positive effects on the beneficiary level will continue to last (e.g., better qualification) while sustainability on the institutional level and the systematic level is lacking. Firstly, after the interventions ended, the training institutions had difficulty maintaining the technical equipment they had received. Secondly, no diffusion effect in terms of other training institutions adopting the introduced training approaches was found. Thirdly, the majority of support focused on the manufacturing/industrial sector although the largest potential in the future lies in the financial and health service sector).
- **Private sector involvement:** Considering project implementation, the evaluation team concluded that the private sector (represented by associations for example) had only been involved in some of the evaluated programs but a systematic involvement was missing. Looking at the evaluation itself, the private sector was explicitly included through structured group discussions with representatives from industry and business associations. Specifically, the evaluation team aimed at identifying private sector acceptance of DTS as well as diffusion. The study concluded that future TVET projects should involve private sector stakeholders because the top-down approach focusing on the regulatory authority (like TESDA) was not effective.
- **Measurement of employability:** Concerning the employment situation of graduates, the evaluation team included several employability-related questions in the graduate field survey. They assessed whether the graduates were currently employed, how long it took them to find employment, and whether they found the practical tasks during training useful to find a job.

Key Evaluation Findings

The purpose of the IE was to rate the relevance, effectiveness, impact and sustainability of the ten project contributions. Overall, the evaluation team **rated the combined interventions as unsatisfactory**, since none of the overall objectives were fully achieved for the respective target group and the interventions were lacking sustainability.

Considering **objective 1** (related to the employment situation of the graduates of supported institutions), the comparison between the treatment and control group reveals **mixed findings**. While graduates of supported institutions were more satisfied with their salary and promotion opportunities, and their qualifications were rated higher by enterprises, their actual income did not differ from the comparison group. The unemployment rate between the two groups did not differ. Regarding **objective 2**, the interviews with the enterprises showed that they **appreciated the qualification** of the graduates from supported institutions. However, most interviewees could not explain whether or how the improved qualification of workers has led to increased productivity of their companies. Too many **disruptive factors**, such as insufficient resources for investments, ineffective workflows, and difficult market conditions seem to have offset potential positive effects of better-qualified staff. Moreover, only larger enterprises were able to comply with dual training requirements (e.g., minimum wage). Representatives from all interviewed enterprises were **partially not convinced that the benefits outweigh the investments**.

At an institutional level (related to **objective 3**), the empirical data confirmed that the management and training capacities of the training institutions had improved considerably due to the support measures. The training and consulting services, as well as the provision of equipment, contributed to the overall **improvement of the training institutions’ performance**. However, the evaluation also identified **significant problems with the use and maintenance** of the technical equipment provided, which reduced the interventions’ sustainability. Concerning **objective 4**, the

interventions **failed to increase TESDA's capacities**. The interview results indicated that TESDA did not increase its resources allocated to dual training and did not monitor the introduction of DTS.

CEval used the evaluation in the Philippines to **publish a book on TVET impacts**. The **evaluation has contributed to a shift** in the general German policy on TVET. The broad introduction of a DTS became more targeted towards specific regions or sectors. Avoiding oversaturation in a few productive sectors/job profiles remains challenging, but the study also points out potentials.

Conclusion on Methods Used

The evaluation was a **pragmatically "rigorized"** evaluation rather than a rigorous IE. The evaluation team had to **adapt the methodology to the given circumstances** and thus was unable to apply strictly rigorous methods. They chose a multi-method approach including both, qualitative and quantitative methods in order to overcome shortcomings that each method might have in the given context. The application of rigorous methods and quasi-experimental designs requires **specific conditions and assumptions**, which were not fulfilled in the interventions under the study (e.g. randomized participant selection, collection of baseline data before the intervention).

Impact measurement at the macro-level: While the "rigorized" evaluation revealed partial evidence of positive effects on both beneficiaries (e.g., job satisfaction of graduates) and training institutions (e.g., improved technical equipment), when comparing the DTS with other concepts, **no impact on the macro level was found**. For example, an increase in the average income of TVET graduates, an increase in national income or changes in unemployment rates of TVET graduate on the national level were not found. The evaluation team concluded that the interventions' effects might be too small to generate effects on the national level. Only qualitative interviews revealed the insufficient sustainability of the DTS in the Philippines, which also might explain the lack of impact at the national level.

Institutional networks: Empirical data confirmed that the perception of the multiplier role of an institution depends most notably on its connectedness with other institutions or a superior agency (like religious training institutions for example). When **selecting training institutions**, stronger emphasis should be placed on the institution's network and connections. If dissemination effects are intended, selected institutions should be required to prove their connection to other relevant institutions.

Policy and donor coordination: The study did not compare effects of DTS (supported by Germany) to those of competency-based trainings, which in the Philippines were mainly funded by the Asian Development Bank and Australia. However, the study detected that the **integration of the two approaches did not appear to be seamless**, as field interviews revealed compliance only with one approach or the other, but not both. Both approaches aim at addressing the mismatch between the growing numbers of school graduates with qualifications that do not fit the labor market demand. There seems to be an **evidence gap on how the two systems conflict or combine**.

Reference



Silvestrini, Stefan; Garcia, Melody. Joint Expost Evaluation 2010 – Dual Vocational Training, Philippines. Centrum für Evaluation, Saarbrücken 2010. https://www.kfw-entwicklungsbank.de/migration/Entwicklungsbank-Startseite/Development-Finance/Evaluation/Results-and-Publications/PDF-Dokumente-L-P/Phillipines_Dual_Vocational_Training_2010.pdf [accessed online on 07.09.2022]

Joint Ex-post evaluation 2010
– Main Report
Dual Vocational Training, Philippines



4.3 NON-EXPERIMENTAL DESIGNS

4.3.1 CASE STUDY 7: GLOBAL – SKILLS FOR REINTEGRATION

Project Description

Title	Skills for Reintegration
Commissioned by	BMZ
Implementing organization	GIZ
Implementing partners	Dominikus-Ringeisen-Werk (DRW), Arbeiter-Samariter-Bund (ASB), Gambia Technical Training Institute (GTTI), ElGroupConsulting LLC Predprinimatel, business training implementer, Ministry of Economy of Kyrgyzstan, Mexican Agency for International Development Cooperation (AMEXCID), International Organization for Migration, Mesoamerica Regional Program of the Secretariat of the Interior (SEGOB)
Evaluation institute	Mainlevel Consulting AG
Project area	Niger, Kyrgyzstan, The Gambia, Mexico
Target groups	Technical and vocational education management and training staff, staff of governmental institutions, staff of the non-governmental organization (NGO) network. For most activities, returning migrant workers and members of local communities make up the indirect target group.
Project term	12/2016 – 11/2020
Project cost	4,478,000 USD ²⁹
Evaluation term	08/2020 – 08/2021 (12 months). Reporting in 2021. Sample in 2018 (final evaluation)
Evaluation cost	Average total cost of a Central Project Evaluation is 75.000 USD ³⁰
Evaluation design (evaluation approach)	Non-experimental design (mixed-method, but mainly qualitative, theory-based approach using contribution analysis, the Kirkpatrick model and most significant change)
Publication date	05/2022

Project Context and Results

The **Skills for Reintegration project** supported migrants in their **decision to voluntarily return** to their home country and supported their reintegration process. The project's objective was to increase the employability of refugees, internally displaced people (IDP) and members of host communities through **needs-based qualification offers**. The project worked with targeted pilot measures in **The Gambia, Niger, Kyrgyzstan and Mexico** to increase the opportunities for people to gain personal and professional skills. The pilot project in Mexico focused on strengthening a network of NGOs, local, regional and international stakeholders and thus differed from the other pilots. In Kyrgyzstan, The Gambia and Niger, the pilot measures have contributed to newly acquired skills and increased employability while no such effect was found for the project in Mexico.

Evaluation Objectives and Indicators

The evaluation objective was to identify whether the pilot projects successfully **contributed to increased employability of voluntary returnees, IDP and members of host communities** through the provided needs-based training opportunities. Therefore, the pilot measures were assessed based on the OECD/DAC evaluation criteria and the evaluation criteria for German cooperation which was based on: relevance, efficiency, effectiveness, impact and sustainability. Coherence had not been introduced as a sixth criterion at the time of the evaluation.

To evaluate the **effectiveness of the pilot measures**, the following main evaluation questions were posed³¹:

1. To what extent have the agreed project objectives been achieved (measured by the indicators)?
2. How did the project contribute via activities, instruments and outputs to achieving the project objective?
3. Which unintended negative or positive results did the project produce?

²⁹ This is equivalent to about 4,500,000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

³⁰ This is equivalent to 75,000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

³¹ Please note that the following list is not comprehensive but meant to provide a brief expression of the main evaluation questions posed in CPEs with regard to the effectiveness criterion. A comprehensive list, structured according to the DAC criteria, can be found in the report's annex.

The assessment of the **pilot measures** was based on the three following outcome indicators for all countries except Mexico³²:

- Indicator 1: The number of participating national and international institutions providing support for refugees, IDP and host communities, which have implemented additional offers to increase the employability of their target groups.
- Indicator 2: The number of women and girls who participated in an offer to acquire personal and professional skills that are specifically geared to their needs.
- Indicator 3: The number of users (by gender) of the additional offers, who confirm that the qualification measures met their needs.

Evaluation Approach and Methods

This study is a typical **Central Project Evaluation (CPE)**, which is commissioned by the GIZ Corporate Unit Evaluation. The evaluation unit reports directly to the Management Board and is separate from the operating units. This organizational structure strengthens its independence. The evaluation design of all CPEs is based on the previously mentioned, now six-point criteria: relevance, coherence, efficiency, effectiveness, impact and sustainability. Moreover, CPEs typically include a **mixed-method approach** based on qualitative and quantitative data and prescribe a theory-based approach.

In order to answer the **first evaluation dimension** (i.e., comparison of the status and targets of the three outcome indicators), internal project monitoring data, survey data (i.e., tracer studies of 55 graduates) and qualitative data from interviews and discussions were used.

To address the **second evaluation dimension**, a **contribution analysis** was applied. The aim was to identify a plausible relationship between the project and the results achieved, based on methodological/data triangulation. For the evaluation at hand, **three country-specific hypotheses** were developed and tested:

1. Mexico: As a result of the initiation of an NGO network in Mexico working on communication, advocacy and knowledge sharing, the network's efforts to generate a new migration narrative in Mexico (e.g., through campaigning against xenophobia) have led to local communities being more sensitized on the topic of migration and (re)integration.
2. The Gambia & Kyrgyzstan: The needs-based offers on professional skills acquisition implemented in preparation for (re)integration have facilitated access to institutions and development actors.
3. The Gambia, Niger, Kyrgyzstan: The needs-based offers on professional skills acquisition implemented in preparation for (re)integration have led to newly learned skills and increased the employability of refugees, forcibly displaced people, returnees and local communities, resulting in better access to work and decent working conditions.

The contribution analysis included the following **key elements**: the results model including expectations on cause-and-effect relationships; the ToC including hypotheses that can be assessed in the evaluation, and a contribution story that documents changes; and the project's contribution as well as alternative explanatory approaches. To address the third hypothesis, the **Kirkpatrick training effectiveness model** was employed using level 1 (reaction), level 2 (learning), level 3 (behavior) and level 4 (results) through direct questioning of the training participants. In each country, around 20% of the participants of each training program were randomly selected and questioned according to the Kirkpatrick model. Wherever possible, the results were triangulated with the results of the tracer studies implemented by the project in each country.

The **most significant change (MSC) technique** was applied to identify unintended results as mentioned in **evaluation question three**. Therefore, one question on the MSC was included in the graduate survey. The findings were then validated in subsequent interviews.

Monitoring data and project documents: The evaluation team received monitoring data from the GIZ-results monitor (web-based monitoring system of GIZ) for all four countries. Identified risks to the project were not monitored regularly as part of the monitoring system. Other relevant project documents included the project proposal, national strategies, annual progress reports, BMZ country strategies and planning documents, as well as the previous results model and a map of actors.

Semi-structured interviews and FGDs: The evaluation extensively relied on semi-structured interviews and FGDs with project staff, implementing partners, the NGO steering committee in Mexico, ministries, and selected graduates of the trainings. During the inception phase, key institutional actors to be interviewed and key criteria for selecting interviewees within the target group, were identified. The interviews were conducted remotely (online, telephone). FGDs usually includes four to six people and lasts approximately two hours.

³² The pilot in Mexico differed from the other three countries because the focus was on strengthening the NGO network. Therefore, the pilot measure in Mexico was also considered as part of the assessment at the level of the project objective but not formally included in the indicators.

Impact measurement: In order to evaluate the project's impact based on higher-level development results, the CPE mainly focuses on relevant SDGs. Impacts were analyzed for each region. In the Gambia, Niger and Kyrgyzstan productive employment and decent work (SDG 8), and improved living conditions for refugees, IDP, returnees and local communities (SDG 1) was analyzed. In Mexico the social, economic and political inclusion (SDG 10) for the same target groups were analyzed. Therefore, the evaluation team illustrated national trends in unemployment rates. However, lacking data on specific unemployment rates for the target group (e.g. refugees, IDP) were not fully available. A contribution analysis including three hypotheses was conducted for the impact criterion based on the data sources mentioned above.

Limitations: The evaluation team faced several challenges which hampered the evaluation. Firstly, it was difficult to trace the former participants, which reduced the studied sample. Therefore, the MSC method was not fully applied. Secondly, the evaluation team tried to identify a control group that could be compared to the training graduates. However, they could only find a very limited number of suitable cases and thus could not implement a control group setting. Thirdly, at the time of the evaluation, employment effects were not measurable, because the COVID-19 pandemic had postponed the start of apprenticeships/employment. Thus, the evaluation team had to use qualifications and relevance of qualifications as proxy for employability (i.e. assuming that acquired skills will lead to higher employability in the future).

Focal and Cross-cutting Topics

- **Use of existing (M&E or administrative) data:** The evaluation team received monitoring data from the GIZ-results monitor for all four countries. The internal monitoring data was updated and helped the evaluation team to identify whether target values of the objective indicators had been met. Moreover, the evaluation team relied on national statistics on unemployment. However, this data was only partially suitable to address the evaluation questions, as they were not disaggregated to suit the target group.
- **COVID-19 pandemic:** The COVID-19 pandemic affected the evaluation procedures and required fieldwork to be conducted remotely or semi-remotely. The international evaluators collected data virtually while local evaluators collected data semi-remotely, with a few interviews being held face to face in The Gambia and Niger. In the semi-remote evaluation design, local evaluators carry a higher level of responsibility. The international and local consultants constantly reflected on findings gained and shared learning experiences.
- **Measurement of employability:** When the evaluation took place, project effects on employability had not materialized due to postponement related to the pandemic. Therefore, qualifications and relevance of qualifications were used as a proxy for employability.
- **Gender-sensitive impact assessment:** Two out of the three indicators intentionally include a gender aspect to distinguish between male and female beneficiaries. However, the gender aspect was not specifically addressed in the evaluation: neither in the tracer studies nor in the survey including the Kirkpatrick Model.
- **Follow-the-money-approach:** In order to address the efficiency criterion, CPEs usually rely on a follow-the-money approach, which combines information on project costs and project results. The approach is split in two parts: Production efficiency, which compares allocated resources and outputs; and allocation efficiency, which compares allocated resources and outcomes. GIZ's Corporate Unit Evaluation has developed an excel tool to standardize this efficiency analysis, which allows allocated resources (financial and human) to be linked to certain outputs (production efficiency) and outcomes (allocation efficiency). This analysis is complemented with evidence from interviews/discussions with project staff.

Key Evaluation Findings

Overall, the global project Skills for Reintegration was categorized as **moderately successful**. The evaluation team concluded that all three project objective indicators were fully achieved by the end of the project. The evaluation team concluded that the third hypothesis was confirmed, as the **pilot projects have contributed to newly learned skills and increased employability** of the trainees attending the capacity-building activities and, according to the data analyzed and the interviews, better access to work and working conditions. The pilot measure in Mexico was formally not included in the indicators but still considered. Due to the issues with the lack of exchange between NGOs and the government, the **pilot project in Mexico cannot be regarded as fully successful**. In general, the target values of the indicators were very low, making them easy to achieve.

The evaluation team identified the following **challenges** to project implementation:

- Lacking clarity on project objectives and target group
- Changes in the project planning, such as the choice of partner countries, target groups and pilot measure approaches
- Limited adaptation of the project design to the implementation reality
- COVID-19 pandemic and the security situation in Niger hampered the project implementation

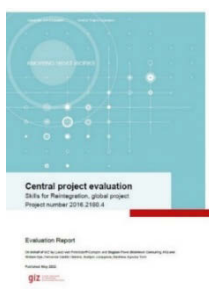
Conclusion on Methods Used

GIZ has **standardized the design of CPEs**, which specifies a theory-based approach, the use of mixed-methods for data collection, a contribution analysis and a follow-the-money approach. GIZ considers the defined approach as robust. The aim is to **identify a plausible theoretical relationship** between the project and achieved results and to gather sufficient evidence (methodological/data triangulation) that the results are **more likely attributed to the project**.

