Federal Ministry
for Economic Cooperation
and Development

Impact Evaluation Guidebook for Climate Change Adaptation Projects

Published by giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

In cooperation with UNDP Empowered lives. Resilient nations.

Centrum für Evaluation CEval Center for Evaluation

Content

List of tables

List of figures

List of abbreviations

AAP	Africa Adaptation Programme
ADB	Asian Development Bank
BGR	Federal Institute for Geosciences and Natural Resources
BMZ	German Federal Ministry for Economic Cooperation and Development
BMUB	German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety
CCA	Climate change adaptation
CIDA	Canadian International Development Agency
CO2	Carbon dioxide
DAC	Development Assistance Committee
DfID	Department for International Development
EU	European Union
GCF	Green Climate Fund
GHG	Greenhouse gases
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH
IFAD	International Fund for Agricultural Development
KfW	Kreditanstalt für Wiederaufbau
LDC	Least developed county
LDCF	Least Developed Countries Fund
M&E	Monitoring and Evaluation
NAMAs	Nationally Appropriate Mitigation Actions
NAP	National Adaptation Plan
NC	National Communication
NGO	Non-governmental organisation
OECD	Organisation for Economic Co-operation and Development
PSM	Propensity score matching
RDD	Regression discontinuity design
RCT	Randomised controlled trial
REDD	Reducing Emissions from Deforestation and Forest Degradation
SEM	Structural equation model(ling)
SIDA	Swedish International Development Agency
ToC	Theory of change
UNDP	United Nations Development Programme
UNEP	United Nations Environment Programme
UNFCCC	United Nations Framework Convention on Climate Change
UNICEF	United Nations Children's Fund
UNIDO	United Nations Industrial Development Organization
USAID	United States Agency for International Development
WFP	World Food Programme

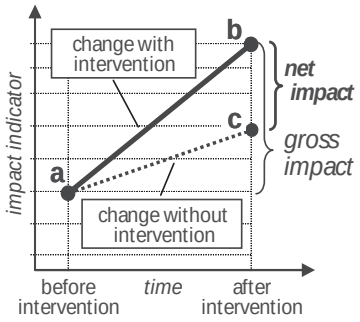
Glossary

Accountability	Obligation to demonstrate that work has been conducted in compliance with agreed rules and standards or to report fairly and accurately on performance results vis-à-vis mandated roles and/or plans. This may require a careful demonstration that the work is consistent with the contract terms.
Adaptation	Adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities. Various types of adaptation can be distinguished, including anticipatory, autonomous and planned adaptation: **Anticipatory adaptation** – Adaptation that takes place before impacts of climate change are observed. Also referred to as proactive adaptation. **Autonomous adaptation** – Adaptation that does not constitute a conscious response to climatic stimuli but is triggered by ecological changes in natural systems and by market or welfare changes in human systems. Also referred to as spontaneous adaptation. **Planned adaptation** – Adaptation that is the result of a deliberate policy decision, based on an awareness that conditions have changed or are about to change and that action is required to return to, maintain, or achieve a desired state.
Adaptive capacity	The ability of a system to adjust to climate change (including climate variability and extremes) to moderate potential damages, to take advantage of opportunities, or to cope with the consequences.
Attribution	Ascribing the cause of an effect (change) to a specific intervention. An approach to (➔) rigorous impact evaluation that concerns clear cause-effect relationships, i.e. establishing causal links and drawing explanatory conclusions about observed changes (whether anticipated or not) and the concerned intervention. Focusing on clear causation implies considering the (➔) counterfactual to assess the (➔) net impact of an intervention by a comparison of what has occurred with the intervention implemented (the ➔ factual) with the situation without the intervention (the ➔ counterfactual).
Climate change	Climate change refers to any change in climate over time, whether due to natural variability or as a result of human activity. This usage differs from that in the United Nations Framework Convention on Climate Change (UNFCCC), which defines 'climate change' as: 'a change of climate which is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and which is in addition to natural climate variability observed over comparable time periods'.
Counterfactual	The situation that would have happened if the intervention had not taken place ('without' situation). As the counterfactual is not directly observable, the unobservable potential outcome of the counterfactual situation is estimated via the situation of an equivalent control/ comparison group that is not affected by the intervention. This is done using (➔) randomised controlled trials (RCT) or (➔) quasi-experimental designs.
Evaluability	Extent to which an activity or a programme can be evaluated in a reliable and credible fashion.
Experimental design	A study in which individuals (or groups of individuals) are randomly allocated (by chance alone) to an intervention group (treatment) or a control group (not being part of the intervention). Experimental studies (➔ RCTs) seek to measure an intervention's (➔) net impact by comparing the two groups' situations before and after the intervention (➔ attribution).
Exposure	The nature and degree to which a system is exposed to significant climatic variations.
Factual	The observed outcome of an intervention ('with' situation). (➔ counterfactual)
Impact[1]	Positive and negative, primary and secondary long-term effects produced by an intervention, directly or indirectly, intended or unintended (➔ net impact)
Impact evaluation	An evaluation that looks beyond the immediate results of an intervention, project, programme or policy. Impact evaluations may focus (i) on higher outcomes rather than intervention outputs, (ii) on longer-term results, (iii) on a group of interventions within a given sector or geographical area, or (iv) explicitly on the impacts produced by an intervention, i.e. attributing impacts to an intervention (➔ Rigorous impact evaluation)
Monitoring	A continuing function that uses systematic collection of data on specified indicators to provide management and the main stakeholders of an ongoing development intervention with indications of the extent of progress and achievement of objectives and progress in the use of allocated funds.

[1] For GIZ colleagues, a brief explanation of how this impact definition fits into GIZ's results model (GIZ 2013): The results model is an expression of GIZ's Managing for Development Results approach. The results model forms the detailed basis for GIZ's subsequent operational planning and for establishing the results-based monitoring system at the start of implementation. Within this model GIZ understands results as the 'intended or unintended, positive or negative changes in a situation or behaviour as the direct or indirect consequence of an intervention'. Results include impacts, outcomes and outputs. Impacts are defined as 'the long-term, overarching development results. They are usually located at the level of the development cooperation programme objective'. This definition largely corresponds to the one used in this document. However, at some points in this publication, the authors use the term 'impact' synonymously with the term 'outcome' when applied to GIZ's results model. Within GIZ, outcomes are defined as the 'expected or actually achieved direct short-term and medium-term results of a measure's outputs'.

Net impact

Effects exclusively produced by an intervention (without effects caused by other possible external causes); effects that can be (➔) attributed to a specific intervention by clear (➔) causation.
The net impact (b-c) is the difference between what has occurred with the intervention implemented (b-a), the (➔)
factual, and the situation without the intervention (c-a), the (➔) counterfactual.



Technically speaking, the net impact is defined as follows: an effect δ caused by a treatment T (intervention) is the difference between the outcome Y_1 under a treatment T (T=1) minus the alternative outcome Y_0 that would have happened without the treatment T (T=0):
$\delta = Y(X, T = 1) - Y(X, T = 0) = Y_1 - Y_0$

Partial least squares analysis

A multivariate statistical approach for the estimation of causal relationships. It is a variance-based, non-linear (i.e. no assumptions regarding the value distribution of a variable required) iterative method based on a linear regression model. It allows the estimation of concrete values for latent (i.e. non-observable/measurable) constructs with the help of manifest (i.e. observable/measurable) indicators.

Quality (rigorous) impact evaluations

As there is (i) a mistaken belief that (➔) rigorous impact evaluations have to use (➔) randomised controlled trials (RCTs) on the one side, and (ii) the fact that most rigorous impact evaluations are being carried out by economists and econometricians (not evaluators) resulting in technically competent studies but not automatically very good evaluations on the other side, more and more the term 'quality' impact evaluation is used to avoid misunderstanding and to focus on quality.

Quasi-experimental designs

An attempt to uncover a causal (➔) attribution, even though a random pre-selection processes is not possible. A quasi-experimental design is a type of experimental design (➔ RCT) where a randomised control group could not be established for either ethical or practical reasons, and therefore the intervention group's situation is compared with those of a similar group of individuals not receiving intervention activities (comparison group).

Randomised controlled trial (RCT)

An (➔) experimental study in which individuals (or groups of individuals) are randomly allocated (by chance alone) to an intervention group (treatment) or a control group (not being part of the intervention). RCTs seek to measure an intervention's (➔) net impact by comparing the two groups' situations before and after the intervention (➔ attribution).

Resilience

The ability of a social or ecological system to absorb disturbances while retaining the same basic structure and ways of functioning, the capacity for self-organisation, and the capacity to adapt to stress and change.

Rigorous impact evaluation

An (➔) impact evaluation with strong emphasis on 'produced by', focusing on clear causation/causal attribution by establishing the counterfactual: Assessing the effects produced by an intervention, the (➔) net impact, requires a comparison of what has occurred with the intervention implemented – i.e. the (➔) factual – with the situation without the intervention – i.e. the (➔) counterfactual. The term 'rigorous' is added to differentiate this approach from many – more traditional – (➔) impact evaluation approaches.

'Rigorised' impact evaluation

A (➔) rigorous impact evaluation conducted in a real-life situation where necessary conditions are not (fully) given and which is therefore conducted as rigorously as is feasible.

Sensitivity

Sensitivity is the degree to which a system is affected, either adversely or beneficially, by climate variability or change. The effect may be direct (e.g., a change in crop yield in response to a change in the mean, range or variability of temperature) or indirect (e.g., damages caused by an increase in the frequency of coastal flooding due to a rise in sea level).

Vulnerability

Vulnerability is the degree to which a system is susceptible to, and unable to cope with, adverse effects of climate change, including climate variability and extremes. Vulnerability is a function of the character, magnitude, and rate of climate change and variation to which a system is exposed, its sensitivity, and its adaptive capacity.

Executive Summary

Monitoring and evaluating impacts is usually both costly and laborious. Often, it is also a challenging process, particularly when complex causal linkages or uncertain framework conditions are involved. All of this applies to climate change adaptation (CCA) projects, which present further methodological and practical challenges that complicate the assessment of concrete adaptation results. However, providing evidence about the impact of an intervention is often indispensable when it comes to generating knowledge about what works and what doesn't, initiating organisational learning processes, monitoring the progress made or simply being accountable for the use of resources. **This Guidebook seeks to support project managers by providing an overview of different impact evaluation methods and how they can be applied to climate change adaptation projects.** The application of the Guidebook is further illustrated by a case study of an adaptation project in Bangladesh.

Evaluations and rigorous impact evaluations – What's the difference?

While many evaluation approaches claim to provide some kind of indication about project impacts (e.g. most significant change approach, participatory rapid assessment), robust evidence can only be provided by sophisticated evaluation designs that comply with scientific standards and are based on valid empirical data. Evaluations that use such sophisticated designs and involve the collection of a considerable amount of empirical data are also called **rigorous impact evaluations (RIEs).** The advantage of RIEs in contrast to 'softer' designs is that they make it possible to clearly **attribute observed changes** to a particular intervention or at least make it possible to **quantify the contribution** an intervention has made to these changes. This Guidebook provides practitioners in the field of CCA with a selection of such RIE designs, differentiated according to the type of impact (i.e. micro, meso or macro-level impacts) they are able to measure.

> Rigorous impact evaluations (RIEs) make it possible to **attribute observed changes** to a particular intervention or at least **quantify the contribution** an intervention has made to these changes.

The difficulties involved in measuring the impacts of CCA projects are widely recognised by practitioners and the adaptation community (e.g. Bours et al. 2014). The latter acknowledges the limited evidence on global adaptation progress and the gaps in the evolving adaptation science (Ford/Berrang-Ford 2015). For climate change mitigation, progress can be tracked with reference to the global concentration of greenhouse gases (GHGs). For climate change adaptation, though, there is no such single metric. This explains the need to equip practitioners with the right methods to assess adaptation results and demonstrate the added value of their projects.

Establishing causality on different levels

According to the OECD's Development Assistance Committee (DAC), impact evaluations take into account intended and unintended, positive and negative as well as expected and unexpected changes. They are supposed to not only provide information on all possible changes that have occurred during the implementation of an intervention but also to link these observed changes to their causes. Therefore, the **establishment of causality** (cause/effect relationship) is crucial in order to understand why particular incidents occurred during and after a project or programme. In particular, the question 'What would have happened without the project or pro-

Photo: © GIZ/Guenay Ulutuncok

gramme?' needs to be investigated in detail. At this point, a so-called **counterfactual assessment** (i.e. the comparison between what actually happened and what would have happened in the absence of the intervention) must be considered. Since it is impossible to collect data on what would have happened if the intervention had not been implemented, many RIEs are based on designs that include comparison data, i.e. on individuals in a similar situation who were not exposed to the intervention.

The RIE designs presented in this Impact Evaluation Guidebook are differentiated according to the **type of impact** they are able to measure, i.e.

- micro (individual),
- meso (institutional) or
- macro (systemic)-level impacts.

The aim of the Guidebook is to address the challenge of producing more and better impact evaluations, giving practitioners the necessary know-how in order to plan, implement and steer an RIE. It enables practitioners to identify which of the RIE designs are suitable for successfully evaluating their particular project or programme, and also reveals the respective potentials and limitations of the different designs when it comes to their application in the field of CCA.

The benefit of implementing RIEs for CCA projects

Why are RIEs increasingly being considered as an option for evaluating a CCA project? The Paris Declaration on Aid Effectiveness, which calls for more and better impact evaluations, has created a certain amount of pressure and consequently a rising demand to use RIEs. This also explains why more funding is becoming available and the literature is growing in this area. During the research for this Guidebook, one crucial step was to understand how CCA projects were evaluated in the past and how they are currently evaluated. So far, RIEs have rarely been undertaken for CCA projects. Additionally, respected independent evaluation departments have seldom implemented RIEs with counterfactuals for CCA projects up to now. Since there is limited evidence on global adaptation progress to date, RIEs have the capacity to redress the situation and generate learning. This is because they are conducted according to scientific quality standards that

i. ensure the correctness of measurement findings (reliability),
ii. exclude as far as possible alternative causes for an observed finding (internal validity) and
iii. shift findings from an observed sample to a larger population (external validity).

No 'one size fits all' RIE design

The numerous **(methodological) challenges** faced by CCA **projects** have led to intensive discussion among experts on how to adequately monitor and evaluate them. The main challenges are

1. the lack of a conceptual agreement on definitions, including what actually constitutes successful adaptation and therefore
2. the non-existence of a universal indicator (unlike for mitigation);
3. changes in the climatic context during a project's lifetime and the related problem of shifting baselines;
4. uncertainty about actual climate change patterns and their effects and
5. the long-time horizon of potential climate change impacts (see 2.2 for more details).

Individually, none of these challenges are unique to CCA; they also exist in projects in other sectors. Together, though, they represent a fundamental difficulty for practitioners in terms of monitoring and evaluating project impacts. Since adaptation projects vary widely in their scope and in the sectors they cover, there is no such thing as a 'one size fits all' evaluation design. Whether and which RIE design(s) are to be used depends on various factors such as the **level** (micro, meso or macro) at which the project generates an impact, the availability of **data** and the **time** and **resources** available to the project. For example, at the systemic (macro) level, a climate policy that is still at an early stage of implementation and project activities that have so far mainly focused on consultations and capacity building exercises may not be appropriate for a counterfactual assessment. Nonetheless, most CCA projects follow a multi-level approach, which implies that a combination of two or more evaluation approaches may be required to measure the total impact of a project or programme.

The selection of evaluation designs presented in this Guidebook is based on two considerations. Firstly, the design needs to comply with scientific standards in order to produce robust findings. Secondly, the selected designs should cover a wide range of different project approaches and all possible impact levels. In particular, six types of designs are discussed in detail:

- experimental and quasi-experimental designs,
- matching techniques,

Photo: © GIZ/Steinberg

- pipeline approach,
- regression discontinuity design (RDD),
- time-series designs and
- structural equation modelling (SEM).

Which evaluation design to choose – creating evidence at the individual level

At the individual level, if a baseline is available (ex-ante and ex-post data from both the treatment and control group), three evaluation designs can be considered for evaluating the CCA project: an **experimental**, **quasi-experimental** or **regression discontinuity design**.