The following conclusions were drawn by the CPE-evaluation team:

1. To ensure comparability and broader use of evaluations, the CPE-system requires that all reporting procedures be followed, all chapters are filled, and a strict page limit is fulfilled. In the view of the evaluators this led to an **inflexible evaluation process** that did not meet the project-specific needs in this case. As a result, the project was dissatisfied with the evaluation process, and the recommendations were not considered helpful in drawing practical conclusions, as there was also no follow-on project.
2. A useful tool is the **pre-defined evaluation matrix**, which covers all aspects of the DAC criteria and the sub-criteria. However, the **rating is done rather subjectively**.
3. The evaluators added the **Kirkpatrick Model** to assess training results. The model helped to **estimate the impact on employment**, which had not been visible at this evaluation stage. The team worked pragmatically with an adapted questionnaire including **open and closed questions**.
4. The **use of evaluation results is not clear** to the evaluation team. As there is no direct follow-up project, the **recommendations might remain unused** in this case.

Reference



Petersdorff-Campen, Lukas von; Pavel, Bogdan; Njie, William; Valdés Herrera, Fernanda; Jussupova, Arailym; Ayoub Tinni, Bachirou (2022): Skills for Reintegration, Global Project; Evaluation Report. Central Project Evaluation 2016.2180.4. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, <https://mia.giz.de/qmlink/ID=249722000> [accessed online on 07.09.2022]

4.3.2 CASE STUDY 8: EGYPT – EMPLOYMENT PROMOTION

Project Description	
Title	Employment Promotion in Egypt (EPP) and Enhancement of the Egyptian Dual System (EEDS)
Commissioned by	BMZ
Implementing organization	GIZ
Implementing partners	Ministry of Education and Technical Education (MoETE)
Evaluation institute	Madiba Consult GmbH
Project area	Egypt
Target groups	Political entities, TVET authorities, the private sector, direct beneficiaries (students, graduates, unemployed youth, returnees)
Project term	EPP: 01/2016 – 06/2020; EEDS: 12/2015 – 06/2020
Project cost	EPP: 14.4 million USD; EEDS: 14.8 million USD ³³
Evaluation term	February – November 2020
Evaluation cost	Average total cost of a Central Project Evaluation is 75.000 USD ³⁴
Evaluation design (evaluation approach)	Non-experimental design (qualitative theory-based approach, the CPE used contribution analysis, most significant change and follow-the-money approach)
Publication date	March 2021

Project Context and Results

The two **German-Egyptian projects—Employment Promotion in Egypt (EPP) and Enhancement of the Egyptian Dual System (EEDS)**—were evaluated simultaneously. Both projects were modules of the Sustainable Economic Development for Employment program in Egypt. Due to the close interlinkages and because EPP and EEDS will be merged into EPP III, the projects were evaluated together.

EPP and EEDS both respond to the labor market needs by contributing to capacity development of the Ministry of Education, TVET institutions and youth (students and graduates) in Egypt. The objectives of the two projects evaluated were to **better qualify TVET students and unemployed youth** for the demands of the labour market (EEP objective) and to **increase the number of students enrolled in dual education** of adequate quality (EEDS objective). Overall, both projects were rated successful in achieving their objectives. However, the gender-related targets were not fully achieved.

Evaluation Objectives and Indicators

The evaluation assessed the two projects in accordance with the OECD/DAC evaluation criteria and the guidelines of GIZ's CPEs. Therefore, the relevance, efficiency, effectiveness, impact and sustainability³⁵ was evaluated. One of the evaluation objective was to identify whether the two projects successfully achieved their objectives, which is in line with the effectiveness criteria and described in more detail in this case study.

To evaluate the **project measures' effectiveness**, the following main evaluation questions were posed³⁶

1. To what extent have the agreed project objectives been achieved (measured by the indicators)?
2. How does the project contribute via activities, instruments and outputs to achieving the project objective?
3. Which unintended negative or positive results does the project produce?

The assessment of the two projects was based on the **following outcome indicators**:

EEP	EEDS
1 Six strategic policy recommendations for action in TVET and labor-market policy targeting youth, (originating from the employment dialogue), have been submitted to the political decision-makers for strategic adoption	1 At nine locations, the number of young men and women in promoted dual education courses is to increase by 30% overall. 2 The percentage of women in vocational education in the promoted courses is to increase by 20%.

³³ This is equivalent to the volumes 14.5 million EUR for EEP and 14.9 million EUR for EEDS in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

³⁴ This is equivalent to 75,000 EUR in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

³⁵ Coherence had not been introduced as a sixth criterion at the time of the evaluation.

³⁶ Please note that the following list is not comprehensive but meant to provide a brief expression of the main evaluation questions posed in CPEs with regard to the effectiveness criterion. A comprehensive list, structured according to the DAC criteria, can be found in the report's annex.

EEP	EEDS
<p>2 75% of the benefiting 47,000 young people (20,000 of whom are women) assess the Active Labor Market Policies (e.g. vocational orientation, short-term training, job placement services, skills competitions) as beneficial (rating 1 or 2 on a scale of 1 to 6) for improving their individual employment prospects.</p> <p>3 52% of the youth from 200 supported institutions have found employment within 6 months after graduation (30% of whom are women).</p>	<p>3 75% of the 100 training companies surveyed confirm that the promoted dual education courses meet their needs in content and organization.</p> <p>4 69% of the 4,500 graduates of the 6 promoted locations are in dependent employment or self-employed 6 months after graduation.</p>

Evaluation Approach and Methods

The projects were assessed using the **standardized CPE approach** developed by GIZ. As mentioned in Case Study 7, the CPEs rely on a theory-based evaluation approach, which includes: a **study of the project indicators** (i.e., quality check and achievement of target values); a **contribution analysis** guided by five selected hypotheses per project (three hypotheses at output to outcome level and two hypotheses from outcome to impact level); **MSC methods**; as well as the **follow-the-money approach**. The CEP is guided by a **predefined evaluation matrix**.

The evaluation team used qualitative data as well as results-based monitoring data and tracer study data from the two projects.

Monitoring data and project-related documents: the evaluation team reviewed relevant project documentation, final reports, survey reports, and existing quantitative and qualitative monitoring data. These sources were used for data triangulation (i.e., compare findings from interviews and FGD).

Interviews and FGD: The evaluators used open and semi-structured interviews, FGD and workshops with the following people and groups: the project team; the implementing partners (e.g., relevant ministries and training centres; universities; civil society; and direct beneficiaries (e.g., training graduates, beneficiaries from career guidance) to collect qualitative data. Qualitative content analysis was used for the analyses of the interviews and FGD. Most interviews and focal group discussions were held in Arabic by the national evaluation experts. Therefore, additional measures had to be taken by the evaluation team to set up a mutually comprehensive overview of the evaluation stage and results.

Quantitative data: Tracer studies had already been conducted by the projects and were available to the evaluation team. Graduates were asked to rate their skills on a scale from 1 to 5. Standardized interviews and the development and distribution of questionnaires or even online surveys were not considered for this evaluation purpose. In correspondence with the project teams, these options were rejected, when taking into consideration the difficulties to get respective permissions from the Egyptian Government.

Focal and Cross-cutting Topics

- **Use of existing (M&E or administrative) data:** The provision of project-related monitoring data is common for CPEs since the achievement of target values of monitored indicators is one central aspect of the evaluation.
- **Follow-the-money approach:** Today, GIZ uses KOMP (Cost-output monitoring and prognosis) to assign costs to outputs during the project implementation. However, KOMP was not systematically used for the monitoring of costs per output in the projects starting before mid-2019. Therefore, the evaluation team reconstructed together with the project teams retrospectively a rough estimation of the percentage of costs allocated to the respective outputs to enable the evaluators to complete the efficiency tool of CPE as a data basis for the efficiency analysis. Human resources could be allocated more precisely to the outputs.
- **Gender-sensitive impact assessment:** Three out of the seven project indicators specifically address female participation. Thus, gender-specific data were gathered by the internal monitoring systems of the project and then provided to the evaluation team in order to analyze the projects' effectiveness. Besides the project indicators, the gender dimension was not specifically addressed in the evaluation.
- **Implications of the COVID-19 pandemic:** The selection of stakeholders to be interviewed and project sites to be visited had to be reorganized for the remote evaluation. During the remote evaluation mission, the international experts conducted all English interviews and FGD, and the national experts all Arabic interviews and FGD. Over the course of the remote evaluation, onsite observations could not complement the findings of interviews, FGD and document analyses. The evaluation team assessed that the lack of observation possibilities compromised the quality of the evaluation to a certain extent: non-verbal communication, environmental settings and atmosphere could have provided important additional information.

Key Evaluation Findings

Overall, **EPP was rated as successful (level 2 out of 6 levels) based on the five DAC criteria**. Concerning effectiveness, EPP achieved its planned target values for indicators 1 and 2. However, EPP was unable to comprehensively fulfil outcome indicator 3, since only 45% of the youth in the supported institutions found employment after their graduation, instead of 52% as set in the indicator. With regard to women, the achievement is even less with only 19% were women, instead of the envisaged 30%. Target groups as well as beneficiaries reported an increased self-awareness of their strengths and their capabilities, which increased their self-confidence and ambitions. Interviews with GIZ suggested that after completion of their training, many graduates found jobs in the informal sector because of the possibility to earn slightly more money there.

Regarding **EEDS, the project was also rated successful based on the five DAC criteria**. Concerning effectiveness, EEDS achieved its planned outcomes according to outcome indicators 1 and 4. However, there were some reservations about outcome indicator 3 because the necessary survey to prove the fulfilment of the indicator could not be completed due to security measures related to COVID-19. Nevertheless, other assessments and surveys completed by EEDS during the implementation period showed that training enterprises were satisfied with the contents. The only indicator EEDS could not fulfil was outcome indicator 2 (increase the percentage of female students in promoted courses by 20%). EEDS only achieved an increase from 22% to 27% (equivalent to an achievement of 38%). EEDS did not pursue expansion into female-dominated occupations (such as the textile industry) due to the very poor working conditions. A positive unintended result was the involvement of the private sector in the form of public-private partnerships with Siemens through the Siemens Academy as well as the integrated development partnership with the private sector.

Conclusion on Methods Used

The **standardized CPEs allow the evaluation teams to rely extensively on existing monitoring data** from the projects, but the evaluators may judge the objective achievement differently, if arguments are presented. This kind of data is usually triangulated with additional qualitative and quantitative data, mainly in the form of interviews and FGD with relevant stakeholders and beneficiaries. In this case, the evaluators were also able to use existing data from tracer studies.

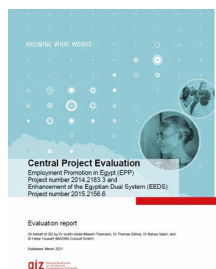
The following conclusions were drawn by the CPE-evaluation team regarding rigorous result/impact evaluations:

Baseline and tracer study data should be generated along with the project implementation to measure the results in a more rigorous way. This data availability, as well as a clear documentation on how participants were selected, could enable a research design which compares the treatment group to a control group. In addition, **qualitative data and analyses are important for the German cooperation** to understand the intervention context, to identify the contribution of the project and to analyze alternative explanations.

This evaluation was **one of the first CPEs that had to be carried out remotely** because the pandemic situation and thus had to conduct remote interviews and FGD. The case is quite similar to the CPE in Case Study 7. However, in Egypt **two ongoing projects were evaluated simultaneously** in order to draw lessons learnt for a joint follow-up measure. Results on effectiveness and impact were analyzed separately by project and therefore, the report became rather long and **unsuitable to provide practical recommendations**. After this CPE, GIZ has started to use separate reports for each project, even of those evaluated together, to make the evaluation reports more readable. Moreover, the follow-on project had already been developed and the **evaluation findings were not used in the design of the follow-up project**. GIZ uses these final evaluations for the implementation and steering of the follow-on project.

The evaluation report was published, but to the knowledge of the evaluation team in this case **not officially shared with the partner institutions**. However, it is part of the standard CPE process that projects/country offices get the final reports and are asked to share these with the partners. Thus, the learning effect from this evaluation report may be low for the partner country. Ideally, partners should be part of the complete standard evaluation process, which strengthens the evaluation approach and enables learning throughout the whole process. Generally, GIZ publishes next to the main evaluation report, summary reports and one-pagers of CPEs that are meant to communicate evaluation results to the **policy level and the public**.

Reference



Abdel-Massih-Thiemann, Judith; Döhne, Thomas; Saleh, Bahaa (2021): Enhancement of the Egyptian Dual System (EEDS). Central project evaluation 2015.2156.6. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, giz2021-0094en-projectevaluation-employment-promotion-egypt.pdf [accessed online on 07.09.2022].

4.3.3 CASE STUDY 9: GLOBAL – SUCCESS FACTORS FOR GENDER EQUALITY IN VOCATIONAL EDUCATION AND TRAINING

Project Description

Title	The Future is Equal: Success Factors for Gender Equality in Vocational Education and Training
Commissioned by	BMZ
Implementing organization	GIZ
Implementing partners	GIZ, KfW, German NGOs (incl. PLAN International)
Evaluation institute	Eva Dietz (independent consultant), Anna Emil (Mainlevel Consulting AG), Svenja Müller (GIZ)
Project area	Global scope
Target groups	Survey among project teams
Project term	Unknown
Project cost	Unknown
Evaluation term	February – July 2021
Evaluation cost	Confidential information (about 40 work days in total)
Evaluation design (evaluation approach)	Non-experimental design (qualitative meta-study approach to identify best practices using interviews and existing project documents)
Publication date	2021

Project Context and Results

This study researched **250 TVET projects** implemented under German and international DC. The study focused on projects, which clearly pursued a gender perspective in vocational education, had at least a gender equality (GG) marker and/or addressed specific **gender-relevant aspects** and produced readily accessible information to prove it. Around 30 selected project examples were examined in detail. The study examined projects implemented by the German development organizations GIZ and KfW as well as measures implemented by non-governmental actors, including NGOs and faith-based agencies, as well as international and regional networks.

Evaluation Objectives and Indicators

The analysis and its conclusions were based on the following **key questions**:

- How can DC better identify obstacles and potentials for gender equality in TVET?
- What are the major challenges and greatest potentials in this sector?
- Which success factors help to overcome the obstacles and leverage potentials?
- What recommendations can be drawn from this?

Around **30 selected project examples** were examined in detail as part of this study.

The central categories of analysis were:

- The three-fold approach of BMZ: gender mainstreaming, women's empowerment, and policy dialogue
- The five elements of the GIZ Gender Strategy 2019: political will and accountability, corporate culture, gender competence, process adjustment, and gender equality within the company
- Internal and external anchorage of methodology for mainstreaming gender equality

Evaluation Approach and Methods

The study used a **qualitative social science approach to identify key success factors** in TVET projects that focus on gender. The project selection and analysis were based on the following steps:

1. Definition of the **research question**.
2. The BMZ categories (1 to 3) for the gender sensitivity of the projects were used as the **first grid for project selection** and supplemented by online research.
3. A **key variable in selecting projects** was also whether project information was easily available and whether contact persons were willing to provide information.
4. **The projects were then categorized** by phases (entry phase, training phase and staying in the job), by intervention level (macro, meso, micro levels), by geographical region and other topics/sectors (such as new technologies, agriculture, and health).
5. To **reduce the number of selected projects**, a 2-page interview sheet was sent to selected projects.
6. **In-depth interviews** (approx. 30 projects) were conducted, evaluated and weighted. The project teams were asked about the success factors for overcoming hurdles.

7. A **feedback loop** with projects (digital format) was integrated in order to receive their agreement on the findings.

Evaluation: The abovementioned key questions and categories were used to map the analytical framework³⁷ for evaluating selected project examples. Evaluation itself consisted of **examining various project documents**. This information was triangulated in the 30 selected projects through **interviews** with project managers, gender focal points, desk officers and/or other project staff. The structured interviews used the criteria set out in the analytical framework. The evaluation of the project documents and interviews subsequently provided the basis for identifying the **success factors**. Large parts of the results were not published but used for GIZ's **internal programming**.

Limitations: No target group interviews or FGD were conducted. The inclusion of the target group (i.e., women that benefited from the evaluated programs) would have provided deeper insights into the success factors. Interviews were only conducted with project staff and stakeholders involved in the implementation.

Focal and Cross-cutting Topics

- **Use of existing (M&E or administrative) data:** Although project-related documents were used for the study, M&E or administrative data were not considered.
- **Gender-sensitive impact assessment:** The whole study only included TVET projects that covered gender-relevant aspects.
- **Implications of the COVID-19 pandemic:** The entire evaluation had to be conducted remotely including virtual interviews with the stakeholders. A on-site study would likely have provided deeper insights.
- **Measurement of employability:** Beneficiaries/graduates were not considered in this study and thus their employability was not investigated.

Key Evaluation Findings

Based on the document analysis and the interviews conducted for the 30 selected projects, the study identified success factors that enhance gender equality in TVET. The key outcome of this study consists of the following **15 success factors for gender equality in TVET**: four key success factors and an additional 11 factors, which are graphically showcased through project examples.

Amongst other things, the study puts forward four key success factors: increase awareness of human and women's rights and overcome discriminatory gender stereotypes; promote gender-sensitive training materials and women teacher; support gender-responsive infrastructure (e.g., sanitary facilities, child care); and minimize violence-related risks for women (e.g., safe means of transportation to training centers). These first four success factors are regarded as pivotal for overcoming discriminatory conditions. They are presented on their own, without any concrete project examples.

With its many **recommendations for practical action**, the study fuels support for equality. It addresses TVET specialists, especially implementers and decision-makers engaged in DC. The study concludes that gender equality must become an even stronger quality attribute and eligibility criterion in DC. The study sets out recommendations for TVET in German DC, targeting both the **policy level and the design of development projects** and programs. The recommendations for action were derived from the success factor analysis. In addition, the study puts forward concrete proposals for implementing these success factors.

Conclusion on Methods Used

The study did **not apply a rigorous evaluation approach**. The goal of the evaluation was to identify **success factors** of the projects based on document analysis and interviews. The study did not aim at identifying the project's impact or effectiveness. Therefore, the study focused on the **usability and usefulness** to address specific **project needs and formulate practical recommendations**.

Research on site and with the involvement of participants/beneficiaries is vital and can hardly be replaced by remote methods. The evaluation period was rather short and it was **conducted remotely**. For deeper insights, the interviews would have to be done in presence, including FGD.

Reference



Dietz, Eva; Emil, Anna; Müller, Svenja (2021): The future is equal: Success factors for gender equality in vocational education and training. Bonn: GIZ, [giz2021-0297en-future-is-equal.pdf](#) (English Version) and [giz2021-0228de-future-is-equal-gleichberechtigung-berufliche-bildung.pdf](#) (German Version) [accessed online on 07.09.2022]

³⁷ An analytical framework was developed to guide the structured interviews and to evaluate the selected project examples. The framework is based on the key questions and the central analysis categories. Since this document was not available for the case study, it could not be analyzed in detail.

CHAPTER 5 REFLECTION QUESTIONS PRIOR TO IMPACT EVALUATIONS

The following table provides six clear steps with multiple reflection questions, which should be analyzed prior to any (rigorous) impact evaluation (IE). Table 3 is intended as a toolbox to provide practitioners with guidance. The table briefly explains the reasons why these questions are important for the success of results and IEs. It is based on the evaluation guidelines by [ADA \(2020\)](#) and was adjusted using the findings from the literature review and from the analysis of the case studies and interviews conducted by the present study.

Table 3: Six steps with reflection questions prior to conducting an impact evaluation

STEPS	REFLECTION QUESTION	REASON FOR THE IMPORTANCE OF THIS STEP AND THE RESPECTIVE QUESTION
1. Frame the evaluation interest and use	<p>Should this project/program be evaluated? Would an evaluation be useful?</p> <p>Pro-arguments:</p> <ul style="list-style-type: none"> • Specific knowledge interest • Pilot projects/programs or innovative approaches with potential for replication or scaling-up • Project/program is considered for a subsequent phase • Project/ program addresses evidence gaps and has potential for learning • Project/program is of strategic importance • The findings are expected to be useful and used <p>Contra-arguments:</p> <ul style="list-style-type: none"> • Small projects • Only small contribution by the respective donor and other actors conduct a similar evaluation already (duplication) 	<p>Not every project/program should be evaluated, especially in case of insufficient evaluation interest and if the evaluation findings, conclusions and recommendations are not expected to be used.</p> <p>Step 2 provides further insights for the assessment of the feasibility of an evaluation.</p>
2. Balance scope, budget and time available (contextual factors)	<p>Scope: What do we want to know (with what kind of certainty)? What are the results/impacts we try to measure? What is inside and outside the scope of the IE?</p> <ul style="list-style-type: none"> • Geographic aspects/sample size: Which countries, regions, areas, districts, target communities should be part of the IE and which should be excluded? How large is the sample size? <p>Timing: When do I plan to conduct the IE along the project/program life cycle? When would the findings, conclusions and recommendations be available?</p> <ul style="list-style-type: none"> • Thematic/structural aspects: What specific period, current project/program cycle, multiple cycles should be considered? Is ethical randomization possible ex-ante the project implementation? Does the project ask for adaption of the intervention (e.g., strategy) during project implementation (as far as predictable)? <p>Resources: How many resources (budget and know how) may be invested in a (rigorous) IE?</p> <ul style="list-style-type: none"> • Evaluability aspects: Is there enough data available and are key informants 	<p>There has to be the right balance to ensure high quality IEs:</p> <ul style="list-style-type: none"> • Scope: The scope must be realistic in terms of time and budget. The scope determines the amount of budget needed. • Timing: The timing of the IE is key to gain purposeful insights. The timelines of findings is crucial for the uptake and use of the IE findings. Furthermore, experimental designs are usually not applied to ex-post settings. DC often asks for ex-post IEs, so that accompanying and ex-post IEs using quasi-experimental and non-experimental designs are more frequently used. Ideally, the preparation for an IE should take place as incremental part of the project planning; this enables the potential use of all IE designs, including the so called “gold standard” of rigorous IE and the use of RCTs for causal attribution. However, other preconditions are necessary to conduct RCTs, e.g., large samples, ethical randomization and non-adaption of the treatment during implementation. • Resources (budget/knowhow): What resources are available and required? The IE budget must realistically reflect the workload needed. Adequate IE budget depends on the purpose, objective, scope, design and approach for the IE (see Step 3-5). Experimental and quasi-experimental designs typically ask for higher budgets and a different set of methodological (statistical) know-how than non-experimental designs using qualitative IE approaches. Experimental IEs (RCTs) may easily cost 6-7-digit figures in USD and quasi-experimental designs low 6-digit figures, while qualitative designs are for 5-digit figures³⁸ or even lower costs. However, the costs (budget) highly depend on the

³⁸ ADA calculates with about 25,000 EUR-90,000 USD for evaluations depending on the size of the project/program and they typically earmark at least 3 percent of the respective program or project budgets for evaluations (ADA 2020, EC 2022).

STEPS	REFLECTION QUESTION	REASON FOR THE IMPORTANCE OF THIS STEP AND THE RESPECTIVE QUESTION
	accessible to enable solid data collection and evidence generation?	context (e.g., Case Study 2 shows that an experimental design, RCT, which was possible to be conducted with limited resources in the Indian context). The main costs of IEs are the personnel costs, especially for the data collection and data analysis, which can be reduced if existing data of sufficient quality can be used, or researchers are funded by separate public funds (e.g., by universities). Experimental and quasi-experimental designs usually ask for high personnel costs due to new, extensive and high-quality data collection of large samples.
3. Detail purpose and objectives	<p>What is the main <u>purpose</u> of the results/impact evaluation (for practical use)? Why is the IE undertaken and for whom? Why is the IE undertaken now? Who is asking for it? What are the intended benefits of the IE for whom? Who is going to use the findings and recommendations?</p> <ul style="list-style-type: none"> • Learning: Do we want to know why particular development interventions have worked or not? • Steering: Do we want to supply credible and reliable findings for evidence-based decision-making at strategic and operational levels? • Accountability and communication: Do we want to give account of the use of public funds and results/impacts achieved to partners, donors or the larger public? <p>What is the main <u>objective</u> of the results/impact evaluation? What does the IE seek to accomplish and how should the results be used to benefit the project/program/intervention/organization at large?</p>	<ul style="list-style-type: none"> • Clarify the main purpose of the IE, which can be learning, steering or accountability and communication. If this is unclear the IE might have a wrong focus. • The objectives logically follow the purpose and provide more details on what the IE seeks to accomplish. • Specify the intended users of the IE.
4. Specify the main evaluation questions	<p>What is/are my <u>RQ/RQs</u>?</p> <p>Which OECD/DAC criteria should be assessed (e.g., effectiveness, impact)?</p> <ul style="list-style-type: none"> • Apply the OECD/DAC criteria thoughtfully and selectively as a guiding framework for developing IE questions. These are the two most relevant OECD/DAC criteria for measuring the results and impacts of an intervention: • Effectiveness: Is the intervention achieving its objectives? • Impact: What difference does the intervention make (in the long-term)? 	<ul style="list-style-type: none"> • In this step, the purpose, objective and scope of the results/impact evaluation are translated into specific and clear RQs. • The research/evaluation questions are particularly important because they drive the entire evaluation (incl. the evaluation design, methodological approach and methods for data collection and analysis, as well as the evaluation budget required, data needs and the timing of the evaluation).

STEPS	REFLECTION QUESTION	REASON FOR THE IMPORTANCE OF THIS STEP AND THE RESPECTIVE QUESTION
5. Outline the evaluation design and approach	<p>What IE design does this question ask for? Am I interested in (rigorous) IEs? How do I define rigorous IE?</p> <p>Does the question ask for quantitative or qualitative approaches and methods or both?</p>	<p>Define the IE design and approach through which the re-search/evaluation questions will be answered (<i>see Chapter 2</i>). To identify the best possible IE design and approach, it is necessary to balance what is best³⁹ (in terms of accuracy) and what is feasible in DC practice. The purpose, objectives, context, available resources and timeframe have to be considered.</p> <p>IE design⁴⁰:</p> <ul style="list-style-type: none"> • Experimental design • Quasi-experimental design • Non-experimental design <p>IE approach⁴¹:</p> <ul style="list-style-type: none"> • Quantitative approaches and methods⁴² • Qualitative approaches and methods⁴³ • A mixed method approach is recommendable to increase the variety of information and insights. Triangulating data, sources and methods is important to promote credibility and use of evaluation results.
6. Consider taking a participatory approach for the evaluation process	<p>What kind of options for participation exist (for multiple stakeholders, like project planners, project implementers, evaluation departments, partners and the target group) during the results/impact evaluation process? How can we ensure various stakeholders have a voice in the evaluation process? How can partners and the target groups be involved in the IE and in what ways and formats can results be shared with them?</p>	<p>Effective options for participation in the IE process:</p> <ul style="list-style-type: none"> • Inception/ kick-off workshop can be used to identify the information needs and interests of project implementers and relevant stakeholders to increase the utility of the evaluation and understand each other's perspectives and build common ground for the IE, update the ToC and clarify communication processes) (<i>see Chapter 6.2</i>) • Participatory evaluation approach, like outcome mapping, MSC or MAPP (<i>see Chapter 2.3</i>): • Validation or debriefing workshop can be used to communicate the findings, consult and reflect on findings, conclusions and recommendations in a participatory manner (<i>see Chapter 6.2</i>) • Commenting and quality check of IE reports by project implementers and evaluation units before finalization by the evaluator (<i>see Chapter 6.2</i>) • (Senior) Management responses and regular monitoring could be foreseen from the beginning of an IE as an effective tool for the increased utilization of IE results (<i>see Chapter 6.3</i>)

³⁹ Robust methods and data collection tools and triangulation are preconditions for obtaining solid and reliable data, which is the basis for credible and useful findings of results and impacts.

⁴⁰ The evaluation design is the overall strategy chosen for assessing, analyzing and estimating the causal results and impacts (change).

⁴¹ The evaluation approach is methodological approach, incl. the selection of methods for the data collection and analysis.

⁴² Quantitative approaches and methods measure and assess what can be studied with numbers. These answer the 'what' questions and use structured approaches that provide precise data that can be statistically analyzed.

⁴³ Qualitative approaches and methods analyze and explain what can be studied with words. These use semi-structured techniques to provide data than can provide an in-depth understanding of attitudes, perceptions and behaviors.

CHAPTER 6 STRATEGIES ON HOW TO USE THE FINDINGS FROM IMPACT EVALUATIONS IN THE SKILLS DEVELOPMENT SECTOR

One of the **main purposes of IEs is learning**, as stated by the evaluation policies of all DACH DC actors (*see Chapter 3*). Measuring results and impacts of an intervention enables it to prove its ToC and to **identify what worked or what did not and why and learn from that**. The findings and lessons learned can then be used to improve the design, strategy, and/or implementation of future project interventions and, ultimately, enhance DC results. IE findings can also be used to drive transformation in a particular sector by fostering political will and support for innovative approaches (e.g., cooperative TVET), based on evidence and lessons learnt. Furthermore, IEs may also be used for academic and scientific research purposes, for example, when the analysis addresses the need to validate causality hypotheses or a ToC in order to infer conclusions that confirm assumptions or refute common fallacies. These type of studies and evaluations can be then used by the conception of DC interventions—for example, to provide evidence supporting the application or scaling-up of particular intervention approaches at the regional, national or international level.

During the interviews conducted for the present study, it was identified that **researchers and development practitioners have fundamentally different understandings of IE or “learning about what works and why”**. On the one hand, a primary goal of researchers is to publish their research results in peer-reviewed journals with strict minimal requirements for scientific rigor, resulting in a tendency to prefer experimental and quasi-experimental designs. After publication, they rarely know who reads or uses their research findings for DC practice. On the other hand, the main interest of DC practitioners (project planners, implementers, or policy makers), is to foster learning processes with stakeholders at different levels and to translate evaluation insights into implementable recommendations and entry points for action. Compared to researchers, practitioners have fewer (and less stringent) requirements in terms of evaluation methodology. DACH DC actors tend to use non-experimental designs and qualitative approaches most frequently in practice, also for the evaluation of TVET interventions. Although the evaluation methodology is crucial for the significance and validity of the recommendations, practitioners rarely question it. **“Many practitioners tend to consider a weak and a very rigorous evaluation design as equally informative”** (derived from the interview about Case Study 6 and based on long-term evaluation experience in German DC). This phenomenon may be due to multiple reasons. Practitioners face various hurdles, such as: limited access to the confidential data collection documentation on which the evaluation results are based; non-existent or very broad description of the methodology in the evaluation reports; lack of time for a thorough review of the evaluation methodology; or because practitioners (being experts in their areas of work), do not necessarily have sufficient knowledge of evaluation methodologies, particularly of quantitative ones. Therefore, they contract experienced evaluators and experts and trust their methodological capabilities.

To discuss strategies related to the use of IE findings, the **different needs of various groups of DC stakeholders** have to be addressed separately. Focus will be placed on the following three stakeholder groups: project planners (*see Chapter 6.1*), project implementers (*see Chapter 6.2*) and policymakers (*see Chapter 6.3*).

6.1 USING META-EVALUATIONS FOR PROJECT PLANNING

When designing a DC intervention, project planners need to be well informed about the context and situation of both the given sector and the target groups. Besides other documents, project planners consult the progress reports of a project/program for the planning of a next project phase or a new intervention in the same intervention area. In addition, IE reports of similar DC interventions will also be considered as a source of tested hypotheses to build the ToC of the new intervention. This is not new in the field of project planning, neither in the planning of TVET interventions. However, in recent years, commissioning parties (e.g., BMZ) and implementation organizations (e.g., GIZ) are increasingly requiring project planners to provide evidence to support the methodological approach of a planned DC intervention. Therefore, studies and evaluation reports of projects/programs with a similar objective are the main source of **evidence for the causality hypotheses reflected in the ToC of a planned intervention**.

However, it is **difficult to generalize the findings of a single evaluation** and it is **risky to apply the singular lessons learnt of a project implemented** in a similar – but not exactly equal – context to another one or to a new target group. Contextual factors may strongly influence the results and impacts of an intervention. These may include: the macro-economic and political situation; changing levels of fragility or violence in a region; new challenges due to climate change effects; or particular changes in the sector, like in the TVET-governance structure or in the labor market demand for TVET profiles. In addition, intervention details (e.g., geographical regions, financial volume and management capacity of TVET implementing partners), can be critical to the achievement of results and impacts of a TVET intervention. Therefore, it is recommended that project planners invest their (often very limited) time to **derive conclusions from SRs (meta-evaluations)** which summarize the findings of many impact evaluations.

In this context, it should be emphasized, firstly, that **SRs of rigorous IE** can be used for project planning, provided that they exist for the topic in question. The **DEP and RED database** include SRs of impact evaluations, which fulfill requirements of scientific rigor. However, research within the present report led to the conclusion that there are **no SRs of IEs in DACH DC in the TVET sector**. Secondly, the present report highlights that **SRs of project experiences**, do not fulfill

standards of scientific rigor (like Case Study 9). However, practitioners may still perceive these as very insightful and useful for practice (if relevant RQs are addressed). Thirdly, project planners may also derive some useful insights from result and impact evaluations of previous project phases⁴⁴ or very similar single projects/programs from a similar context, even if the evaluation design does not fully close the causal attribution gap (like rigorous IE). Evaluation findings should ideally be available before the next phase is planned, which is seldom the case (*see Case Study 7 or 8, covering GIZ CPEs*).

Finally, bearing in mind that the OECD/DAC evaluation criteria states that **efficiency** can only be reached if the results are **relevant**, the present report concludes that the use of existing SRs of evidence on results and impacts is a must to ensure **results orientation** in the design and planning of future DC interventions. Contracting organizations are advised to carefully consider the corresponding additional working time for consultants in the preparation of an evaluation. For the specific case of the **TVET sector**, **there is a clear need for more and easily understandable SRs and efforts from the side of the DACH DC to fill the evidence gap, also with a focus on LAC.**

6.2 USING PARTICIPATORY APPROACHES FOR PROJECT IMPLEMENTATION

Impact Evaluations are interventions by their own right yet must also comply with the internationally agreed upon standards and principles for DC interventions, including the Do-No-Harm principle. In addition, the DeGEval Standards (2016) for Evaluation⁴⁵ as well as the DEval Standards (2018)⁴⁶ include professional codes of conduct for evaluators, aiming to avoid possible negative impacts on the cooperation structure, the motivation of project stakeholders or staff, and to safeguard impartiality and the confidentiality of the information sources, among other aspects. Nevertheless, **project implementers frequently fear evaluations** or perceive evaluators' judgements as too strict, subjective or arbitrary, especially if these are not sufficiently explained (as observed through the interviews for Case Study 7 and 8 where GIZ CPE scores and percentage figures lead to this impression). This could prevent, for example, project staff or partners from contributing fully and transparently to the evaluation process and may compromise its results. If project staff and partners do not feel the evaluation results reflect the reality of their project or that the evaluation findings are arbitrary, insufficient, and not comprehensive enough or wrong, this affects the **usefulness of the evaluation** (the first DEval and DeGEval evaluation standard).