In an **experimental design**, also known as a **randomised controlled trial (RCT)**, a treatment and control group are compared with each other at two points in time – before and after the intervention – in order to estimate the counterfactual (what would have happened in the absence of the intervention). To be classified as an 'experimental' design, the beneficiaries need to be randomly selected during the planning phase of a project, i.e. each person has the same chance of being a member of either the control or the treatment group. If a random selection cannot be realised (e.g. due to self-selection bias or conscious selection), then such a design is called **quasi-experimental**. In that case the group without treatment is called a 'comparison group'. This means that some kind of matching technique needs to be applied during data analysis in order to control for selection bias.

Thus, both experimental and quasi-experimental designs are applicable to CCA projects as long as project activities aim at creating a direct impact at individual (micro) level. The use of an experimental or quasi-experimental design calls for technical, logistical and financial resources that mainly depend on the size and accessibility of the target group. The data selection process is often quite costly since a large team of enumerators is required. Moreover, there are two potential biases that may emerge and hamper the process of attributing the observed change to project activities: spill-over and contagion effects. The former implies that the control/comparison group indirectly receives a benefit from the project or programme, and the latter occurs when different factors (e.g. projects operated by different agencies) affect the situation of the target population. Nonetheless, in terms of study design, RCTs show the highest internal validity[2] and enable the clear attribution of interventions to impacts (the latter is also true of the quasi-experimental and pipeline approach). **Regression discontinuity design (RDD)** is a quasi-experimental evaluation method that can be applied if beneficiaries of the project are selected based on a special characteristic (e.g. income) of relevance to the desired impact that distinguishes them from the non-beneficiaries. At the same time, beneficiaries are compared with the comparison group in a number of other respects (e.g. location). Being a quasi-experimental design, RDD does not require randomisation.

If no baseline data have been collected beforehand, two further evaluation designs can be considered at the individual level: the **pipeline approach** and the **panel design**.

The application of a **pipeline approach** is possible if a project is implemented sequentially (e.g. in different regions). Thereby, the selection of the target population that is not yet affected can serve as a quasi-comparison group for the group that is already affected. It is important, though, that the treated groups of each phase are comparable. Each data collection phase is just as labour-intensive and time-consuming as a quasi-experimental approach. **Panel designs**, a type of time-series design, are in principle applicable for measuring impacts on all levels, but require comparably larger sample sizes in order to be able to perform the necessary statistical calculations. The special characteristic that differentiates the panel design from all the others mentioned above is that it requires data to be collected from identical units (e.g. persons, households) at each point in time. This makes it prone to attrition effects (i.e. decreasing sample sizes) over time. It should also be mentioned that a panel design only makes it possible to estimate the contribution made by an intervention to observed changes but not to attribute these changes unambiguously.

[2] Internal validity is defined as the extent to which the variation of a dependent variable can be explained by the variation of a specific independent variable; i.e. the extent to which alternative explanations for the variation of the dependent variable can be excluded.

Creating evidence at institutional and system level

If an intervention aims to generate an impact at the **institutional** and system **levels**, two further evaluation designs can be taken into account: **time-series designs** and **structural equation modelling**.

Time-series designs can be of particular use when it is necessary to cover longer periods than those set by a project or programme. This makes them highly suitable for CCA projects, where changes or results may unfold at later stages. If data were collected repeatedly over a longer time period during the project, time-series designs might be the best option. It should be borne in mind, though, that these designs require a large number of repeated observations (at least about 10) in order to be able to perform the necessary statistical analyses for providing robust findings. Finally, if none of the designs fit, **structural equation modelling (SEM)** could be an option for providing meaningful evidence about the impact of a project or programme. It may be applied to measure large-scale policy-based programmes that aim to affect an entire sector, country or region. SEM makes it possible to measure the statistical relationship between several influencing factors (e.g. public investments, project funding, disaster resilience of a population). This can be used to estimate the contribution of each of these factors, one of which may be an intervention. It should be pointed out that SEM and time-series designs only serve to estimate the contribution made by an intervention to observed changes.

Depending on the reader's objective, the different sections of the Guidebook may be of more or less interest. However, in order to obtain a comprehensive understanding of the subject matter, we strongly recommended that you go through the **'must reads'**: Section 1.2, which provides practical guidance for the planning and implementation of an RIE; the summary at the end of Section 3 that contains an overview of the pros and cons of the different evaluation designs and leads you through the process of deciding which design to choose; and Section 3.2 that focuses on the requirements for collecting large-scale quantitative data. Once a particular evaluation design has been chosen, the reader may consult the respective section that describes the chosen design in more detail (Sections 3.1.1 to 3.1.6), and, as the case may be, the further explanations in the annex. If any of the technical terms or concepts are unclear, the reader can consult the glossary. M&E specialists who are not so familiar with the characteristics of CCA projects are advised to consult Section 2, in particular Section 2.1, which discusses the key challenges of CCA projects with regard to their evaluation. Finally, the case study of an adaptation project in Bangladesh illustrates the application of this Guidebook.

Conclusion and reading recommendation

The overview of available evaluation designs shows that there are several options for measuring impacts, each with different potentials and limitations as well as methodological and data requirements. This Impact Evaluation Guidebook provides practitioners with guidance on selecting the appropriate approach for a particular CCA project, based on its characteristics and the available resources. It also contains references to

- definitions and conclusions **(blue framed boxes)**,
- CCA project characteristics **(green framed boxes)**,
- practical tools and instruments **(purple framed boxes)**,
- further reading material **(orange framed boxes)**, and
- checklists summarising key issues **(red framed boxes)**.

Photo: © GIZ/Ranak Martin

1 Introduction

Are you planning to provide meaningful evidence of the impact of your climate change adaptation (CCA) project, based on a sound empirical foundation? Do you want to find out what difference it makes, what changes can be observed in the field of intervention and how these changes are related to the project activities? Or do you aim to gain knowledge about the inherent cause-and-effect relationships of a project, learn from the past for the future, further develop the project design or be accountable for the use of funds? The decision about **how to evaluate a project** depends on a number of aspects such as the objectives of the evaluation, its target groups and stakeholders or the available financial, technical, human and time resources. So the question 'how' is closely related to several further aspects, e.g. which design to implement, which data collection instruments to apply or which analytical methods to choose.

In this Guidebook we assume that you have decided to implement a **rigorous impact evaluation (RIE)**. As the name says, the focus of an RIE lies on the measurement and assessment of the impacts of a programme or project. According to the OECD/DAC evaluation criteria, impact means any positive and negative, primary and secondary long-term effect produced by an intervention, directly or indirectly, intended or unintended (OECD/DAC 2006: 6). Thus, basically any **observable change** occurring in – and potentially beyond – a field of intervention presumably related to the evaluated programme or project may be subject to an RIE. Another main characteristic of an RIE is that it aims to clearly **attribute** these observable changes to the intervention being examined or at least to provide sufficient evidence to **estimate the contribution** the intervention has made to these changes. Finally, the term 'rigorous' implies that the measurement and attribution, or contribution analysis, is conducted in line with scientific quality standards. This means that the findings of an RIE must meet the methodological demands that apply to scientific research projects (i.e. that they should be reliable, internally and externally valid, and objective.[3] These requirements – the rigorous measurement of impacts and attribution/contribution analysis – determine the options for choosing a specific design, or specific instruments and methods.