At the same time, evaluation departments would like to see project staff and partners interested in and welcoming of **more critical reflections of their work as an opportunity to learn and improve**. Ideally, all stakeholders of an evaluation should be convinced about the **benefits of evaluations for their work and the achievement of planned results**. Therefore, to ensure the use of evaluation results, it is strategical relevant to address the needs of implementers and the benefits that they can derive from an evaluation together with them at the beginning of every evaluation. Equally important is to involve the project implementers in the evaluation process, by using participatory data collection and/or analysis methods, in order to **foster their ownership on the evaluation results**. This is a good practice that **increases commitment** and the **probability that evaluation results will be properly used**, especially for steering. Moreover, the **information needs** of different stakeholders may differ from the standardized OECD/DAC evaluation matrix. They should be identified prior to the beginning of an evaluation, in order to develop a proper design with suitable approaches and methods. Standardized reporting templates (like CPE templates) should allow additional sections to include findings on project implementers specific information needs. This finding was derived from the interviews related to case study 7 and 8. Unfortunately, standardized reporting templates do not always allow adjustments or additional sections, so that valuable information gathered by the evaluation many get lost.

The most basic **participatory approach** includes conducting an inception or **kick-off workshop** with the project implementers to identify their (and other relevant stakeholders like partners or target groups), **information needs to increase the utility of the evaluation**. **Understanding each other's perspectives** (e.g., requirements from the research and practice perspectives), contributes to building a common ground for the evaluation. This inception workshop can also be used to train project staff in rigorous IE methods and create, actualize or improve the ToC (log frame/results matrix), which is needed for conducting the results/impact evaluation and may lead to additional insights for project implementers. The causality hypotheses (reflected in the ToC) combined with the needs (of more or less accurate evidence of their validity) at the side of the implementers will make clear to what extend qualitative and/or quantitative approaches are appropriate to cover the needs and expectations of the results and impact evaluation.

No matter how accurate the evaluation results are, if they are not used, it was not a good evaluation. Therefore, making use of the findings and lessons learnt by reaching the audience in an appropriate manner is critical to the success of an IE. To foster the use of evaluation finding for improving the implementation of an intervention, a **de-briefing and validation workshop** with relevant stakeholders and the project team is strategically critical. Besides **communicating the findings**, the evaluation team should invite the audience to reflect on them, discuss openly on their validity, highlight potential blind spots and complete or argument on them. The conclusions and recommendations can

⁴⁴ For the planning of a follow up phase, IE of the previous phase is seldom available as it is normally conducted ex-post the completion a project or phase.

⁴⁵ The DeGEval Standards for Evaluations are Usefulness, Feasibility, Fairness and Accuracy (DeGEval 2016).

⁴⁶ The DEval evaluation standards are organised according to the criteria of utility (U); evaluability (E); fairness (F); independence and integrity (I); accuracy, scientific rigour and comprehensibility (A); as well as comparability (C) (DEval 2018).

then be derived jointly, in a participatory way, which increases the ownership and commitment of the stakeholders to apply the lessons learnt.

To **increase the quality and acceptance of evaluation findings, conclusions and recommendations**, all evaluation reports should be quality checked by evaluation units and project staff before being returned to the evaluators for finalization and inclusion of feedback from all actors involved. Evaluation reports should ideally be published in a timely manner. For example, CPEs of GIZ are made available via an online database, but it is not clear who reads and uses them (especially if these are published with a delay of one year).

6.3 USING IE FINDINGS FOR POLICYMAKING

Policy is crucial for the development of a sector as it: establishes the framework for action by institutions and organizations (and projects); defines the requirements for actors or approaches; and constitutes the enabling environment for capacity development at all levels (individual, organization and society) by opening or closing the possibilities for innovation. Therefore, most German DC projects and programs follow a holistic approach, addressing the **three levels of intervention** (macro, meso and micro). This is also the case for most TVET interventions.

However, **policy is usually a very sensitive field**, where many influential factors may play a role. Political and economic interests and power or influential groups exist in any sector of the economy and the TVET sector may not be an exception. As policy changes can alter power relations within or between state institutions, or between the state and the private sector, the policy-making process is often long and tortuous. It can involve a variety of actors, in many cases including the parliament if the policy changes need to be approved by law. This is also the case for TVET, which many partner countries are trying to reform and modernize in order to make TVET more labor-market oriented and therefore, more effective in reducing youth unemployment. Germany's DC, for example, supports partner countries in these efforts, by implementing cooperative TVET projects. There are **many structural factors** (varying according to the partner country in question) that need to change at policy level, and in the national TVET systems, to allow the scaling-up of such an approach at the national level.

In this context, impact evaluations can play a crucial role to support innovation efforts. If communicated clearly and adequately, policy makers may use evaluation findings, conclusions and recommendations, for **evidence-based/results-based argumentation and decision-making**. Therefore, all DACH DC actors have internal mechanisms for management response processes and communication formats of evaluations to support policy-making. These can include:

DACH DC foresees creating management responses and the regular monitoring of these, as an effective tool to facilitate the use of evaluation results (SDC 2018, ADA 2020, BMZ 2022). These provide relevant stakeholders with the opportunity to react to the evaluation recommendations and determine the way forward. Recommendations may be accepted in full, partially accepted, or rejected. In case recommendations are accepted (in full or partially), an action plan should be defined, which includes the specific measures, responsibilities and a timeframe for implementing the recommendations. A plain template of an ADA management response can be found in annex 10 of ADA 2020⁴⁷. The implementation status of the management response needs to be regularly monitored and documented to ensure timely implementation (ADA 2020). One part of these management responses (concerning the management's feedback on recommendations) is usually published, while the other part, covering the action plan, is usually used as an internal document that should be followed up at least annually (SDC 2018, ADA 2020, BMZ 2022). The BMZ comments on all DEval evaluations in a **public text-based statement** format that summarizes the relevance of the evaluation and draws key conclusion from the recommendations in an evaluation summary (BMZ 2022), which differs from the **table-based formats** used by ADA and SDC (ADA 2020, SDC 2018, SDC 2019).

The SDC considers the **reporting and communication of evaluation results** as a prerequisite for institutional learning, steering, transparency and accountability. SDC is particularly strong in using **participatory exchange formats** such as capitalization workshops, debriefing as well as discussions of findings with the evaluators, which aim at increasing the acceptance and usefulness of evaluation findings. SDC declares to take senior management responses very seriously and keeps track of the implementation of recommendations on an annual basis. It strive for **transparency** and ensures that evaluation reports with management responses are publicly available (based on interview with SDC and SDC 2018). Similarly, ADA perceives the **communication of evaluation findings, conclusions and recommendations** as key for evidence-based decision-making. They emphasize that different audiences need to be addressed differently and therefore they try to successfully **reach policymakers with short policy briefs or executive summaries and infographics** (based on interview with ADA and ADA 2020).

⁴⁷ ADA Guidelines for Programme and Project Evaluations, Annex 10

CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS

In this chapter, the main findings of the study are summarized, conclusions are derived from these findings and related recommendations are presented. First, general conclusions and recommendations are presented for measuring results and impacts of skills development interventions. These have been obtained from the evaluation designs and approaches used to measure results and impacts of DC-interventions (see *Chapter 2*) and the evaluation policy and trends within Development Cooperation of Germany, Austria and Switzerland (DACH-DC) (see *Chapter 3*). Next, the main findings, conclusions and related recommendations are presented for each of the **focal topics** of particular interest for Inter-American Development Bank (IDB), German Federal Ministry for Economic Cooperation and Development (BMZ) and *Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH* (GIZ).

The interviews and **case studies** examined in this study focused on concrete examples of existing impact evaluations (IEs) of Technical and Vocational Education and Training (TVET) interventions, mainly in German DC, which enabled **conclusions to be made for measuring the results and impacts of skills development interventions**. The research team was unable to find any IEs performed by German DC in Latin America and the Caribbean (LAC). The only presented case study from the LAC region—Case Study 5 from Brazil—was financed by IDB. However, the methodological designs, approaches and methods examined are similarly applicable to other country or regional context like the LAC region. **Methodological IE insights are transferable to other contexts, including LAC, whereas single technical IE findings are very context specific and cannot be transferred to other contexts**, e.g. between regions, countries or even within countries. The research team did not find systematic reviews of rigorous IEs of skills development interventions performed by German DC. If such systematic reviews were found, they could allow generalizing conclusions from a sufficiently large number of existing technical findings, which could potentially be transferred to the LAC context.

1. General Conclusion: The understanding and use of the term rigorous IE differs in research and practice.

There are different definitions and understandings of what a “rigorous” IE is in quantitative research and DC practice. On the one hand, **quantitative prone researchers understand experimental and quasi-experimental designs as the core of rigorous IE designs** (German Institute for Development Evaluation, [DEval 2021](#)). This understanding is in line with requirements for publishing research in peer-reviewed journals. DEval ([2022d](#)) concludes that systematic reviews and evidence gap maps of existing rigorous IE are non-experimental designs, but these are closely linked to rigorous IE because rigorous IEs are the underlying studies for these. On the other hand, **DACH DC practitioners prefer a more comprehensive definition of rigorous or rather robust IE**, which also includes non-experimental designs and thus qualitative, theory-based approaches such as contribution analysis, which is the standard tool most frequently used by Austria’s and Germany’s DC. Austrian Development Agency (ADA) highlights that there is “**no single right or best evaluation design or approach**” emphasizing that each design or approach “needs to be **tailored to the evaluation purpose, objectives and questions**” ([ADA 2020](#)). Many practitioners are of the same opinion, specifying that rigorous methods should stand for the **use of an appropriate method** in the given context (*Case Study, CS 6*).

Related recommendations:

- Consider establishing a **common understanding and usage of the term “rigorous” and “robust” IE** among stakeholders of the evaluation, including what kind of evaluation designs and approaches may be considered when conducting a rigorous or robust IE (see *Chapter 2 and 3*).
- Before selecting an appropriate **IE design, approach and methods**, take the following steps (see *Chapter 5*):
 - ◆ Step 1: Clarify the evaluation **interest and use** and conduct a needs assessment
 - ◆ Step 2: Balance **scope, budget and time** and understand the contextual factors
 - ◆ Step 3: Define the **purpose and objective** of the IE
 - ◆ Step 4: Specify your main **evaluation questions**. Rigorous IE is only appropriate if your evaluation questions address the OECD/DAC criteria impact or effectiveness.
 - ◆ Step 5: Derive the **rigorous IE design, approach and methods** from the RQs, purpose, objectives, contextual factors and practical needs identified above.
- Allocate enough time to make use of **systematic reviews of rigorous IEs** to identify generalizable technical insights from the global scientific knowledge base, which may be transferable to other contexts (see *Chapter 6.1*).
- Start planning and **initiating evaluations during project planning** (not only in ex-post settings frequently found in DACH DC) (see *Chapter 6.2*).
- A **participatory approach** in the IE process will ensure that all actors have a voice and may learn from the IE. It also enhances their ownership on the results, which increases commitment and the likelihood of implementation of the recommendations (see *Chapter 5 and 6.2*).
- Use **management responses** and adequate follow-up for translating learning from evaluations into actions (see *Step 6 in Chapter 5 and 6.3*).
- Develop adequate communication formats to **reach different stakeholders** and audiences according to their information needs and the use they may make of the evaluation findings.

2. **General Conclusion: Non-experimental designs, especially contribution analysis, are the standard instruments DACH DC uses to measure results and impacts of TVET interventions.** Quasi-experimental designs were used in multiple pilot projects in the TVET sector and there is a growing trend and support for them in Germany. The present study did not find any experimental designs (most rigorous method) to measure results and impacts of TVET interventions in DACH-DC practice. This is because these are perceived as rather inappropriate and not feasible in DC as they require randomization, large sample sizes, large volumes of high-quality data with high costs for data collection and ex-ante evaluation settings etc. However, international and German research institutes implemented their own TVET evaluations with the purpose to use randomized controlled trials (RCTs), but they deliberately refrained to collaborate with German DC practice for these experimental designs.

A wide range of varied examples were found for measuring results and impacts of TVET interventions in German DC practice and for the involvement of German universities or research institutes.

DACH DC actors mainly use **non-experimental designs** and apply qualitative theory-based IE approaches, especially contribution analysis, which is a standard tool and evaluation method in Austrian and German DC practice. These non-experimental designs are **much more frequently used** in DACH DC practice than experimental or quasi-experimental designs using different quantitative approaches and methods. Non-experimental designs were applied in:

- Case Study 7 – used a qualitative approach used contribution analysis, Kirkpatrick model and most significant change (MSC) in a central project evaluation (CPE);
- Case Study 8 – used a qualitative approach including contribution analysis, MSC and follow-the-money in a CPE; and
- Case Study 9 – used a qualitative approach studying best practices.

GIZ's CPEs follow a standardized approach using contribution analysis, with some flexibility to add further methodological approaches. Therefore, there are many more examples of non-experimental designs using the qualitative approach contribution analysis in the GIZ CPE database. The selected case studies reflect DC-practice-based examples. German DC—namely GIZ—and researchers or evaluation consultants who specialized in conducting these types of result and impact evaluations for GIZ, created the underlying evaluation reports in collaboration. Also, non-experimental designs require adequate time and resources to ensure high-quality evaluation results (CS 9). **Qualitative approaches are feasible in different DC contexts.** These approaches allow an in-depth understanding of the needs and limitations of the target group, the implementing partners, and other stakeholders, and are therefore useful for designing effective CD interventions in the TVET sector. Many contextual, technical and soft factors, need to be considered for project success, and are less likely to be captured with quantitative approaches (CS 9).

Some examples of **quasi-experimental designs** used to measure results and impacts of TVET interventions in DC of DACH countries were identified. Quasi-experimental designs are used in:

- Case Study 3 – used a quantitative approach using DiD and matching;
- Case Study 4 – applied a quantitative approach using linear multivariable regression model similar to DiD; and
- Case Study 5 – used a quantitative approach using a natural randomized experiment
- Case Study 6 – applied a mixed-method approach using quantitative propensity score matching and qualitative explorative interviews.

It can be concluded that these methodological approaches are **suitable and effective to measure results and impact** of capacity development interventions (CS 3). For example, Case Study 3 states that “evidence (...) shows that **labor market matching interventions** have the largest and most consistently positive employment effects in Jordan.” The underlying IE of these Case Studies (CS 3, 4) were conducted in collaboration between German DC and academic research institutes⁴⁸. Case study 5 (quantitative approach using a natural randomized experimental study) is considered a specific quasi-experimental design, since it relies on a natural experiment. The study analyses the impact of a TVET program implemented by the Brazilian Government. Many lessons learned can be derived from these case studies and are summarized as specific recommendations for quasi-experimental and experimental designs below. Regarding trends, the **importance of rigorous IEs is increasing in German DC (GIZ 2018)**. Due to the new funding program for rigorous IE financed by BMZ ([DEval 2022b](#)) there might be more quasi-experimental impact evaluations of TVET interventions in the near future.

The research conducted for the present study could not find any example of DACH DC for conducted **experimental designs (RCTs)** in the field of TVET. Representatives of the DACH-DC evaluation units interviewed by the present study (namely, ADA and Swiss Agency for Development and Cooperation, SDC) provided the following explanation for this. Rigorous impact evaluations do “not prescribe the use of a specific evaluation design” ([ADA/OeEB 2019](#)). There is **not one single right or best evaluation design or approach**, because the evaluation design and approach “need to be tailored to the specific evaluation purpose, objectives and questions” ([ADA 2020](#)). In DACH DC, a **good evaluation is**

⁴⁸ German DC collaborated with the quantitative research institute RWI in Case Study 2 and 4 and with the qualitative evaluation/research institute CEval in collaboration with a KfW-internal economists in Case Study 7.

one that will be used, e.g., for project design, implementation, policy-making or results achievement. Evaluation designs must **balance what is best⁴⁹ and what is feasible in DC practice**. Despite experimental designs (RCTs) being the “gold standard” and most powerful methods for causal attribution, **RCTs face many challenges in practice**, which explain why these have not been used for TVET interventions in DACH DC, yet. Experimental designs require ethical randomization and a large number of observation units (i.e., large sample sizes). In addition, these designs are usually time-, budget-, and knowledge- intensive, as large amounts of high-quality data have to be collected and analyzed, which requires adequate methodological knowledge in econometrics. Furthermore, results-oriented DCs prefer to be flexible during project and program implementation in line with the adaptive, results-based management approach, to be able to modify necessary elements when the conditions in which the project operates change. This may be contradictory to the needs of experimental designs, e.g., *ceteris paribus* condition for as many variables as possible. Experimental designs have been used in:

- Case Study 1 – applied a quantitative approach using an RCT and
- Case Study 2 – conducted quantitative approach using an RCT.

Case Studies 1 and 2 are scientific research projects of German universities or research institutes in collaboration with other international universities⁵⁰, conducted by academic researchers, who declared during interviews, that they had deliberately refrained from cooperation with DC to be able to implement RCTs.

Related general recommendations:

- **Keep flexibility of evaluation formats and allow various IE designs, approaches and methods** to ensure that the evaluation addresses the practitioners’ needs adequately, e.g., for project planning, implementation or policy-making (*see Chapter 3 and 6*).
- **Balance what is best or more accurate and what is feasible in DC practice**, considering the scope of the evaluation, time, budget, knowhow, contextual and ethical considerations, project size, etc. (*see Chapter 2, 3, 5 and 6*)
- **Consider the whole range of evaluation approaches and methods** and not only the most rigorous, but **place evaluation findings in context**, e.g., by describing in detail the methods used and being clear about their level of accuracy in terms of causal attribution inference (*see Chapter 3, 5, 6*)
- **Allocate adequate resources**, namely budget, know-how and time to any IE (*see Chapter 3 and 5*).