As in every scientific research project, the selection of the design further depends on the characteristics of the research subject. With regard to impact evaluation, the main characteristic to be considered is **'what kind of impact'** needs to be evaluated. While impact can be differentiated according to various category schemes (e.g. social, political, environmental impacts), the most decisive aspect in terms of evaluation design is the 'aggregation level' at which it occurs. This applies to climate change adaptation just as to any other field of intervention. A distinction is usually made between three aggregation levels:

- **Individual (micro-level) impacts**, i.e. benefits or drawbacks for particular target groups (e.g. improving the resilience of a population to climate change effects)
- **Institutional (meso-level) impacts**, i.e. changing the resources, capacities, performance, etc. of organisations such as enterprises, governmental or non-governmental institutions (e.g. strengthening the capacities of local authorities to deal with the socio-economic consequences of climate change)
- **Systemic (macro-level) impacts**, i.e. sectoral or regional developments (e.g. increasing the efforts of a government to reduce climate change-related causes of poverty or ill health)

The different accessibility of data at these levels implies that there is no 'one size fits all' design. For example, while comparative analysis, such as is applied in a quasi-experimental design (cf. Section 3.1.1), may be suitable for measuring and assessing the livelihoods of climate migrants in a specified intervention area, the influence on the capacities of basic urban service providers may call for the use of a structural equation modelling approach (cf. Section 3.1.6). In the latter case, there are many competing influences, which makes it impossible to establish a comparison group (see Section 4 for an in-depth

[3] Reliability, internal and external validity, and objectivity are counted among the most essential scientific quality criteria. In empirical studies, reliability indicates the correctness of measurement findings. Internal validity means that the study design excludes alternative causes for an observed effect as far as possible, while external validity means that the findings are in principle transferable from an observed sample to a greater population i.e. can be generalised. Finally, objectivity describes the independence of the measurement findings from their framework conditions (i.e. who collects the data, who analyses them and under which conditions the findings are interpreted).

discussion of this example). Given the complexity of climate change adaptation needs, many CCA projects follow a multi-level approach. Thus, in an evaluation it may be necessary to combine several designs in order to assess the total effectiveness of such a project. The Guidebook therefore aims to provide an operationally viable framework for evaluating CCA projects.

According to this objective, the structure of the Guidebook reflects the points to be considered when planning an RIE, starting with a general overview of how to plan such an exercise (Section 1.2). This is followed by a discussion of the types and key features of CCA projects, the typical challenges in the field of CCA and the way they are currently evaluated (Section 2). Section 3 presents the different options for rigorously evaluating CCA projects, including references to the CCA-specific requirements and challenges outlined above. The methodological prerequisites for generating reliable empirical data on a large scale are also discussed. Finally, two exemplary evaluation designs are outlined on the basis of a GIZ case study (Section 4), including a brief presentation of their practical implementation. The Guidebook closes with an annex containing further introductions to some methodological aspects that need to be considered when designing an RIE, a glossary of technical terms and further literature references.

1.1 How to use the Guidebook

As the Guidebook aims to provide 'hands-on' information, it contains a number of hints, definitions, practical examples, links to further reading material and checklists, which are highlighted as follows:

Definitions and **conclusions** are presented in blue boxes followed by an exclamation mark.

References to CCA projects are presented in green text boxes marked with a magnifying glass.

Practical tools and **exemplary instruments** are presented in purple boxes flanked by two cogwheels.

References and links to **further reading material** are presented in orange boxes highlighted with an open book.

Checklists summarising key practical issues and recommendations for providing high- quality evaluation findings are presented in red boxes marked with a tick.

Depending on the reader's objective, the different sections of the Guidebook may be of more or less interest. However, in order to obtain a comprehensive understanding of the subject matter, we strongly recommend you to go through the 'must reads': Section 1.2, which provides practical guidance for the planning and implementation of an RIE; the summary at the end of Section 3 that contains an overview of the pros and cons of the different evaluation designs and leads you through the process of deciding which design to choose; and Section 3.2 that focuses on the requirements for collecting large-scale quantitative data. Once a particular evaluation design has been chosen, the reader may consult the respective section that describes the chosen design in more detail (Sections 3.1.1 to 3.1.6). Some methodological concepts discussed in these sections require some more methodological background and are therefore further explained in Annexes 5.2 to 5.6. If any of the technical terms or concepts are unclear, the reader can consult the glossary. M&E or evaluation specialists who are not so familiar with the characteristics of CCA projects are advised to consult Section 2, in particular Section 2.1, which discusses the key challenges of CCA projects with regard to their evaluation. Finally, the case study of an adaptation project in Bangladesh illustrates the application of this Guidebook.

1.2 How to plan and implement an RIE for CCA projects

Let's say you have decided to rigorously evaluate a CCA project. How can this be done?

First of all, when planning such an endeavour, you must realise that on the one hand, an RIE needs to comply with scientific standards in order to produce valid and reliable empirical evidence of the impact of an intervention. On the other hand, you may face a number of obstacles, such as time, data and budget constraints (cf. Bamberger 2004), which often make it difficult to adhere to such standards. Planning any evaluation is therefore always a balancing act between fulfilling methodological demands as far as possible and keeping the evaluation feasible in practical terms. To find the right 'middle course' it makes sense to begin the planning of an evaluation by answering a set of questions that determine its general framework:

Questions for determining the evaluation framework

- What is the object of the evaluation? What is the scope of the evaluation in terms of observed period, regions, activities, etc.?
- What are the objectives of the evaluation and which criteria will it use to assess the object of the evaluation?
- Who are the recipients and who are the stakeholders of the evaluation?
- What is the time frame of the evaluation? When are the evaluation findings needed by?
- Which human, financial and organisational resources are available for the evaluation?
- Who will implement the evaluation? What qualifications and experience do the responsible persons have?
- How will the evaluation be implemented? What is the intended evaluation design? Is it feasible with the available resources?
- Which data collection instruments and analysis methods will be applied? Are the people in charge of data collection and analysis familiar with these instruments and methods?
- What tasks need to be performed during the evaluation and who will be in charge?

As all of these aspects are interrelated, it is crucial to answer these questions before actually starting to implement the evaluation. For instance, the scope and objectives depend on the stakeholders and recipients of an evaluation. The design, which determines the choice of instruments and methods, again depends on the available resources and qualifications, and so on. In view of these manifold interdependencies it makes a lot of sense to clarify, document and communicate these issues among all involved stakeholders right from the outset. Usually, when external consultants are commissioned to carry out an evaluation, this is what should be outlined in the terms of reference and agreed upon in the inception report, if not before.

When it comes to evaluating CCA projects, some of these questions might be more difficult to answer than in other evaluations, for instance the second question on the evaluation criteria. Due to the lack of consensus about successful CCA (cf. Section 2.2) it needs to be clarified right from the start what is understood by that concept in the specific context of the project, ideally in line with the project objectives. Another important aspect, which is not unique to the field of CCA, however, is the question of the sectoral expertise needed by the evaluation team. Given the complexity of climate change-related issues, a thorough understanding of the project's theory of change is a crucial prerequisite not only for data collection and analysis but also for integrating the findings into a larger (environmental/development-policy) framework, particularly with regard to the project's contribution to overarching development goals (e.g. MDGs, SDGs, country development goals). Therefore it might be recommended to include a sector expert in the evaluation team who provides technical advice on assessing and interpreting the evaluation findings. Finally, to decide which evaluation design to choose, the project characteristics need to be considered (such as the uncertainty of the framework conditions in which it operates, or the time frame within which its impacts can be observed). Section 3 of this Guidebook therefore provides several suggestions that take these and further characteristics into account.