Specific recommendations for quasi-experimental and experimental designs:

- **Ensure close cooperation and exchange between researchers and practitioners** for rigorous IE in DC, because both parties follow different objectives and have different perspectives (CS 4). It is important to bring together “**project thinking**” and “**research thinking**” to build a common ground for the IE, (i.e., the practitioners’ perspective on the respective intervention and the researchers’ perspective on what constitutes an appropriate rigorous IE design). This requires efforts to understand each other’s objectives, constraints and *modus operandi*. Researchers should learn how interventions work and how these are evaluated using existing M&E data. Practitioners may need to understand why researchers require a control or comparison group, stress the importance of the issues of selectivity, randomization of treatment, large sample sizes and comprehensive data for solid empirical evidence. This collaboration should ideally **start when the intervention is being designed or at its inception**, for example by using the support of researchers when setting up the results logic, so that they can develop more detailed pathways to achieve outcomes and test them empirically using robust experimental or quasi-experimental designs (CS 3, 4).
- **Make sure that practitioners get a brief but proper information workshop on rigorous IE designs and approaches** (e.g., from researchers if they are skilled to train), so that they develop a better understanding of the above mentioned “research thinking”. This will increase their willingness to contribute to the rigorous IE implementation, and their interest on the findings, especially if they do not anymore consider a weak IE design and a very strong rigorous IE as equally informative. This will also ensure the use of the evaluation findings.
- **Earmark and allocate adequate resources to local project managers**, who face additional workload for rigorous IE in DC. They play a critical role for the success of a rigorous IE (CS 4). This most likely requires additional resources on top of the regular M&E staff (CS 3).
- **Integrate the impact evaluation in the project planning and implementation from the very beginning** because researchers need to define the variables of interest very easily, in order to be periodically measured along the implementation phase and beyond it. Therefore, the involvement of researchers should start at the conception phase of the (TVET) interventions (CS 4).
- **Plan sufficient time between the intervention completion and the follow-up data collection** to allow treatment effects to unfold and be able to measure long-term impacts (CS 4).

⁴⁹ Most rigorous method for causal attribution.

⁵⁰ The RCT in Uganda (CS 1) was implemented in collaboration between the German Leuphana University and two Ugandan universities and the RCT in India (CS 2) was implemented by an Australian University (Monash University) and a US American university (Fordham University) and commissioned by the German (academic) Institute of Labour Economics (IZA).

- **Precisely document the selection mechanism of the beneficiaries.** This is an important measure to construct adequate control groups for ex-post IE (quasi-experimental designs), but practitioners are often not aware of this (CS 4, 5).
- **Clearly specify implementation periods and exact evaluation dates.** This helps detecting impacts during rigorous IE research. In situations where the rollout of measures is spread over a long period of time, quasi-experimental or experimental rigorous IE designs are not an appropriate evaluation design. If rollout of DC TVET measures takes place over 2 years, this can mean that some people are just newly trained at the end of the project, which makes detecting impacts less likely (CS 4).

Be aware that rigorous IE requires large volumes of high-quality data, which makes data collection **very important and often expensive** sometimes exceeding the regular M&E budget of DC projects/programs. Existing M&E and administrative data has a potential to reduce the costs, but can be used in rather few cases if multiple challenges are overcome (CS 4, 5).

3. Conclusion on the Focal Topic “Using Existing M&E Data”: M&E data is frequently used in non-experimental designs, but rarely used in quasi-experimental or experimental rigorous IE designs, due to: small sample sizes; incomplete and insufficient data in terms of extend and quality; and lack of suitable control group information. Most quasi-experimental and experimental rigorous IE designs collect their own data instead.

Based on the case studies, it was concluded that **non-experimental IE designs**—and especially the CPE of GIZ—use M&E data to measure results and impact. This reduces the costs of these IE, as only limited amounts of new data need to be collected. Using M&E data is a central aspect of CPE. It reveals whether the target values of the monitored indicators have been achieved (CS 7, 8). For example, a CPE analyzed for the present study used monitoring data from multiple countries available via the GIZ-results monitor (web-based monitoring system of GIZ), which led to the updating of internal monitoring data (*in CS 7*).

The evaluators of Case Study 3 and 6 used **quasi-experimental designs** and tried to incorporate existing M&E systems and data in close collaboration with the GIZ M&E team and local researchers. However, the existing M&E data was insufficient and of limited use (CS 3, 6). For some projects, the monitoring data were unavailable or incomplete and the M&E systems were not established before the implementation phase (CS 6). The M&E data was, for example, not suitable to detect interventions-induced effects on the macro-level, as the monitoring system of each particular project/program followed its own logic and was not designed to measure the overall achievements of the interventions (CS 6). In conclusion, existing M&E systems are usually not geared to fulfil the requirements of tailor-made rigorous IE designs. For example, in case of Case Study 3, local researchers were contracted to collect additional data via surveys and process them.

Case Studies 1 and 2 used **experimental designs** but did not collaborate with DC projects and programs, so that M&E data was not available and not used.

Related recommendations:

- **Set up M&E systems** at beginning of the project/program **with results and impact indicators**, baseline and target values, and **keep the monitoring data up to date**
- **Ensure that monitoring systems of multiple project and programs are aligned** to enable measuring overall achievements.
- **Consider using, updating, collecting and completing existing M&E data** for measuring results and impacts.
- **Earmark and allocate adequate resources for new data collection** in case the M&E data is not or insufficiently available and/or not useful for measuring results and impacts. This is very likely in case you are using quasi-experimental or experimental IE designs.

4. Conclusion on the Focal Topic “Using Existing Administrative Data”: Administrative data has the potential to drastically reduce the costs of rigorous IE. However, these data sources are rarely available in an adequate data quality, and they are not easily accessible. Therefore, the rigorous IE examined have collected additional data or used purely new data.

Three of the case studies examined present examples for the use of **administrative data for quasi-experimental rigorous IE designs** (see CS 3, 5, 6). Case Study 6 from the Philippines used national statistics on vocational education and labor market data⁵¹ for a **preliminary framework analysis** but not for the quasi-experimental IE (CS 6). Case Study 3 from Serbia, constitutes a particular **good practice example for administrative data use for a quasi-experimental IE design** (DiD and matching). Administrative data was successfully used, with the support of national stakeholders, in the following two ways: 1.) the Serbian Institute for Improvement of Education and Upbringing helped to identify comparison profiles and comparison schools, while the Serbian Ministry of Education, Science and Technological Development provided additional administrative data on enrolment scores and established the contact with comparison schools. This enabled the use of a DiD design for the IE of the German-DC Reform of

⁵¹ i.e. from the National Statistics Office, the National Statistical Coordination Board, the Bureau of Labour and Employment Statistics, the National Wages and Productivity Commission and the International Labour Organization.

Vocational Education and Training project. 2.) Access to large-scale administrative data from the Serbian National Employment Service enabled the use of statistical matching methods and creating a control group for the IE of the German-DC Youth Employment Promotion project (CS 3). In Case Study 5 from Brazil, the research team was able to use the **administrative data on training applicants provided by the TVET institutions**. Based on the applicant lists, including names, phone numbers, gender, education, which allowed the research team to track and interview training graduates. Although the main data had to be collected, the **administrative data provided an entry point** for the subsequent data collection. Moreover, the provided applicant lists also contained data on individuals that were not randomly selected for course admission, which allowed the researchers to **create a control group retrospectively**.

Both Serbian IEs highlight the following four **success factors for the use of administrative data for IE**: 1.) The close collaboration with national stakeholders was key—their support and interest in the research enabled the use of existing administrative data. 2.) The availability of existing administrative data in sufficient quality was a precondition for being used in the IE. This required a critical quality assessment of the data as well as technical know-how for understanding and analyzing the administrative data⁵². 3.) Many privacy and data protection concerns were solved to the use of administrative data, e.g., generating large, anonymized datasets. Overcoming these challenges enabled the successful use of administrative data for this IE (CS 3).

The case studies reported **multiple challenges**, which cause administrative data to be rarely used for IE so far. These challenges include: 1.) National statistics may not be available in a highly aggregated form (CS 6) or do not allow a disaggregation for the respective target group, so that the data is not or only partially suitable to address the evaluation questions, e.g., national statistics on unemployment used in CS 7. 2.) Researchers often face challenges accessing administrative data or these data is either not available or of low quality and thus unsuitable for the research purpose, so that experimental designs (CS 1, 2) exclusively collected their own baseline, midline and end-line survey data for the RCTs. 3.) In some cases, useful data from international organizations may be publicly available too slowly and thus perceived as outdated (even for the use of the non-experimental design in CS 9).

Thinking about the TVET sector in LAC and the use of administrative data for evaluations: Many LAC countries have very good administrative data available, especially for the education sector (CS 5). National Statistics offices as well as the Ministries of Education and Ministries of Labor of many LAC countries have developed, in the last two decades, highly technical capacities to collect, analyze and store data. Many LAC countries make this data, which may be disaggregated by gender, age, location, type of school, etc., publicly available on the internet and update it continuously. The level of cooperation with national and regional administrative officers can differ from country to country and from region to region, but based on working experience in LAC countries it has been observed that in most cases cooperation works well.

Related recommendations:

- **Closely collaborate with national stakeholders** to request their support in accessing administrative data and **assess the possibility to use existing administrative data for IE** (e.g., of employment effects), as administrative data can be a cost-efficient alternative or rather a complement to newly collected data.
- **Critically analyze the availability and quality of administrative data** before using it, incl. the reliability, validity, and timeliness, as well as the level of data aggregation.
- **Allocate adequate technical knowhow for understanding and analyzing administrative data** and for overcoming privacy and data protection challenges.
- **Offer the national institutions that contribute to the evaluation providing data to inform them on the findings**, as these could be also useful to improve their work in the sector.

5. Conclusion on the Focal Topic “Gender”: Some DC projects and studies focus on women, so that targeted indicators were used to indicate progress for women only. Most DC-projects/programs and IEs target men and women and tended to use corresponding disaggregated indicators. Despite existing progress in gender-responsive M&E, there is room for improvement in IE reporting. Gender-specific results and impacts are not always reported, even though gender-disaggregated indicators exist.

In the selected TVET case studies, the main characteristics of vulnerable groups were gender (women), age (youth) and people from poor or immigrant backgrounds, like IDPs living in immigrant settlements. **TVET interventions may target women** or specific groups of women, such as women aged 18 to 39 years, exclusively (see CS 2 using an RCT). The selected IE measure the results and impact of training participation for women exclusively. In Case Study 2, **women-targeted indicators** were used and a disaggregation by gender did not apply in this context. The RCT from Case Study 2 found that TVET raised the employability and earnings of women in the target area of India. The IE was conducted in a gender-sensitive manner, involving female enumerators and looking at barriers to women’s participation and continuation of training. Some main findings refer to the lack of credits/resources, lack of adequate childcare and security/safety concerns in reaching trainings centers which is also applicable to the participation of women in TVET in the LAC context (CS 2).

⁵² This includes software skills for accessing specific dataset formats and quantitative data analysis.

Qualitative approaches may focus, explicitly on the **underlying barriers and success factors for gender equality** (focusing mainly on women and girls) in TVET. For example, Case Study 9 used a non-experimental design and a qualitative meta-study approach to identify success factors from a literature review. Ideally, this desk research should have been accompanied by on site research to verify the reported results, especially because project reports and evaluations do not contain sufficient information. To better understand the access barriers to TVET or education for girls, the target groups and their family members (especially parents) need to be involved in the data collection to identify these, e.g., safety on their way to school, in school or in internships, gender discrimination within the family, traditional gender roles, etc. For employment project/program evaluations, it is necessary to consider that a large proportion of women and girls work in the informal sector especially in rural areas in LAC countries, what makes them harder to reach, e.g., for enrollment in TVET measures, etc. (CS 9).

In situations where the intervention does not exclusively target beneficiaries of a single gender, it is necessary to measure the results and impacts for all genders. Experimental or quasi-experimental IE designs in the case studies included **gender-disaggregated indicators** and operationalized gender as a dichotomous variable with the two characteristics, namely men or women. This makes it possible to differentiate the effect of the variable gender on the causal relationship between the intervention and outcome/impact variable (CS 1, 3, 4, 5, 6). If no significant difference is observed using quantitative IE approaches (*as in CS 1 or 6*), the IE reports do not mention that at all or do not explain this observation in detail. For example, in case of Case Study 6, the results show that the salary of female and male graduates did not differ, except for the first salary, but no further explanation is provided. In most quantitative studies that assess the effects and impacts of TVET interventions, gender is considered as one of multiple socio-economic characteristics of respondents, such as age, previous work experience, etc. (CS 4). In case of significant and robust differences between genders, further in-depth studies (like qualitative interviews or FGD) would be necessary to better understand the reasons underlying that findings (CS 5). For example, Case Study 5 revealed a positive impact on female TVET graduates but no effect on males. However, the researchers were unable to identify the mechanisms causing the gender heterogeneity.

Although **non-experimental IE designs** may offer the opportunity to examine gender-specific effects in depth, this is not automatically the case and was not the case in the selected CPE Case Studies 7 and 8. The internal M&E system of the DC-projects contained multiple **gender-disaggregated indicators**, so that differences were monitored at the project/program level and forwarded to the evaluators. The CPEs analyzed by this study reported on the indicators, including those disaggregated by gender, but did not **specifically address gender in the evaluation report** (CS 7, CS 8), which leaves room for improvement.

Related recommendations:

- **Consider that “gender equality” addresses all genders, not only women**, and be aware that the related results/impacts may be highly context-specific.
- **Measure gender as a targeted or disaggregated indicator.** In case of disaggregation by gender, assess the possibility of using dichotomous (female/male) or categorical variable (female/male/diverse), for example.
- **Use quantitative assessments to identify a significant or insignificant effect of the control variable “gender” on the causal relationship** between the intervention and the effects and impacts on the variables of interest.
- **Use additional qualitative approaches and methods to gain in-depth insights** about significant differences and surprising findings and the reasons for them.
- **Include reflections on the use of specific or disaggregated indicators** and on any gender-specific differences or surprising findings and explanations in evaluation reports and include the reasons.
- **Report if no differences are observed between genders**, as this is also a relevant finding.

6. **Conclusions on the Focal Topics “Measurement of ‘Employability’, ‘Entrepreneurship’ and ‘Non-cognitive skills’”:** While no coherent definition of employability can be identified, most studies measure employability in terms of TVET graduates’ labor market outcomes such as current employment, time until employment was found, and wage earnings. So far as observed in the present study, fewer studies include additional outcomes like non-cognitive skills and entrepreneurial skills of TVET graduates. These concepts can be measured with specific instruments based on questionnaires and personal interviews about graduates’ behaviors in different scenarios. The majority of the studies focuses on individual employability of graduates while the perspective of potential employers is considered less frequently.

One of the main purposes of TVET projects is to contribute to the reduction of youth unemployment by providing **occupation-related knowledge and skills**. TVET projects aim to improve the **employability of youth**, which refers to formal and informal employment as well as to **self-employment** (entrepreneurship training). Thus, evaluations of TVET projects usually investigate the impact of TVET projects on employability-related outcomes of TVET graduates (CS 2, 3, 4, 5, 6, 8). Some evaluations focus on the skills that were gained through the trainings (CS 1, 5). Most of the case studies **address employability from the graduates’ perspective** while only Case Study 6 additionally includes employers or potential employers.

There is no explicit and clear definition of the concept “employability” that could be derived from the interviews conducted and case studies analyzed. **Employability is usually measured in terms of outcomes of TVET graduates** rather than measuring the propensity of students to obtain a job. The case studies in the present report provide several examples of measuring such **employability-related outcomes**. Employability-related questions were integrated in field surveys and assessed whether the graduates were currently employed, how long it took them to find employment after the training, and whether they found the practical tasks during training useful to find a job (CS 1, 3, 4, 5, 6). Other studies also included aspects of work earnings, formal employment, working hours, and the field of employment (CS 2, 3, 4, 5). In terms of **self-employment**, Case Study 2 measured outcomes such as ownership of a sewing machine and membership in a savings and credit association to assess whether graduates were aiming at starting a business. Thus, these approaches to employability operationalization rely on measuring the concept in terms of outcomes where characteristics such as the job type (i.e., formal/informal, full-time/part-time, occupation in area of training) and the timing (i.e., employment was found in certain period after training completion) are assessed, usually in quantitative quasi-experimental or experimental studies.

So far as observed in the present study, fewer studies measure the **TVET impact on acquired knowledge and skills** (CS 1, 5). However, one relevant conclusion from interviews was to not only look for labor market outcomes, but for non-cognitive skills to understand the channels and mechanisms of TVET projects in more depth (CS 5). For example, Case Study 1 used quantitative methods to measure **entrepreneurial skills** such as action knowledge, entrepreneurial goal intentions, action planning and entrepreneurial self-efficacy. Through personal interviews and questionnaires, the graduates were asked to identify suitable actions for different scenarios, to rate their likeliness to pursue start-up activities, and about their business plans. Case Study 5 measures the effect of the TVET intervention on **non-cognitive skills** such as the agreeableness⁵³, conscientiousness⁵⁴, and extraversion⁵⁵. These skills were measured based on the Social and Emotional Nationwide Assessment inventory in Brazil, which was developed by the Ayrton Senna Institute. Interestingly, both Case Studies hypothesize that the **acquired skills function as a channel through which training effects translate into labor-market outcomes** (CS 1, 5). Those skills might increase the propensity of graduates finding a job and therefore deserve greater attention in TVET evaluations.

Another measurement approach to **employability relies on subjective self-assessment**. In those cases, TVET graduates were asked to self-assess their acquired skills and qualification, the usefulness of those skills as well as their preparedness for the job market based on their own judgement (CS 7, 8). For example, in Case Study 8, graduates were asked whether they thought that the intervention was beneficial for improving their individual employment prospects. Since the employment effects of the intervention had not materialized at the time of the evaluation due to COVID-19 related training postponements, Case Study 7 asked TVET graduates about their perceived relevance of the acquired qualification and used this as a proxy for employability. Those studies usually apply a qualitative research approach including interviews/FGD.

Most of the case studies address employability from the graduates’ perspective but it is also important to **consider employers and potential employers**. Case Study 6 studied the interventions’ impact on TVET graduates’ labor market outcomes quantitatively but also includes qualitative interviews and FGD with enterprises. The representatives from the enterprises were asked whether they employ TVET graduates and whether they are satisfied with their knowledge and skills.

Related recommendations:

- **Consider employers and potential employers** as main source of information for measuring the employability of TVET graduates, because they know best what knowledge, competencies, hard and soft skills TVET graduates must have to qualify for employment, thus being “employable.” TVET graduates’ self-assessments are an additional and more subjective source of information.
- **Check the existing IE literature and good practices before inventing a new definition and operationalization of a TVET concept.** Use definitions and operationalization of core TVET concepts, which were tested and found useful in practice.
- **Consider the use of proxy variables** to measure the impact on employability and entrepreneurship.
- **Consider evaluating the impact of TVET interventions on additional outcomes such as non-cognitive skills and entrepreneurship skills.** This evaluation may deliver deeper insights into the mechanisms and channels.
- **Decide whether a qualitative approach including graduates’ self-assessment** could provide valuable information to answer your research questions.

⁵³ i.e., tendency to act cooperatively

⁵⁴ i.e., tendency to be organized and responsible

⁵⁵ i.e., orientation towards external world

7. **Conclusion on the Focal Topics “COVID-19”, “Green Transformation”, and “Technological Change”:** As in many other areas, the COVID-19 pandemic has led to increased use of internet connectivity, digital tools and platforms, etc. to continue DC operations, including the conduction of remote evaluations. Video calls were used for coordination, (incl. kick-off and validation workshops), as well as for data collection, (such as virtual interviews and focus group discussions). Reduced travel and technological change are beneficial in terms of green transformation and climate change. However, purely remote data collections can lead to blind spots (e.g., due to insufficient observations and limited access to confidential data), which can only be overcome by on-site evaluation activities.

Many IEs presented as case studies were **affected by the COVID-19 pandemic**, especially if the **data collection** had not taken place before March 2020⁵⁶. The pandemic led to a reduction and/or delays of on-site data collection. During 2020 and 2021, most IEs had to be conducted fully remotely or semi-remotely, because international evaluators could not travel and local evaluators were still able to travel just in few countries. Some faced many restrictions to meet, especially for group discussions, depending on how restrictive the national regulations to control COVID were. Therefore, most data collection had to be operationalized virtually and with strong local support. This enhanced the role of local evaluators and led to higher transaction costs for the evaluation, in terms of increased need for close virtual coordination and communication within the international and national evaluation team to adjust procedures, reflect on findings and share learning experience, like in Case Study 7, which utilized semi-remote CPE. Other evaluations had to be entirely conducted remotely, meaning all data was collected virtually. Fully remote evaluation missions of DC projects/programs did not exist before COVID-19, and therefore constituted a major challenge for the evaluation teams, the evaluation management, and the contractors, with international and national experts conducting FGD and interviews online. Onsite observations were not possible, which was a major constraint and compromised the quality of the evaluations, as important information was missing. In other cases, planned evaluations were simply cancelled.

The pandemic affected not only the data collection and processes for IE, but also the **DC interventions were affected drastically and had to adjust strategy, operational planning, implementation of activities, as well as time schedules and budget**. Many projects were no longer able to reach their target groups. In the case of TVET, **implementation delays were significant, as schools were closed for months and agreed internships for students in companies were cancelled. In many countries, COVID-19 regulations did not allow for onsite meetings with ministries officials and policy reform got stuck**. Some DC organizations brought their **international personnel back to their home countries** and they had to operate remotely with the local team for many months. Some examples of the effects of the COVID-19 pandemic on DC interventions are described in the case studies examined. In Case Study 1, the training implementation was interrupted for more than one year. In case of Case Study 9, more mentoring and guidance (especially for girls) was needed by (female) role models. Planned on-site trainings had to be modified and conducted in online training formats, but were only found to be successful for (refresher) trainings of trainers, not for participants of the target group. The project learned from this experience, that different methods are required in online trainings compared to physical trainings to make these interesting for participants (CS 9). Further, many TVET interventions target vulnerable population, who do not have the necessary hardware, software, and sufficient internet connectivity to attend trainings online. In addition, the already precarious situation of many TVET participants became even more critical, increasing the number of dropouts from TVET schools. At the same time, the pandemic has fostered technological change, being an opportunity for the development of new apps, a more extensive and diversified use of mobile phones and a driver for the digitalization of all sectors, including education. Therefore, digitalization in the TVET sector has become a priority for many donor and partner countries and this political is turning into funding and is becoming an integral part of any TVET intervention being planned recently.

Related recommendations:

- **Monitor new rigorous IE publications** as there were delays of interventions and of IE during the COVID-19 pandemic.
- **Critically assess whether international travel is necessary** to conduct IE or whether semi-remote evaluations with the support of national experts would work as well. This could be beneficial for capacity development in the partner countries for enhancing ownership on evaluation results and a contribution to the green transformation.
- **Keep some degree of flexibility in the evaluation design**, as to be able to adjust methodology if changes occur or exogenous factors impede the implementation of the evaluation approach as initially planned.
- **Be aware that new apps have been developed** in the last two years that could be useful to collect data where the internet and mobile phones are available.

⁵⁶ The World Health Organization declared the novel coronavirus (COVID-19) a worldwide pandemic on 11th of March 2020. Many countries closed their airports for international and national travel in the subsequent weeks.

8. **Conclusions on the Focal Topic “Sustainability of Impacts”:** The case studies present examples of sustained impact measurement at the beneficiary, institutional and/or systemic level. This required regular tracing of participants, institutional or systemic actors and the control group since inception, over longer periods of time and beyond the project completion. However, evaluations rarely assess the sustained impact of TVET interventions beyond a period of two to three years after training completion.

Sustainability of impacts is usually measured in terms of **sustained project impact after the end of TVET interventions**. In the selected case studies, the sustainability of impacts was measured **at the individual, institutional and/or systemic level**.

It is difficult to measure the **sustainability of impacts of capacity development measures at the individual level**, because after the end of the intervention it is often difficult to reach former participants and even more so, the control group. Regular follow-up of participants and the control group must be ensured. Ideally, control and treatment groups should be defined at the very beginning of the project and the periodic tracing of participants of both groups should be part of the intervention since the inception phase (CS 7). The case studies present various examples for measuring the sustainability of impacts. In Case Study 2, the **sustained impacts** after the project end were measured at the **individual level**. The evaluators conducted a **survey 6 months after the training** and **again 18 months after the training** to analyze whether positive training effects for beneficiaries were sustained over a longer period (CS 2). In Case Study 4, **longitudinal data** was collected, including a baseline survey and four interviews with the treatment group, which consisted of graduates from intervention schools who had chosen to participate in the program. The control group, which consisted of graduates from non-intervention school and graduates at intervention schools who did not choose to participate in the program, was interviewed three times. The tracer study collected interview data at different points in time, after 9, 12 and 24 months after treatment, which enabled to assess the **medium-term impact of training and therefore the sustained project impacts up to two years after program end**.

Case Study 1 studied the long-term effects of the training over more than 32 months, which was possible because the Student Training and Entrepreneurial Promotion (STEP) intervention was implemented by the partner institutions and accompanied by data collection over the course of three years. The evaluators found sustained impacts at the **partner/institutional level**, as many partners continued implementing the trainings on their own after the three-years program implementation period, so that the trainings were fully locally implemented. This local institutional ownership guarantees sustainability.

In Case Study 6, the sustainability of impacts was measured at the **individual, institutional and systemic level**. The study **found lasting positive effects at the individual level** (e.g., better qualification) after the support was completed. However, **sustainability was lacking at the institutional level and the systemic level**. Firstly, training institutions seemed to be incapable to maintain the equipment they had received from the project, program or single DC intervention measure, after it ended. Secondly, the study found no diffusion effect in terms of other training institutions adopting the introduced training approaches. Thirdly, the majority of support interventions focused on the manufacturing/industrial sector although the largest future potential is in the financial and health service sector.

Related recommendations:

- **Specify the definition of sustainability of impacts** and whether the study is interested in the sustained impacts at the **individual, institutional and/or systemic level at the beginning**. The level of analysis influences the data collection.
- **Define the beneficiaries, institutions or system** (incl. treatment and control groups in case of an experimental or quasi-experimental design) that will be subject of the IE, ideally at the very beginning of the project. This will allow the beneficiaries, institutional or systemic actors to be periodically traced during implementation and create a database for the evaluation.
- **Consider taking a holistic capacity development** approach from the outset of the project, encompassing capacity development at individual, institutional/organizational and systemic/societal levels. This approach will substantially contribute to the **sustainability of impacts** – due to creating institutionalized structures and perpetuating knowledge and skills.
- **IE need ex-post data of beneficiaries**, but they are difficult to reach after project completion. This **must be ensured during the project period**.

9. **Conclusions on the Focal Topic “Efficiency and Cost-Effectiveness”:** The case studies present examples for using the **value-for-money approach to measure efficiency** and the **follow-the-money approach to measure the cost-effectiveness of TVET interventions**.

GIZ’s CPEs usually apply a **follow-the-money approach** to address the OECD/DAC **efficiency criterion** and combine information on project costs and project results (CS 7, 8). The approach is split in two parts: 1.) Production efficiency, which compares allocated resources and outputs; and 2.) Allocation efficiency, which compares allocated resources and outcomes. GIZ’s Corporate Unit of Evaluation has developed an excel tool to standardize this efficiency analysis, which allows to link allocated financial and human resources to certain outputs (production efficiency) and outcomes (allocation efficiency). The evidence for this analysis is collected in interviews and discussions with project staff (CS

7). In Case Study 8, the **GIZ's KOMP (cost-output monitoring and prognosis) tool**, which assigns costs to outputs during the project implementation, was not systematically used for the monitoring of costs per output in the projects starting before mid-2019. Instead, the evaluation team reconstructed together with the project teams retrospectively a rough estimation of the percentage of costs allocated to the respective outputs to enable the evaluators to complete the efficiency tool of CPE as a data basis for the efficiency analysis. Human resources could be allocated more precisely to the outputs.

In Case Study 4, the **value-for-money approach** was used. This approach led to the conclusion that the **cost-effectiveness ratio** of the Kenyan Association of Manufactures program was 0.00021-0.00036 jobs per EUR invested or 2,778-4,762 EUR per job⁵⁷ depending on the assumptions made. The value-for-money approach is a common tool to evaluate, but not implement, the cost-effectiveness of interventions. It summarizes a complex intervention in a ratio of total impact to total costs and allows comparisons of interventions easily. The report described and applied a step-by-step guide for cost-effectiveness analyses to the Kenyan Association of Manufactures program (*see evaluation report of CS 4 pages 243-275 for more information*).

Related recommendations:

- **Consider applying the follow-the-money approach to assess the efficiency** of the TVET intervention. Examples for the application of this approach can be found in the underlying reports of Case Study 7 and 8.
- **Consider applying the value-for-money approach to assess the cost-effectiveness** of the TVET intervention. If this approach will be used, follow the step-by-step guide in the underlying report of Case Study 4.

10. Conclusions on the Focal Topic “Private sector”: The present study on measuring the results of skills development interventions provides some exemplary insights for the involvement of the private sector in the implementation and evaluation of skills development interventions. However, these insights cannot be generalized for other contexts, like the LAC region.

Special attention should be placed on two interesting impact evaluations in relation to private sector involvement:

In Case Study 6 - Philippines, the evaluators concluded that the private sector, represented by associations, had only been involved in some of the evaluated programs but a systematic involvement was missing. Looking at the evaluation itself, the private sector was explicitly included through **structured group discussions with representatives from industry and business associations**. Specifically, the evaluators aimed at identifying private sector acceptance of the dual training system as well as its diffusion. The study concluded that future TVET projects should involve private sector stakeholders because the top-down approach focusing on the regulatory authority, like Technical and Vocational Education and Skills Development Authority (TESDA), was not effective.

In Case Study 4 - Kenya, the BMZ's Employment and Skills for Development in Africa (E4D) initiative was implemented by the Kenyan Association of Manufactures in collaboration with national training providers and member companies, so that the private sector was strongly involved in the implementation of these TVET interventions. The two program components of the Kenyan Association of Manufactures, work readiness training and internship placement, improved access to jobs, economic opportunities and labor market outcomes for youth as well as increased jobs in the manufacturing sector. The work readiness training rather than the internship placement program mainly drove the positive effects on labor market outcomes. The quantitative impact evaluation described in this Case Study collected data from graduates exclusively. However, there has been an additional qualitative evaluation, which gathered data from the companies involved in the implementation using **semi-structured interviews with company representatives**. The qualitative study (not covered in the presented Case Study) focused on investigating how a change in the internship stipend funding from E4D to companies affected companies' ownership and sustainability of internship placements of the Kenyan Association of Manufactures program.

Related recommendations:

- When implementing a skills development project/program with private-sector involvement, **consider the systematic involvement of the private sector in the program design** to ensure context suitability, labor market relevance, and intervention quality.
- In these cases, it is crucial to raise **the perspectives of the private sector** (e.g., industry and business associations or single companies) **during the data collection of impact evaluations**, (e.g., using semi-structured interviews or focus group discussions with representatives of the private sector). This may lead to additional insights and understanding their perspectives in-depth.
- **Consider taking a mixed-methods approach**, because purely quantitative impact evaluations are likely to conduct surveys with the target group and therefore participants or graduates of TVET measures only.

⁵⁷ One EUR is equivalent to 0.9951 USD in November 2022. The exchange rate of the EC was used for the conversion (EC 2022).

ANNEXES

ANNEX 1 EVALUATIONS SELECTED FOR ANALYSIS

#	COUNTRY	CITATION (WITH FIRST NAMES)
1	Uganda	Gielnik, Michael M.; Frese, Michael; Kahara-Kawuki, Audrey et al. (2015): Action and Action-Regulation in Entrepreneurship: Evaluating a Student Training for Promoting Entrepreneurship. Academy of Management Learning and Education. https://doi.org/10.5465/amle.2012.0107 [accessed online on 07.09.2022]
2	India	Pushkar Maitra, Subha Mani, Learning and Earning: Evidence from a Randomized Evaluation in India. Published in: Labour Economics (Special Issue on Field Experiments in Labor Economics and Social Policies), 2017, 45: 116-130, https://www.sciencedirect.com/science/article/abs/pii/S0927537116303384 or https://docs.iza.org/dp8552.pdf [accessed online on 01.03.2023]
3	Serbia (incl. Jordan, Rwanda)	Bachmann, Ronald; Kluve, Jochen; Martinez Flores, Fernanda; Stöterau, Jonathan (2019): Employment impacts of German development cooperation interventions: A collaborative study in three pilot countries, RWI Projektberichte, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen. Project report commissioned by "Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH". Final report. August 2019. http://hdl.handle.net/10419/215904 [accessed online on 07.09.2022]
4	Kenya (incl. Uganda)	Ebert, Cara; Flörchinger, Daniela; Frohnweiler, Sarah; Ihring, Stephanie; Rosadio Cayllahua, Karen Micaela (2021): Employment and income effects of skills development interventions: An impact evaluation of three employment promotion measures in Eastern Africa within GIZ's employment and skills for development program, RWI Projektberichte, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen. http://hdl.handle.net/10419/251877 [accessed online on 07.09.2022]
5	Brazil	Camargo, Juliana, Lima, Lycia, Riva, Flavio and Souza, André Portela. "Technical Education, Non-cognitive Skills and Labor Market Outcomes: Experimental Evidence from Brazil" IZA Journal of Labor Economics, vol.10, no.1, 2021, pp.-. https://doi.org/10.2478/izajole-2021-0002 . https://www.sciendo.com/article/10.2478/izajole-2021-0002# [accessed online on 07.09.2022]
6	Philippines	Silvestrini, Stefan; Garcia, Melody. Joint Expost Evaluation 2010 – Dual Vocational Training, Philippines. Centrum für Evaluation, Saarbrücken 2010. https://www.kfw-entwicklungsbank.de/migration/Entwicklungsbank-Startseite/Development-Finance/Evaluation/Results-and-Publications/PDF-Dokumente-L-P/Philippines_Dual_Vocational_Training_2010.pdf [accessed online on 07.09.2022]
7	Global	Petersdorff-Campen, Lukas von; Pavel, Bogdan; Njie, William; Valdés Herrera, Fernanda; Jussupova, Arailym; Ayoub Tinni, Bachirou (2022): Skills for Reintegration, global project; Evaluation Report. Central Project Evaluation 2016.2180.4. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, https://mia.giz.de/qlink/ID=249722000 [accessed online on 07.09.2022]
8	Egypt	Abdel-Massih-Thiemann, Judith; Döhne, Thomas; Saleh, Bahaa (2021): Enhancement of the Egyptian Dual System (EDDS). Central project evaluation 2015.2156.6. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. https://mia.giz.de/cgi-bin/getfile/53616c7465645f5fb5cf642a8008ef7d20953d965e4cb0fac53163ed2bc47afdda56a32c01fafee34b05456eb6ff291fc4e93cbd056c96311b2f0e76f610ef97/giz2021-0094en-projectevaluation-employment-promotion-egypt.pdf [accessed online on 07.09.2022]
9	Global	Dietz, Eva; Emil, Anna; Müller, Svenja (2021): The future is equal: Success factors for gender equality in vocational education and training. Bonn: GIZ, https://mia.giz.de/qlink/ID=248444000 [accessed online on 07.09.2022]

ANNEX 2 PERSONS INTERVIEWED

NAME	ROLE
Lukas Petersdorff	Evaluator, Mainlevel GmbH
Fernanda Martinez Flores	Evaluator, RWI Essen Fellow
Eva Dietz	Evaluator
Cara Ebert	Evaluator, RWI Essen Fellow
Michael Gielnik	Researcher, Leuphana Universität Lüneburg
Judith Abdel-Massih-Thiemann	Evaluator, thiemannconsulting
Stefan Silvestrini	Evaluator, CEval GmbH
Wolfgang Meyer	Evaluator, Centrum für Evaluation (CEval). Universität des Saarlandes.
Juliana Camargo	Evaluator, Fundação Getúlio Vargas (FGV EESP Clear), São Paulo, Brazil
Pushkar Maitra	Evaluator, Monash U Clayton Campus
Subha Mani	Evaluator, Associate Professor of Economics, Fordham University
Romana Tedeschi	Swiss Agency for Development and Cooperation (SDC), Head of the Evaluation and Controlling Unit
Sigrid Breddy	Austrian Development Agency (ADA), Director Evaluation & Statistics

ANNEX 3 SEMI-STRUCTURED QUESTIONNAIRES FOR INTERVIEWS

Measuring Results and Impact of TVET Interventions

Interview guideline

Objective: To collect feedback on perceptions, approaches and experiences with different impact evaluation methods of TVET projects/programs. The methodological focus is on impact evaluation, while effectiveness and sustainability measurements will be considered as well.