Photo: © GIZ/Florian Kopp

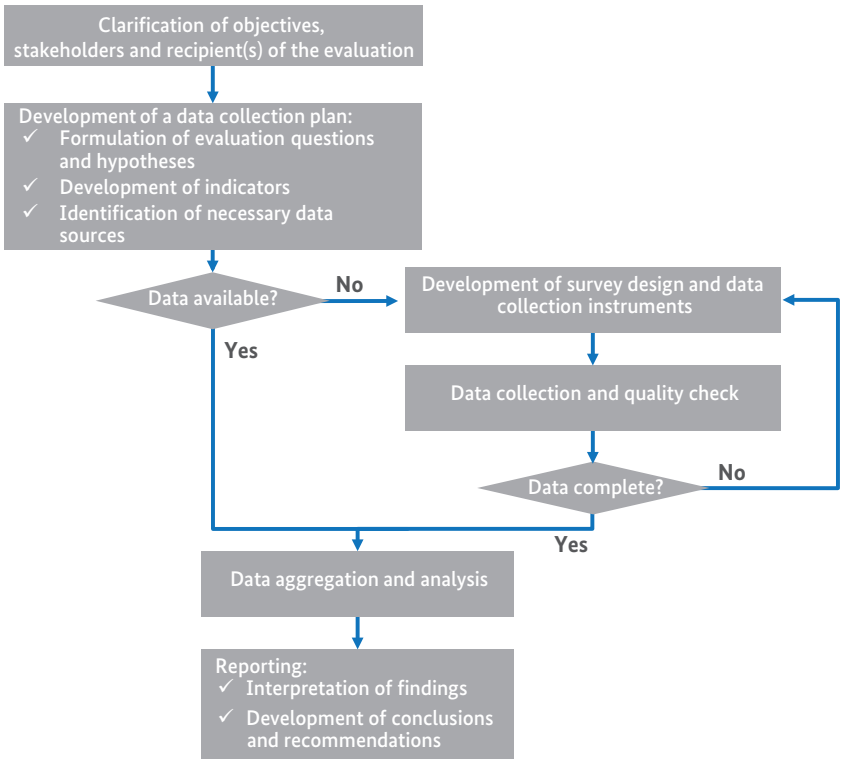
Once these questions have been answered, the first step of any evaluation is to develop a **data collection plan** (also called an evaluation matrix), structured according to the previously defined evaluation criteria. A data collection plan is the reference document that guides the evaluation from instrument development to data analysis. It provides a tabular overview of the evaluation questions and hypotheses to be answered/tested, the required indicators, data sources and availability, data collection instruments, sampling procedures, timing of the data collection, data analysis methods, responsibilities and required resources. The following table gives an example of how such a data collection plan is usually structured:

Table 1 Exemplary data collection plan

Analysis dimension	Hypothesis to be tested/Question to be answered	Indicator	Data source	Sample	Data type	Data collection instrument	Timing of data collection	Data analysis method	Responsible	Required resources
Direct outcomes	The livelihoods of climate migrants have improved	Proportion of climate migrants who have a job	Target group (climate migrants)	5% random sample of treatment and comparison group	Quantitative	Questionnaire	Start of project, end of project, 3 years after end of project	t-test for independent samples	Evaluator no. 1	1 working day (wd) for development of questionnaire, 2 wds for pre-testing, 3 wds for implementing survey, X € for technical implementation
		Income of respondent								
		Proportion of climate migrants who live in formal settlements								
		...	...	...	...	...	...	...	...	...
	Working opportunities have been generated	No. of working opportunities	City corporations	n.a.	Quantitative	Official statistics	Continuous monitoring	Time series analysis	Project officer	n.a.
		Types of working opportunities		2 representatives of each involved city corporation	Qualitative	Interview with corporation representatives	Start of project, then each 6 months update	Qualitative summary	Evaluator no. 2	3 wds for conducting interviews in each intervention area
		Use of working opportunities by target group	Target group	5% random sample of treatment group	Quantitative	Questionnaire	End of project	Descriptive statistical analysis	Evaluator no. 1	See above
		Reasons for acceptance or non-acceptance of working opportunities	Target group	3 groups of 6 to 10 persons in each intervention area	Qualitative	Focus group discussion	End of project	Qualitative data analysis	Evaluator no. 2	1 wd for conducting FGD in each intervention area
		...	...	...	...	...	...	...	...	...
	Access to health services has been improved	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...

Besides the construction of the indicators, another important issue during the development of a data collection plan is the identification of the data sources. If data is already available, the question of whether they are sufficient for answering an evaluation question or whether additional data need to be gathered must be resolved. If the available data are insufficient, it needs to be decided which further data sources can be tapped and which instruments are applicable. There are a number of factors to be taken into account when deciding for or against a certain instrument, including the costs incurred and the human resources and time required to complete the data collection. For this it is necessary to draw up a plan of staff, time and a finance, which also shows who will be carrying out the respective tasks at what times and what costs will be involved. The following flow chart summarises the typical sequence of an evaluation:

Figure 1 **Flow chart for implementing an RIE**



During any evaluation, a number of precautions need to be taken to ensure that the evaluation findings are not only valid and reliable but also well-understood and widely accepted. Therefore, all involved stakeholders should be informed that an evaluation is being conducted and why. Furthermore, the roles and responsibilities of the different stakeholders need to be clarified. A good starting point for considerations of this nature is provided by the feasibility standards established by the American Evaluation Association (AEA), which are intended to ensure 'that an evaluation is planned and conducted in a realistic, prudent, diplomatic and frugal way'.[4]

Feasibility standards of the American Evaluation Association (AEA)

- **F2 – Practical Procedures:** Evaluation procedures should be practical and responsive to the way the program operates.
  ➔ Refers to the problem of the feasibility of scientifically ideal collection procedures with regard to the costs they incur, and to ethical implications.
- **F3 – Contextual Viability:** Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups.
  ➔ Points out the significance of taking into account the interests of all the stakeholders in a balanced way. This is important, since the use of the evaluation findings depends to a high degree on acceptance by the various stakeholders and because access to the relevant information often depends on people's willingness to cooperate.
- **F3 – Resource Use:** Evaluations should use resources effectively and efficiently.
  ➔ Points to the necessity of taking into account the cost-benefit ratio in the implementation of evaluations. In calculating the costs, it is important to consider not only the consumption of tangible (financial) resources but also the intangible outlay (e.g. the use of time and deployment of human resources), though this can mostly be converted back into financial costs. More difficult by far is the quantification of the benefit, since it is not possible to provide any specific information about the expected findings in advance.

[4] Cf. http://www.eval.org/p/cm/ld/fid=103

When evaluating CCA projects, it is especially important to consider the question of contextual viability as interventions may not always generate directly tangible benefits for the target groups. All the more important then is to highlight the long-term objectives of a project when approaching the target groups for data collection. Furthermore, due to the usually large number of different stakeholder groups in CCA projects, it may be necessary to start by identifying the different needs, capacities, objectives and strategies of the different groups as well as their interrelations before collecting data. For instance, when a project focuses on adaptation to diminishing natural resources, it is necessary to include not only the project beneficiaries but all groups who in principle depend on these resources or have a right to use them. Ignoring the often complex actor constellations could lead not only to considerably biased evaluation findings (e.g. overestimation of impact) but also to the refusal of stakeholders to provide important information.

Besides these formal aspects, a number of further organisational issues need to be taken into account during the evaluation process, such as the confidentiality of data. Given the need to maintain the anonymity of those who provide the information, certain rules must be adhered to when providing feedback on the evaluation findings. Apart from the fact that there is a moral obligation to the 'informants' (i.e. interviewees, participants in written surveys, etc.) not to abuse their willingness to cooperate, a number of statutory provisions also exist.

Regardless of how well an evaluation is planned and organised, unforeseeable events can have negative effects on its scheduling and findings. The most common obstacles that must be anticipated when conducting the evaluation include the refusal of the informants to cooperate (refusal to be interviewed, etc.) and the occurrence of negative, unintended effects in data collection (e.g. a rapidly deteriorating work climate due to the respondents not having been given sufficient information about the evaluation). It is advisable in all cases to seek an opportunity to talk to those concerned and if appropriate to arrange a meeting at which everyone is given the opportunity to present their point of view and work together to find solutions, for instance through alternative survey instruments or questionnaires. In this situation it is of key importance that the evaluators are able to give a credible impression of their independence and neutrality.

A potentially controversial question is whether or not recommendations should be included in the report, and if so, how they should be presented. If it is decided when clarifying the assignment that recommendations will be made, it is necessary to point out that they are intended as aids to orientation. In other words, it must be made clear what the evaluation can achieve and what it cannot, in order to avoid unreasonable expectations. This is even more true of the evaluation of CCA projects as due to the uncertainty of future climate developments, a recommendation given at one point in time can prove to be wrong a few years later. Therefore it is even more important to develop such recommendations together with the stakeholders and probably further climate experts in a participatory manner, e.g. in the context of a workshop. Furthermore, it may also be necessary to clarify the spatial and temporal validity of these recommendations and the assumptions under which they were made as regards the development of the framework conditions. In any case it is necessary to differentiate between the evaluation findings and the conclusions and recommendations drawn from them.

We recommend the following books and articles for further reading about practical evaluation requirements:

Bamberger, M., Rugh, J., Church, M., Fort, L., Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints, American Journal of Evaluation, 25(1), 2004, pp. 5 – 37.