List of questions for evaluators of selected evaluations (case studies)

Questions about the evaluation methodology:

1. Does a methods section or (inception) report exist which could be of interest to us in terms of a more detailed description of the impact evaluation methods used?
2. Which method(s) did you choose for evaluating the impact of projects/programs? Why did you choose the respective impact evaluation method?
 - a. What was in favour of this quantitative/qualitative method? What are strengths of the methodological approach?
 - b. What are weaknesses of the methodological approach? What were major challenges?
 - c. What were arguments against other impact evaluation methods?
 - d. By whom and how were the methods selected?
 - e. What makes your methodological approach to impact measurement innovative?
3. How rigorous do you consider the methodology used?
 - a. Is the target group reached (treatment group) compared to a control group not reached?
 - b. How were project outcomes attributed to impacts? How and over what time periods are graduates tracked? (e.g. survey, tracer study, alumni networks, company surveys, expert reports, ...).
 - a. How are external factors filtered out?
4. Which difficulties and potentials of the method do you see?
 - a. What limits validity and how can it be increased?
 - b. How can the cost-effectiveness of impact evaluations be increased?
 - c. What methods would you use today to measure impacts of TVET?

Questions about contextual factors of the evaluation:

5. What was the project/program volume to be evaluated?
6. What were the financial costs of the project/program evaluation?
7. How long did the evaluation last (expert working days and time span)? When did the evaluation start (project planning, project start, beginning/mid/end of project implementation, after project end)?
8. How were the evaluation results validated and communicated?
9. How were the evaluation results used? (Who is worth talking to at the client/funder to follow up on how recommendations were used?)

Questions about focal areas:

10. How was existing data used, e.g. from the project/program's M&E system or other stakeholders (institution of administration, vocational training or labor market research, or other publicly available sources)?
11. How did your evaluation...
 - d. Measure the sustainability of impacts? How?
 - e. Measure the effectiveness of the impacts? How (e.g. describe approach to cost-benefit analysis)?
12. Did your evaluation...?
 - a. Captured specific impacts for women and girls?
 - b. Captured specific impacts for vulnerable group
 - c. Used digital applications (digital tools)
13. Did the Corona pandemic influence the evaluation or impact the use of evaluation results?

Questions about impacts:

14. What concepts were used for the impact evaluations of TVET projects/programs? (e.g. employability, skills improvement, capabilities, life skills, ...).
 - a. How do you conceptualize/ define and operationalize/ measure the impact (impact) "employability in the labor market"?
 - b. How do you conceptualize/define and operationalize/measure the impact "income changes of graduates"?

- c. How do you conceptualize/define and operationalize/measure the impact (outcome) "skills improvement of graduates"?
15. What are the most important impacts you identified in your evaluation of the TVET project/program? What are lessons learned, success or failure stories in the evaluation of the TVET project/program?

Questions about how the results will be used:

16. How will the results from this (impact) evaluation be used by a TVET system/program? What other opportunities do you see (related to this evaluation and/or in general)? (e.g. use of evaluation results for project planning, steering, evidence-based decision making or learning).
17. What impacts and methods can be transferred specifically to the Latin American context (LAC region)? Why?
18. *Our GOPA team is designing a case study after the interview. May we reach out to you again in case of any questions?*

Measuring Results and Impact of TVET Interventions

Interviewleitfaden

Zielgruppe: Mitarbeiter*innen von Evaluationsabteilungen deutschsprachiger bilateraler Entwicklungsorganisation

Zielsetzung: Feedback zu Auffassungen, Ansätzen und Erfahrungen mit verschiedenen Methoden der Wirkungsmessung von Berufsbildungsprojekten/-programmen erhalten. Der methodische Schwerpunkt liegt auf der Wirkungsmessung (*impact evaluation*). Zusätzlich sollen Effektivität und Nachhaltigkeitsmessungen berücksichtigt werden.

Fragenkatalog an Mitarbeiter*innen von Evaluationsabteilungen

Allgemeine Fragen zu (rigoroser) Wirkungsevaluation:

- 1) Wie werden in Ihrer Institution Wirkungen (Impacts) von Projekten/Programmen gemessen?
- 2) Haben Sie Erfahrungen mit rigorosen (oder evidenz-basierten) Wirkungsevaluationen? Was versteht Ihre Institution unter rigoroser Wirkungsevaluation (rigorous impact evaluation)? Wie definiert Ihre Institution rigorose Wirkungsevaluation?
 - a. Könnten Sie bitte Ihre aktuellste (interne) Definition oder Policy zu (rigoroser) Wirkungsevaluation mit uns teilen?
- 3) Werden rigorose Wirkungsevaluationsmethoden (rigorous impact evaluation methods) in Ihrer Institution angewendet?
 - a. Wenn ja, wie häufig?
 - b. Welche Methoden?
 - c. Gibt es gute Beispiele zu (rigorosen) Wirkungsevaluation aus dem TVET Sektor? Sind die Evaluationsberichte öffentlich zugänglich?
- 4) Welche methodischen Ansätze sind für die Messung von Wirkungen von Projekten/Programmen innerhalb ihrer Institution am effektivsten (oder nützlichsten)? Warum?
 - a. subjektive Einschätzung z.B. im Hinblick auf:
 - i. Steuerung von Projekten/Programmen und evidenzbasierte Entscheidungsfindung
 - ii. Lernen für zukünftige Projekte/Programme und
 - iii. Rechenschaftspflicht (*Accountability*)
- 5) Welche methodischen Ansätze zur Messung von Wirkungen von Projekten/Programmen haben sich besonders nützlich erwiesen, um geschlechterspezifische Wirkungen (und spezifische Wirkung für marginalisierte Gruppen, wie Jugendliche, Menschen mit Behinderung, Arme, Geflüchtete, etc.) zu erfassen?

Allgemeine Fragen zu Evaluationssystemen:

- 6) Wann werden Wirkungsevaluationen begonnen/geplant (z.B. Projektplanung, Projektstart, Beginn/Mitte/Ende der Projektimplementierung, nach Projektende)? Welche Art von Wirkungsanalysen?
- 7) Welcher budgetäre Rahmen steht für Wirkungsevaluationen ungefähr zur Verfügung?
- 8) Welchen Beitrag leisten Monitoring Systeme bei der Wirkungsmessung?
- 9) Welchen Beitrag leisten vorhandene Partnersysteme für die Wirkungsmessung (Statistiken, administrative Daten, Zensus, Surveys...)?
- 10) Welche Rolle haben die verschiedenen Projektbeteiligten bei der Planung und Durchführung von Wirkungsevaluationen?

- 11) Wie werden die Ergebnisse von Wirkungsevaluationen genutzt? (z.B. im Bereich Steuerung/evidenzbasierte Entscheidungsfindung, Lernen oder Rechenschaftspflicht)

(Optional) spezifische Fragen mit TVET-Bezug:

- 12) Was sind die wichtigsten Wirkungen von TVET Projekten/Programmen in Ihrer Institution?
- 13) Welche Konzepte wurden durch im Rahmen von Wirkungsevaluationen von TVET Projekten/Programmen geschärft? (*employability, skills improvement, capabilities, life skills, ...*)?
- a. Wie konzeptualisieren/definieren und operationalisieren/messen Sie die Wirkung (Impact) „Beschäftigungsfähigkeit (*employability*) auf dem Arbeitsmarkt“?
 - b. Wie konzeptualisieren/definieren und operationalisieren/messen Sie die Wirkung (Impact) „Einkommensveränderungen der Absolvent*innen“?
 - c. Wie konzeptualisieren/definieren und operationalisieren/messen Sie die Wirkung (Outcome) „Kompetenzverbesserung (*skills improvement*) der Absolvent*innen“?
- 14) Wie werden die Ergebnisse von Wirkungsevaluationen von Berufsbildungssystemen/-programmen genutzt oder übertragen? Welche weiteren Möglichkeiten sehen Sie? (z.B. Nutzung von Evaluationsergebnissen zur Projektplanung, Steuerung, evidenzbasierte Entscheidungsfindung oder Lernen)
- 15) Welche Konzepte und Methoden eignen sich speziell in Lateinamerika und der Karibik? Warum?

ANNEX 4 DEFINITION OF KEY CONCEPTS

CONTROL GROUP	A control group is an “untreated” research sample, which is constructed to study the counterfactual. The control group is not affected by the intervention and thus can be compared to the treatment group, which receives the intervention. The assignment to the control and treatment group is random, which ensures that those groups are similar on average (e.g. in age, income, education). The random assignment of control and treatment groups typically forms the basis for experimental designs.
EVALUABILITY	The extent to which an activity or project can be evaluated in a reliable and credible fashion (OECD/DAC). Assessment of how far the object of an evaluation (a measure, project, program, instrument, strategy or organization) can be evaluated in a reliable and plausible way. It requires an ex-ante appraisal to ascertain whether the objectives set have been appropriately defined and the results achieved can be verified.
EVALUATION	Systematic and objective analysis and evaluation of an ongoing or completed development measure. This investigation includes the conception, implementation and in particular the results of the development measure and should contain action-relevant findings and, in appropriate cases, recommendations for improvements to the design. Evaluation is sometimes understood as the process, sometimes as the result of analysis and evaluation.
EVIDENCE GAP MAP	A (web-based) interactive tool that provides an overview of and quick access to existing evidence on a topic or (sub-)sector of international cooperation. Evidence maps make gaps and focal points of the existing evidence base visually clear and can thus, for example, support decisions on where a systematic review is expedient.
EXPERIMENTAL DESIGN	Experimental impact evaluation designs include different variations of randomized controlled trials (RCTs). In an RCT, units of observation from a population of interest (e.g., households) are randomly assigned to two groups: (1) the treatment group, which experiences a development intervention, (2) the control group, which does not experience the intervention. The randomization approach ensures that both groups have similar characteristics on average. The difference in the outcome of interest (e.g., employability and employment) between the two groups after the intervention thereby represents the impact of the intervention (e.g. TVET participation). A common approach in field experiments is to randomly assign applicants for a training to treatment and control groups (see Chapter 4.1 CS 1 and CS 2).
IMPACT	Impact is part of the OECD/DAC evaluation criteria and it refers to significant positive or negative, intended or unintended higher-level effects of an intervention. The term applies to long-term and potentially transformative social, environmental and economic effects of an intervention. An intervention’s impact is broader in scope than the results under the effectiveness criterion because it aims to capture consequences beyond an intervention’s immediate results.
IMPACT EVALUATION (IE)	Impact evaluations (IEs) refers to evaluation designs that try to measure the causal effect of an intervention (e.g., a TVET program) on an observed variable of interest (e.g., skills improvement, employability or income). IEs should ideally allow the formation of robust conclusions about the impact caused by an intervention and therefore causal attribution. According to OECD-DAC, impact can be defined as a “positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended”. In the report, we use the superordinate term “IE” for the three main design options in IEs, namely experimental, quasi-experimental and non-experimental IE designs.
META-EVALUATION	Evaluation of one or more evaluations in order to assess quality based on a recognized standard e.g. OECD/DAC or DeGEval, based, comprehensible analysis grid.
NON-EXPERIMENTAL DESIGN	There are multiple non-experimental impact evaluation designs, which are less rigorous qualitative impact evaluation approaches. Most of these qualitative approaches are based on a theory, which states the results hypothesis or how project activities determine the outcomes and impacts of an intervention. In a theory-based impact evaluation, all steps and underlying assumptions in the causal chain linking activities and outcomes are spelled out and tested. A common theory-based non-experimental design is based on the contribution analysis (see Chapter 4.3 CS 7 and CS 8).
OECD/DAC EVALUATION CRITERIA	The internationally agreed OECD/DAC criteria should guide any evaluation and are important for the validity of an evaluation, learning and accountability of development cooperation evaluations. An evaluation matrix may be prepared which asks questions about each of the OECD/DAC criteria: relevance, coherence, effectiveness, efficiency, impact, sustainability.
OUTCOME	Outcomes measure the achieved short-term to medium-term changes and effects on beneficiaries, produced by the intervention outputs. Outcomes are linked to the effectiveness criterion of the OECD/DAC evaluation criteria.
OUTPUT	The outputs of an intervention refer to the immediate and concrete consequences of project activities. Concrete consequences are products, capital goods and services, which result from development interventions.

QUALITATIVE APPROACH	Qualitative approaches and methods analyse and explain what can be studied with words. Qualitative methods are applied to understand people's beliefs, experiences and attitudes, generating large amounts of non-numerical data. Semi-structured interviews, focus group discussions and case study research are common qualitative methods.
QUANITATIVE APPROACH	Quantitative approaches and methods measure and assess what can be studied with numbers. Quantitative research methods focus on describing characteristics of a population using structured approaches that provide precise data that can be statistically analysed. Questionnaires and survey containing closed-ended questions are common data collection instruments.
QUASI-EXPERIMENTAL DESIGN	Similar to experimental designs, quasi-experimental research designs test causal hypotheses by identifying a control group that is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics. The key difference between an experimental and quasi-experimental design is that the latter lacks random assignment, and the assignment often takes place ex-post the intervention. By measuring the variable of interest in both, the control and treatment groups, the control group states what the situation of the variable of interest (outcome, for the intervention group) would have been if the program or policy had not been implemented (i.e., the counterfactual). They can also often be applied when the intervention has already started, whereas RCTs must be prepared before the intervention has started. Quasi-experimental designs include for example different matching techniques (see Chapter 4.2 CS 6), difference-in-differences estimation (see Chapter 4.2 CS 3 and CS 4), and natural experiments (see Chapter 4.2 CS 5).
RESULT	According to the OECD/DAC terminology, the superordinate term results refers to outputs, outcomes and impacts of development interventions. Each of the three elements contributes to the next one.
RIGOROUS IMPACT EVALUATIONS (RIGOROUS IE)	Rigorous impact evaluation (IEs) include evaluation designs that measure the causal effect of an intervention based on counterfactuals. The narrow definition of rigorous IEs includes experimental and quasi-experimental evaluation designs that use control or comparison groups as the counterfactual situation to compare what happened due to the intervention and what would have happened without the interventions. This helps to identify if an intervention works and to analyze the causal impact of an intervention. According to the BMZ guidelines for evaluations, the rigorous IEs also require adequate assurance of the independence and quality of the investigation. The broader definition of IEs also includes non-experimental designs including for example contribution analyses.
SKILLS DEVELOPMENT	Skills Development refers to the productive capabilities acquired through all levels of learning and training, occurring in formal, non-formal, informal and on-the-job settings. It enables individuals to become fully and productively engaged in livelihoods, and to have the opportunity to adapt these capabilities to meet the changing demands and opportunities of economy and labour market. The acquisition of such capabilities depends on many factors, including a quality lifelong learning system and a supportive learning environment. The types of skills required for employment can be divided into: (I) Basic and foundation skills (acquired through the primary and secondary formal school system or through non-formal and/or informal learning processes; (II) Transferable skills (incl. the abilities to learn and adapt, solve problems, communicate ideas effectively, think critically and creatively and the ability to manage self and others); (III) Technical and vocational skills (specialized skills, knowledge or know-how to perform specific duties or tasks, mainly in a professional environment); and (IV) Professional and personal skills (incl. individual attributes relevant to work such as honesty, integrity, reliability, work ethic and judgement).
SYSTEMATIC REVIEW	Evaluation or aggregation of content-related findings from RIEs or impact analyzes according to a strict protocol of selection and aggregation along a study protocol, such as that of Cochrane or Campbell, or according to comparable standards.
TECHNICAL AND VOCATIONAL EDUCATION AND TRAINING (TVET)	Technical and Vocational Education and Training (TVET) is understood "as comprising education, training and skills development relating to a wide range of occupational fields, production, services and livelihoods". In this sense, it is "used as an equivalent term for vocational education and training (VET)". In the context of this study, the term "TVET" is used as an overarching concept to describe all kinds of formal and non-formal training and learning for work provided by public and private institutions, formal and informal providers (e.g., workshops) and learning locations.
TREATMENT GROUP	A treatment group is a "treated" research sample, which is constructed to study the counter-factual. In contrast to the control group, the treatment group is part of the intervention and affected by it. The assignment to the control and treatment group is random, which ensures that those groups are similar on average (e.g. in age, income, education). The random assignment of control and treatment groups typically forms the basis for experimental designs.

(Sources: [ADA 2020](#), [BetterEvaluation 2020](#), [DEval 2021](#), [DEval 2022d](#), [Duflo et al. 2008](#), [Gertler et al. 2016](#), [NCVER 2013](#), [OECD 2009](#), [SIDA 2018](#), [UNESCO 2015](#))

ANNEX 5 FURTHER LITERATURE

LITERATURE	INSTITUTION
ADA/OeEB (2019): <i>Evaluation policy of the Austrian development cooperation</i> , https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Evaluierung/Englisch/Evaluationpolicy.pdf [accessed 24.11.2022].	ADA/OeEB
ADA (2020): <i>Guidelines for Programme and Project Evaluations</i> , https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Evaluierung/Evaluierungs_Leitfaeden/Guidelines_for_Programme_and_Projekt_Evaluations_ADA_2020.pdf [accessed 24.11.2022].	ADA
ADA (2022a): Education, https://www.entwicklung.at/en/themes/education [accessed 24.11.2022].	ADA
ADA (2022b): <i>Support of agricultural cooperatives as an effective means to reduce poverty? Impact study on Austrian Development Cooperation (ADC)'s engagement from 2010 to 2020, with a focus on Armenia and Georgia</i> , https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Publikationen/Studien_u_Analysen/ADA_Impact_study_Agri_Coop_Final.pdf [accessed 24.11.2022].	ADA
ADA (2022c): <i>Evaluability Assessments in Austrian development cooperation. Guidance Document</i> , https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Evaluierung/GL_for_Evaluability_Assessments.pdf [accessed 24.11.2022].	ADA
Better Evaluation (2013): <i>Contribution Analysis</i> , https://www.betterevaluation.org/en/plan/approach/contribution_analysis [accessed 24.11.2022].	
Better Evaluation (2016a): <i>Realist Evaluation</i> , https://www.betterevaluation.org/en/approach/realist_evaluation [accessed 24.11.2022].	
Better Evaluation (2016b): <i>Process Tracing</i> , https://www.betterevaluation.org/en/evaluation-options/process-tracing#:~:text=Process%20tracing%20is%20a%20case,adjudicate%20between%20alternative%20possible%20explanations. [accessed 24.11.2022].	
BMZ (2021): <i>Guidelines for bilateral Financial and Technical Cooperation with cooperation partners of German development cooperation</i> , https://www.bmz.de/resource/blob/92794/7639a6b5542630243f506a36978faaa8/guidelines-for-bilateral-financial-and-technical-cooperation-data.pdf [accessed 24.11.2022].	BMZ
BMZ (2022): <i>BMZ responses to DEval evaluations</i> , https://www.bmz.de/de/ministerium/evaluierung/bmz-responses-19422 [accessed 24.11.2022].	BMZ
Center for Effective Global Action/University of Berkeley (2017): <i>Module 2.3: Randomized Promotion</i> , https://edge.edx.org/assets/courseware/v1/9ca9292205786060f477587c5373b35d/c4x/BerkeleyX/CEGA101AIE/asset/Module_2.3_Randomized_Promotion.pdf [accessed 24.11.2022].	CEGA
Davies, R.; Dart, S. (2005): <i>The 'Most Significant Change' (MSC) Technique A Guide to Its Use</i> , https://mande.co.uk/wp-content/uploads/2018/01/MSCGuide.pdf [accessed 24.11.2022].	
DeGEval (2016): <i>Standards für Evaluation</i> , https://www.degeval.org/fileadmin/Publikationen/DeGEval-Standards_fuer_Evaluation.pdf [accessed 24.11.2022].	DeGEval
DEval (2018): <i>STANDARDS FOR DEVAL EVALUATIONS. DEval Methods and Standards 2018</i> , https://www.deval.org/fileadmin/Redaktion/PDF/05-Publikationen/Policy_Briefs/2018_Methoden_Standards/DEval_Policy_Brief_Methods_Standards_2018_EN.pdf [accessed 24.11.2022].	DEval
DEval (2021): <i>RIGOROUS IMPACT EVALUATION: EVIDENCE GENERATION AND TAKE-UP IN GERMAN DEVELOPMENT COOPERATION. Research Report</i> , https://rie.deval.org/fileadmin/Redaktion/PDF/03_Methoden/RIE/DEval_Research_Report_2021_Rigorous_Impact_Evaluation_in_German_DC.pdf [accessed 24.11.2022].	DEval
DEval (2022a): <i>Rigorous Evidence Database</i> , https://rie.deval.org/rigorous-evidence-database [accessed 24.11.2022].	DEval
DEval (2022b): <i>Rigorous impact evaluation – funding programme</i> , https://rie.deval.org/funding-programme/the-rie-funding-programme/funding-programme [accessed 24.11.2022].	DEval
DEval (2022c): <i>Focelac+</i> , https://www.deval.org/en/evaluation-capacities/current-ecd-projects/focelac [accessed 24.11.2022].	DEval
DEval (2022d): <i>What is Rigorous Impact Evaluation (RIE)?</i> , https://rie.deval.org/what-is-rie/what-is-rigorous-impact-evaluation-rie [accessed 24.11.2022].	DEval

LITERATURE	INSTITUTION
DEval (2022e): <i>What is a Systematic Review (SR)?</i> , https://rie.deval.org/what-is-rie/what-is-a-systematic-review-sr [accessed 24.11.2022].	DEval
Duflo, E., Glennerster, R.; Kremer, M. (2008): <i>Using Randomization in Development Economics Research: A Toolkit</i> , ch. 61, p. 3895-3962 in Schultz, T. Paul and Strauss, John A. eds., Elsevier.	
EC (2022): Exchange rate (InforEuro), https://ec.europa.eu/info/funding-tenders/procedures-guidelines-tenders/information-contractors-and-beneficiaries/exchange-rate-inforeuro_en [accessed 24.11.2022].	EC
Gertler, P.; Martinez, S.; Premand, P.; Rawlings, L.; Vermeersch, C. (2016): <i>Impact Evaluation in Practice</i> , Second Edition. Washington, DC: Inter-American Development Bank and World Bank, https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464807794.pdf?sequence=2&isAllowed=y [accessed 24.11.2022].	IDB, World Bank
GIZ (2018): <i>GIZ's evaluation system. General description</i> , https://www.giz.de/en/downloads/GIZ_EVAL_EN_general%20description.pdf [accessed 24.11.2022].	GIZ
GIZ (2022a): <i>Technical and Vocational Education and Training (TVET)</i> , https://www.giz.de/expertise/html/60014.html [accessed 24.11.2022].	GIZ
GIZ (2022b): <i>Worldwide</i> , https://www.giz.de/en/html/worldwide.html [accessed 24.11.2022].	GIZ
Hudson, J., Fielding, S., Ramsay, C. (2019): Methodology and reporting characteristics of studies using interrupted time series design in healthcare. <i>BMC Med Res Methodol</i> 19, 137. https://doi.org/10.1186/s12874-019-0777-x	
KfW (2016): <i>Group-wide Impact Management</i> , https://www.kfw.de/nachhaltigkeit/About-KfW/Sustainability/Strategie-Management/Sustainable-Finance/Impact-management/ [accessed 24.11.2022].	KfW
KfW (2018): <i>Berufsbildung</i> , https://www.kfw-entwicklungsbank.de/PDF/Entwicklungsfinanzierung/Themen-NEU/Themen_aktuell_Bildung_Berufsbildung_DE_10_2018.pdf [accessed 24.11.2022].	KfW
KfW (2021a): <i>16th Evaluation Report 2019–2020. Evaluate. Measure. Learn.</i> https://www.kfw-entwicklungsbank.de/Bilder/Evaluierungsbericht-2021/Startseite/KfW-Evaluation-report_2019_2020.pdf [accessed 24.11.2022].	KfW
KfW (2022a): <i>Sustainability Guideline. Assessment and management of Environmental, Social, and Climate Aspects: Principles and Procedures</i> , https://www.kfw-entwicklungsbank.de/PDF/Download-Center/PDF-Dokumente-Richtlinien/Nachhaltigkeitsrichtlinie_EN.pdf [accessed 24.11.2022].	KfW
KfW (2022b): <i>Our principles</i> , https://www.kfw-entwicklungsbank.de/International-financing/KfW-Development-Bank/Evaluations/Principles/ [accessed 24.11.2022].	KfW
KfW (2022c): <i>Our results</i> , https://www.kfw-entwicklungsbank.de/International-financing/KfW-Development-Bank/Our-results/ [accessed 24.11.2022].	KfW
KfW (2022d): <i>Education</i> , https://www.kfw-entwicklungsbank.de/International-financing/KfW-Development-Bank/Topics/Education/ [accessed 24.11.2022].	KfW
Kirkpatrick Partner (2022): <i>What is the Kirkpatrick Model?</i> , https://www.kirkpatrickpartners.com/the-kirkpatrick-model/ [accessed 24.11.2022].	
Neubert, S. (2010): <i>Description and Examples of MAPP</i> , Method for Impact Assessment of Programmes and Projects (MAPP) (ngo-ideas.net) [accessed 24.11.2022].	IDOS
OECD (2009): <i>Glossary of key terms in evaluation and results based management</i> , https://www.oecd.org/dac/evaluation/glossaryofkeytermsinevaluationandresultsbasedmanagement.htm [accessed 24.11.2022].	OECD
OECD (2021): <i>Evaluation Criteria</i> , https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm [accessed 24.11.2022].	OECD
Outcome Mapping Learning Community (2022): <i>What is Outcome Mapping?</i> , https://www.outcome-mapping.ca/about/om [accessed 25.11.2022].	
Sammeth, F.; Lakoh, A.; Michel, B.; Hites, G. (2010): <i>Impact of rural poverty reduction strategies: The case of smallholders in Sierra Leone</i> , https://www.researchgate.net/figure/OECD-DAC-Evaluation-criteria-and-the-Logical-Framework-Approach_fig44_254383817 [accessed 24.11.2022].	

LITERATURE	INSTITUTION
SDC (2016): <i>The SDC portfolio in Vocational Skills Development (VSD) – Key figures</i> , https://www.shareweb.ch/site/El/Documents/To%20SORT/Themes/VSD/SDC%20-%20Report%20-%20Analysis%20of%20the%20SDCs%20VSD%20Portfolio%202016%20-%202017(en).pdf [accessed 24.11.2022].	SDC
SDC (2018): <i>Evaluation Policy Swiss Agency for Development and Cooperation SDC</i> , https://www.eda.admin.ch/content/dam/deza/en/documents/resultate-wirkung/20180906-evaluationspolitik-maerz-2018_EN.pdf [accessed 24.11.2022].	SDC
SDC (2019): <i>Independent Evaluation of SDC's Performance in Social Protection 2013 - 2017, Evaluation 2019/2</i> , https://www.shareweb.ch/site/Poverty-Wellbeing/social-protection/Documents/social-protection_EN.pdf [accessed 24.11.2022].	SDC
SDC (2020a): <i>Switzerland's International Cooperation Strategy 2021-24</i> , https://www.eda.admin.ch/content/dam/deza/en/documents/publikationen/Diverses/Broschuere_Strategie_IZA_Web_EN.pdf [accessed 24.11.2022].	SDC
SDC (2020b): <i>Evaluationsberichte</i> , https://www.eda.admin.ch/deza/de/home/wirkung/berichte/evaluationsberichte.html [accessed 24.11.2022].	SDC
SDC (2022): <i>Organigramm SDC</i> , https://www.eda.admin.ch/content/dam/deza/en/documents/die-deza-organigramm_EN.pdf [accessed 24.11.2022].	SDC
SECO (2017): <i>Evaluation Guidelines</i> , https://www.seco-cooperation.admin.ch/dam/secocoop/de/dokumente/resultate/evaluation/eval-policy.pdf.download.pdf/Evaluation%20Policy.pdf [accessed 24.11.2022].	SDC
SECO (2021): <i>Evaluation Policy</i> , https://www.seco-cooperation.admin.ch/dam/secocoop/de/dokumente/resultate/evaluation/eval-policy.pdf.download.pdf/Evaluation Policy.pdf [accessed 24.11.2022].	SECO
SECO (2022): <i>Evaluation</i> , https://www.seco-cooperation.admin.ch/secocoop/en/home/results/evaluation.html [accessed 24.11.2022].	SECO
UN (2022): <i>Global Partnership for Effective Development Cooperation</i> , https://effectivecooperation.org/landing-page/about-partnership [accessed 24.11.2022].	UN
Waddington, H.; White, H.; Snilstveit, B.; Garcia H. (2012): <i>How to do a good systematic review of effects in international development: A tool kit</i> . Journal of Development Effectiveness. 4. 359-387.	
White, H., Sabarwal, S.; de Hoop, T. (2014): <i>Randomized Controlled Trials (RCTs), Methodological Briefs: Impact Evaluation 7</i> , UNICEF Office of Research, Florence, https://www.unicef-irc.org/publications/pdf/brief_7_randomized_controlled_trials_eng.pdf [accessed 24.11.2022].	UNICEF
World Bank (2022a): <i>Difference-in-Differences</i> , https://dimewiki.worldbank.org/Difference-in-Differences [accessed 24.11.2022].	World Bank
World Bank (2022b): <i>Matching</i> , https://dimewiki.worldbank.org/Matching [accessed 24.11.2022].	World Bank
World Bank (2022c): <i>Regression Discontinuity</i> , https://dimewiki.worldbank.org/Regression_Discontinuity [accessed 24.11.2022].	World Bank
World Bank (2022d): <i>Instrumental Variable</i> , https://dimewiki.worldbank.org/Instrumental_Variables#:~:text=and%20weak%20instruments.,Overview,measurement%20error%2C%20or%20simultaneity . [accessed 24.11.2022].	World Bank

Websites and Databases

NAME AND LINK	INSITUATION
Bridging Innovation and Learning in TVET Library https://unevoc.unesco.org/bilt/BILT+Library	UNESCO-UNEVOC International Centre for Technical and Vocational Education and Training UN Campus
Datenbank zur internationalen Berufsbildungszusammenarbeit (Go VET) https://www.govet.international/de/2358.php	Bundesinstitut für Berufsbildung
Donor Committee for dual Vocational Education and Training https://www.dcdualvet.org/en/	DCDUALVET
Donor Committee for Enterprise Development (DCED) https://www.enterprise-development.org/what-works-and-why/evaluations-of-agency-psd-work/	OECD
3ie Development Evidence Portal (DEP) https://developmentevidence.3ieimpact.org/	International Initiative for Impact Evaluation
Economy + Education Network https://www.shareweb.ch/site/EI	SDC
European Training Foundation https://www.etf.europa.eu/en	ETF
EvalParticipativa https://evalparticipativa.net/en/	
Evaluation Reports Database https://luxdev.lu/en/documents/section/eval	Luxemburg Development Cooperation Agency
FOCEVAL+ http://foceval.org/documentos-y-enlaces/	BMZ/DEval
GIZ Evaluation database https://www.giz.de/en/aboutgiz/516.html	GIZ
Impact Evaluation: Resources https://www.iadb.org/en/topics-effectiveness-improving-lives/impact-evaluation	IDB
IZA Institute of Labor Economics https://www.iza.org/en	IZA
KfW Evaluation database https://www.kfw-entwicklungsbank.de/Evaluierungsbericht/Evaluierungen/index-2.html	KfW
OECD DAC Evaluation Resource Centre (DEReC) https://www.oecd.org/derec/	OECD
Randomized evaluation database https://www.povertyactionlab.org/evaluations	J-PAL
Red de Seguimiento, Evaluación y Sistematización de Latinoamérica y El Caribe, y su acrónimo https://www.relac.net/	
Rigorous Evidence Database https://rie.deval.org/rigorous-evidence-database	DEval
Skills, knowledge and employability http://www.ilo.org/global/topics/skills-knowledge-and-employability/lang--en/index.htm	ILO
VET Toolbox Coordination HUB https://www.vettoolbox.eu/en/knowledge	

Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Sitz der Gesellschaft / [Registered offices](#)
Bonn und Eschborn / [Bonn and Eschborn](#)

Friedrich-Ebert-Allee 32 + 36
53113 Bonn, Deutschland / [Germany](#)
T +49 228 44 60-0
F +49 228 44 60-17 66

E info@giz.de
I www.giz.de

Dag-Hammarskjöld-Weg 1 - 5
65760 Eschborn, Deutschland / [Germany](#)
T +49 61 96 79-0
F +49 61 96 79-11 15

Im Auftrag des



Bundesministerium für
wirtschaftliche Zusammenarbeit
und Entwicklung