Silvestrini, S. , Organizational Aspects of Evaluations in: Stockmann, R. (ed.), A Practitioner Handbook on Evaluation, Edward Elgar, Cheltenham, 2010.

Stockmann, R. (ed.), A Practitioner Handbook on Evaluation, Edward Elgar, Cheltenham, 2010.

2 Evaluating climate change adaptation projects

The first section of this chapter (2.1) provides an overview of the types and key features of CCA project interventions with regard to each level (micro, meso, macro). The second section (2.2) describes the challenges when it comes to evaluating CCA projects, given their complexity. Finally, the third section (2.3) summarises currently used designs and data collection methods to evaluate the results of CCA projects and shows their main shortcomings with regard to the quality of the evaluation findings.

2.1 Types and key features of climate change adaptation projects

While most CCA projects follow a multi-level approach, specific interventions (measures, activities) of such projects usually focus on a singular impact level, be it on a particular target group such as climate migrants (individual level), local authorities (institutional level) or the legal and policy framework (system level). It is also important to understand that all three impact levels are interconnected. For instance, legal or policy changes (system level) will necessarily have effects on the institutional level (e.g. local authorities, companies) and the individual level (e.g. residents of a country or region). The individual benefits perceived by a specific target group (e.g. households in an intervention area) may eventually be scaled up to the institutional and system level (e.g. through spill-over or other dissemination effects). The reason why it is important to make this distinction, though, is that the choice of the appropriate evaluation design depends on the targeted impact level.

2.1.1 Adaptation projects addressing the individual level

A particular type of adaptation intervention primarily affecting the individual level is the so-called community-based adaptation (CBA) approach, which is often facilitated by a local organisation and can be implemented like a small-scale development project. The primary objective of such an intervention is to improve the capacity of local communities to adapt to climate change, applying an integrated approach that often combines traditional knowledge with innovative strategies. Capacity building and awareness-raising are essential elements of such interventions. In fact, UNDP has found that communities and even policy-makers in target countries have limited knowledge of climate change issues, especially of adaptation (UNDP 2009). Thus, in order to be able to participate effectively, locals must be supported to develop a good understanding of these issues, which is why a lot of investments are being made in capacity building. However, working with communities can be a process that requires considerable time, financial resources and energy. But it pays off because community engagement generates ownership (at best) and enables empowerment, giving communities a voice and the chance to participate. Although CBA projects are situated at local level, development cooperation aims to promote action at all levels to achieve systemic and sustainable change. Relevant sectors for adaptation include rural devel-

Photo: © GIZ/Guy

opment, agriculture, water resources management, coastal management, disaster risk management and public health. Typical practices address adjustments in the behaviour of individual groups as well as in the use and development of technologies (e.g. construction of large reservoirs, adjustments to traditional farming practices, and risk reduction for the rural population).

The Community-Based Wetland Management Project (BIRAM) (UNDP 2015a), implemented by Boudhi Investigate and Research Assembly of Men, supported by UNDP with funding from the Least Developed Country Fund, the Government of Bangladesh and the UK Department for International Development, is an example of a project that focuses on the micro level in five villages in Bangladesh (community level). The indigenous peoples living in the project areas are confronted with declining rainfall, rising temperatures and decreased water levels due to climate change. A nearby stream is the main source of irrigation and fish farming. Climate change forecasts predict that temperatures will continue to rise, generating aridity in the target regions, which will have negative consequences on ecosystems and livelihoods. The project focuses on promoting sustainable crop varieties, improved agricultural practices and improved water collection. Moreover, it will train community members in alternative income-generating activities to reduce pressure on natural resources and diversify income sources.[5]

The Sustainable Agricultural Programme (PROAGRO I & II) (GIZ 2015a) – a trilateral cooperation arrangement between Bolivia, Germany and Sweden – is an example of a project that focuses on the individual level. Its objectives are the following: (1) to increase the resilience of smallholders in arid and semi-arid regions of Bolivia to climate change risks, (2) to conserve and distribute scarce water resources as part of integrated watershed management, and (3) to increase income from agricultural production. A fourth component – i.e. adaptation to climate change – was integrated into PROAGRO II, implemented by GIZ on behalf of BMZ. A vulnerability assessment was conducted in 2013, following the approach of GIZ's 'Vulnerability Sourcebook' (Cordero 2014). The main climate risk is water scarcity, which will probably increase in future and will have a negative impact on agricultural production. The findings of the vulnerability assessment show that PROAGRO reduced the vulnerability of smallholder farmers by adjusting crop types and sowing dates and replacing old irrigation technologies with more appropriate ones, which enhanced water efficiency in the parcels (id.). A mid-term evaluation was also conducted in 2013, using a mainly qualitative evaluation design (analysis of primary and secondary data; semi-structured interviews and surveys with national counterparts, strategic partners and target groups; direct observation; focus group discussions) (Kronik 2013).

2.1.2 Adaptation projects addressing the institutional level

Effective institutions are a major factor when it comes to the ability to respond to growing climate risks, since they are designed to perform a set of functions related to decision-making and implementation (Dixit et al. 2011). They play a critical role in increasing society's capacity to adjust as conditions shift and as new climate change knowledge appears. In a constantly changing climate, the process of institutional change represents an important aspect of building adaptive capacities, i.e. the ability of a national government and other entities and individuals to design and implement effective adaptation strategies or to react to negative climate stresses (Dixit et al. 2011). Since national policymakers, international negotiators and funders assist in developing methods and guidelines for adaptation planning, it is essential that they include a focus on building institutional capacity to adapt to climate change impacts. One of several ways in which climate-related challenges may require institutions to make significant adjustments involves meeting the needs of the most vulnerable people, who tend to be poor or marginalised. They usually have few resources with which to adapt, and little say in public decision-making processes.

Various CCA projects comprise interventions that aim at making an impact at institutional level. One example is the National Adaptive Capacity (NAC) framework launched by the World Resources Institute, which assists governmental agencies in includ-

Photo: © GIZ/Andreas K.

[5] More UNDP community-based adaptation projects can be found under the following link: http://www.undp-alm.org/projects/spa-community-based-adaptation-project.

ing institutional capacity development in their adaptation planning processes. It enables its users to systematically assess institutional strengths and weaknesses that may help or hinder adaptation. National adaptation plans may then be better designed to make best use of strengths or remedy weaknesses (Dixit et al. 2011). Another example is the Public Investment and Climate Change Adaptation (IPACC) project (GIZ 2015b) conducted by GIZ on behalf of BMUB in Peru. The aim of the project is to raise awareness – among national and regional policy-makers and relevant technicians – of the potential environmental, social and economic costs of climate change and to encourage decision-makers to consider climate-relevant criteria when formulating and approving public investments. Hence, methods were designed to integrate disaster risk management and CCA considerations into a country's national investment planning to increase the adaptive capacity and reduce climate-related risks (GIZ 2012). Next to initiatives in the policy field, there are also CCA projects carried out by non-governmental organisations. The DANIDA-funded project Capacity Strengthening in Least Developed Countries (LDCs) for Adaptation to Climate Change (CLACC) (IIED 2015), which is operated under the auspices of the International Institute for Environment and Development (IIED), strives to strengthen the capacity of civil society organisations in LDCs to adapt to climate change and foster adaptive capacity among the most vulnerable people. This implies integrating adaptation to climate change into the work of NGOs and simultaneously setting up a system to share knowledge and experience among the adaptation community. In this context, IIED helped to establish the International Centre for Climate Change and Development (ICCCAD) in Bangladesh, which provides training on CCA and development for NGOs, donors, media, government staff and the private sector.

2.1.3 Adaptation projects addressing the systemic level

CCA projects aiming to generate system-wide impacts often follow a top-down approach and comprise interventions at global, regional or national level. Such interventions target policy changes (e.g. sector/development strategies), changes in public opinion on a particular topic (e.g. climate change adaptation requirements) or changes in economic framework conditions (e.g. introduction of climate change- adapted value chains [6]). With regard to CCA projects, two potential intervention types fall under this category: country level and global or regional initiatives. At the country level, CCA projects provide policy advice to support the elaboration of national climate change strategies and action plans or to create awareness and facilitate consultation among (government) stakeholders. The aim is to give consideration to climate change risks in laws, planning, policies and negotiations. Such interventions also support the design of targeted policy strategies, e.g. to climate-proof [7] the agricultural sector and introduce new crops, cropping methods or efficient irrigation technologies.

An example of a CCA project that comprises such interventions at the federal level is the Climate Support Programme (CSP) (GIZ 2015c) launched by GIZ on behalf of BMUB in South Africa in 2009. The project's aim is to assist the Department of Environmental Affairs (DEA) in reducing South Africa's carbon footprint, mitigating climate change impacts and strengthening the country's resilience to climate change (GIZ 2013). The project supported the development and implementation of the Government's white paper on the national climate policy [8] by providing expertise, contributing also to consensus-building among important stakeholders. Section 12 of the white paper clearly mandates the development of an adaptation M&E system. Moreover, sector departments have started the process of reviewing their policies to align them with the white paper. Additionally, together with the BMZ-funded project 'Effective adaptation finance (M&E Adapt)', GIZ is supporting the development and implementation of a monitoring and evaluation (M&E) system for adaptation.

[6] See for instance the project 'Promoting a Value Chain Approach to Climate Change Adaptation in Agriculture in Ghana', which aims to help reduce climate-induced risks and thereby contribute to the achievement of food security and income-generating objectives for rural communities, by focusing on the improvement and adaptation of the cassava value chain. www.thegef.org/gef/sites/thegef.org/files/gef_prj_docs/GEFProjectDocuments/Climate%20Change/Ghana%20-%20%284368%29%20-%20Promoting%20Value%20Chain%20Approach%20to%20Adaptation%20in%20Ag/1-17-2012%20%20ID4368%20%20%20%20Ghana%20SCCF%20Full%20Project%20Document%20for%20Re-submission%20%20Jan%202012%20clean.pdf.

[7] 'Climate-proofing' is a methodological approach applied by GIZ that integrates climate change aspects into development planning. The reason for climate-proofing projects or programmes is the following: even if GHG emissions are drastically re-duced, climatic changes will still occur, which will be fatal for some populations and ecosystems and will have negative con-sequences on the country's economy. These climatic changes will gradually become more visible. Consequently, when de-signing a project, it is of the utmost importance to consider climate change aspects, especially if the project is supposed to run for a longer period of time (Hahn, M., 2010:2-6).

[8] See White Paper of the National Climate Change Response Policy (published in 2011): http://rava.qsens.net/themes/theme_emissions/111012nccr-whitepaper.pdf.

CCA projects with a global dimension include various international players who work together to achieve a common goal. They mainly focus on conceptual work, capacity building and knowledge exchange. The institutions involved are responsible for the development of methods and tools for novel topics such as national adaptation plans (NAPs) and also for piloting these tools and methods in partner countries, thereby cooperating with existing projects within these countries. In a best-case scenario, the results and experiences of the CCA projects are shared with the CCA community to enable an exchange of knowledge and good practices through webinars, workshops or online platforms.[9] One example is the National Adaptation Plan Global Support Programme (NAP-GSP) (UNDP 2015b) for Least Developed Countries, which is a UNDP-UNEP programme financed by the Least Developed Country Fund (LDCF) and the Special Climate Change Fund managed by the Global Environment Facility (GEF). It helps LDCs[10] and developing economies to advance with their NAP[11] process through technical assistance, provides tools and training to support the NAP process within the country and facilitates an exchange of lessons and knowledge through South-South and North-South cooperation. The formulation and implementation of NAPs helps to identify the medium- and long-term adaptation needs of a technical, institutional and financial nature and simultaneously develop and implement strategies and programmes to address these needs.[12]

2.2 Key challenges of climate change adaptation projects

CCA projects (including those with a partial focus on adaptation) pose specific methodological challenges[13], which need to be taken into account throughout the whole project cycle. These challenges make it necessary to fine-tune development agencies' current M&E frameworks, especially with regard to the development of indicators, baselines, milestones, targets and the timing of M&E activities, which need to be adjusted to the longer time horizon of the majority of adaptation initiatives. These challenges are outlined below.

First of all, the remaining **ambiguity of the concepts** used may impede the development of appropriate, verifiable and measurable indicators. There is no clear and commonly accepted definition of adaptation, adaptive capacity and climate-resilient livelihoods, which makes it difficult to define objective indicators. There is also a lack of consensus about what actually constitutes successful adaptation. Consequently, there are **no universal indicators** for CCA projects, since they depend not only on the individual project, but also on the context, scale, sector and location. Unlike mitigation, for which the amount of GHG emissions in the atmosphere serves as a universal M&E indicator, the success of an adaptation intervention cannot be measured by a single indicator. Qualitative assessments are as important as quantitative ones, since many aspects of adaptation are 'soft' (e.g. institutional capacity, behaviour change, etc.).

The establishment of a **baseline** is considered another major challenge for evaluating climate change adaptation projects. A baseline is crucial to measure the project's impact as it provides a reference point against which a change can actually be measured. In climate change adaptation projects, future climate change effects also need to be taken into account (i.e. a baseline projection of how the climate is going to evolve during the project term and beyond). Since climate change is most likely to unfold differently than assumed in the projection, experts speak of a 'shifted baseline'. Shifting baselines are considered to be problematic in terms of planning since they change the context of the adaptation intervention.

Thus, **uncertainty** about actual climate change patterns and their effects is another aspect that needs to be addressed. It may be impossible to anticipate when, for example, the next flood will occur in order to analyse to what extent adaptation took place *in situ* compared to the situation before the intervention. Moreover, adaptation strategies marked as successful in the short term may have negative impacts on vulnerability in the long term.

[9] Examples of online platforms of this kind are: www.adaptation-community.net; www.undp-alm.org; www.adaptationlearning-mechanism.com; www.climate-eval.org and www.mediation-project.eu/platform. Europe-wide: www.climate-adapt.eea.europa.eu.

[10] 26 LDCs have requested help with their NAP process.

[11] The NAP process was initiated under the Cancun Adaptation Framework (CAF).

[12] This paragraph is based on the information obtained on the homepage (www.undp-alm.org).

[13] It must be mentioned, however, that the methodological challenges described below are by no means unique to CCA projects. For instance, most SWAPs have long time horizons (e.g. when reforming a TVET system it may take several decades for benefits to become visible); uncertainty and shifting baselines are a problem in most fragile contexts, and universal indicators (except maybe for education and economic development) are still lacking in most fields of intervention.

Finally, the **long time horizon** of potential climate change impacts poses an additional challenge that needs to be considered when designing the M&E framework for CCA projects. Long-term impacts are unlikely to result from the project alone. Instead, they are rather generated by a series of factors or other interventions in the specific area. Complexity increases when trying to identify the short-term and long-term outcomes that can be attributed to a particular intervention. For example, the choice of farming practices in a sustainable agriculture project also reflects the thought given to current or future climate change. However, it remains difficult to isolate and assess the individual adaptation components reflected by these choices.

2.3 Review of current methods to evaluate the results of CCA projects

The projects reviewed are conducted in developing countries and emerging economies.[14] For the purpose of this guidebook emphasis was placed on the methodology, i.e. the analysis of currently used evaluation designs (experimental, quasi-experimental, ex-post, etc.), the presentation of the main data collection methods used (focus group discussions, interviews, surveys, direct observation, etc.) and, lastly, whether a quantitative, qualitative or mixed-method approach was used. The findings are summarised in Annex 5.1.

During the research process, two problems became visible: first, little use is made of RIEs in CCA projects, which relates to the second finding, namely the difficulty of ascertaining which evaluation methodology was applied. Often, it was difficult to identify which design was used, since there was no reference to it in the methodology section of the evaluation. Due to this difficulty, we checked for keywords in the evaluation papers (e.g. randomisation, randomised controlled trials (RCTs), control or comparison group, quasi- and experimental design, counterfactual and so forth). Very often, none of these keywords were found in the whole document, which is why there is no reference to the evaluation design in the table in Annex 5.1.

Prowse and Snilstveit came to a similar conclusion in their study Impact Evaluation and Interventions to Address Climate Change: A Scoping Study[15], published in 2010, saying that just a few RIEs were conducted (using mainly a quasi-experimental design). Although Prowse and Snilstveit recognised the challenging nature of CCA projects, this does not impede the use of RIEs. Especially with regard to the increasing financial resources made available for climate change mitigation and adaptation projects, it is important to make use of RIEs, because evidence of the effectiveness of spending is required.

Challenges inherent to evaluations of CCA projects involve the poor quality of the baseline (if there is one at all), randomisation and the development of a sample size, which is often not representative. The possible spreading of benefits from target to non-target groups is another limitation often referred to (see example below). In general, data triangulation, i.e. using a combination of qualitative and quantitative indicators (mixed-method approach), is visible in almost all CCA projects[16], which can be seen as a sign of progress in this field. Just to name one example, the evaluation of the project 'Chronic Vulnerability to Food Insecurity', implemented by the World Food Programme, triangulates data from qualitative and quantitative sources. It uses qualitative data from focus group discussions with community members and key-informant interviews with community opinion leaders, and quantitative data involving about 3,000 randomly sampled households (cf. Annex 5.1).

Another interesting finding concerns the way international organisations conduct evaluations. These evaluations are aligned with the Paris Declaration on Aid Effectiveness, basing the evaluation mainly on the assessment of effectiveness, efficiency, relevance, impact and sustainability (the DAC criteria). There is often no reference to the applied methodology.

To name an example, consider the Pacific Integrated Water Resource Management (IWRM) project (UNDP 2015c), established in 2009, which aimed to improve water resource and wastewater management and increase water use efficiency in Pacific island countries in order to balance overuse and conflicting uses of scarce freshwater resources. The evaluation is based on the DAC criteria for evaluating development assistance, but it does not explain which design

[14] Three different research approaches were used for this study: (a) a simple google search, using the following keywords: 'climate change adaptation, impact evaluation, vulnerability, vulnerability assessments, monitoring and evaluation of climate change adaptation, adaptive capacity, and resilience'; (b) a review of the reference lists of the journal articles found in (a); and (c) the scanning of relevant evaluations carried out by development agencies, international organisations or independent evaluation institutions.

[15] Impact Evaluation and Interventions to Address Climate Change: A Scoping Study, http://www.tandfonline.com/doi/pdf/10.1080/19439341003786729

[16] Although CCA evaluations are sometimes based on desk reviews and qualitative indicators (e.g. stakeholder interviews, focus group discussions, direct observations, etc.) and often lack robust quantitative data (see for example project N° 31 in the table in Annex 5.1).

Photo: © GIZ/Ranak Martin

was applied. In the methodological part, only the data collection instruments are outlined (i.e. desk review and analysis, interviews, site visits). The same is true of many other evaluations within the Evaluation Resource Centre [17] of UNDP's Independent Evaluation Office [18].

In general, when screening evaluations conducted by international organisations (see for instance the database of IFAD and the Global Environment Facility (GEF), it becomes obvious that they have not conducted many rigorous impact evaluations so far; hence, there is not much experience in this field. Take the Independent Evaluation Office of IFAD, for instance: narrowing down the search to look only at impact evaluations, there is just one evaluation that was conducted in Sri Lanka in 2013 (Dry Zone Livelihood Support and Partnership Programme (DZLISPP)) [19]. Although it is the only impact evaluation, there is a clear methodology behind it. For the first time, IFAD undertook extensive data collection and analysis, including a qualitative survey (30 key-informant interviews with project staff and relevant government officials), 41 focus group discussions with beneficiaries and a quantitative survey of over 2,560 households. Due to an absence of baseline data (which is often the case in CCA projects), two strategies were applied: (1) an attempt to reconstruct the baseline through recall methods, and (2) the use of a quasi-experimental design that does not strictly require baseline data. Further challenges experienced were sample selection bias due to targeting and the potential spreading of benefits from target to non-target groups.

The Independent Evaluation Office of the GEF offers more than one impact evaluation, but the overall number is modest. Furthermore, the few impact evaluations that have been conducted only concern climate change mitigation projects (there are none for CCA projects yet). However, although the GEF has not yet implemented many impact evaluations, it is clear that they deal with this topic since there are many publications that emphasise the impact evaluation of climate change projects. It will surely be only a matter of time until there is a spillover to CCA projects, too. In 2014, the GEF Independent Evaluation Office together with a number of multilateral and bilateral organisations including GIZ hosted the Second International Conference on Evaluating Climate Change and Development, which aimed to promote and develop methods and good practices for M&E of adaptation and mitigation [20]. A book of conference proceedings will be published by Springer in 2016.

To be fair, it has to be stressed that many CCA projects focused on policy development and capacity building exercises at institutional level, for which RIEs may not be an adequate evaluation design. The National Adaptation Plan Global Support Programme or the Climate Finance Readiness Programme are examples where RIEs might be complicated. These are global CCA projects that focus on policy development, capacity building at institutional level and making sure that the target country is ready (institution-wise) to receive the financial resources earmarked for CCA projects.

It should also be mentioned that 'learning' is in many cases not the focus of the evaluation. Frequently, evaluations focus rather on the question 'Have we done what we said we would?' (accountability) than on 'What happened and why?' (learning). This is especially true of development organisations (examples are outlined in the table in the annex). Additionally, while screening CCA projects carried out by development organisations, it became clear that many projects are still ongoing. This means that an evaluation still has to be completed. At best, this publication may raise awareness among practitioners to consider RIEs for CCA projects in order to increase the use of this evaluation design, gain relevant experience and simultaneously obtain valuable findings that can be shared with the adaptation community.

[17] Evaluation Resource Centre, http://erc.undp.org/index.html

[18] Independent Evaluation Office, http://web.undp.org/evaluation

[19] Dry Zone Livelihood Support and Partnership Programme (DZLISPP), http://www.ifad.org/evaluation/public_html/eksyst/doc/impact/2013/srilanka/index.htm

[20] Presentations and videos from the conference can be accessed at https://www.climate-eval.org/events/2014-conference

3 Rigorous evaluation designs and examples of applicability in climate change adaptation projects

Having discussed the types and key features of CCA projects and the way they are currently evaluated, this section will now provide an overview of the strategies and prerequisites for evaluating the impact of such projects, bearing in mind their characteristics. The first section thus deals with the potentials and limitations of different research designs used for impact evaluations when it comes to their application in the field of climate change adaptation (3.1). Section 3.2 discusses the methodological requirements for providing reliable data on a large scale, with a particular focus on adequate survey sampling techniques.

3.1 Overview of evaluation designs – potentials and limitations for climate change adaptation projects

Before going further into the topic as such, the term 'design' should be briefly explained as it is occasionally used with different meanings and confused with other terms such as 'method' or 'instrument'.

Generally speaking, a research design describes how a research question is intended to be answered. Evaluation designs usually differ with regard to the points in time when data is collected, the data sources and the way the data is analysed (e.g. by comparing target group with comparison group data). Most common evaluation designs are ex-post facto designs and single-difference designs (i.e. before-and-after comparisons or group comparisons). However, as these designs are not suitable for evaluating impact – at least when used as the only means during an evaluation – they are not discussed further in this Guidebook. Instead, we shall go on to present alternative and more sophisticated evaluation designs that can provide sufficient empirical evidence about the attribution or at least contribution of an intervention to observed changes. We will explain the types of projects to which they can be applied, which time, budget and data resources they require and how informative their findings are. We will also assess how valid and reliable the evaluation findings are, provided the design has been implemented correctly.

> A research design describes **how** a research question is answered.

The descriptions start with **experimental** and **quasi-experimental** designs (3.1.1), which are favoured by a number of evaluation experts (e.g. Duflo/Glennerster/Kremer 2008; see also literature list at the end of this section) and considered by some as the 'gold standard' for impact evaluation. In 3.1.2 an alternative approach for estimating the impact of an intervention is discussed, so-called **propensity score matching**, which may be suitable when potential beneficiaries cannot be randomly assigned to a treatment and comparison group. The next section deals with the so-called **pipeline approach** (3.1.3). This approach is of particular use if a project is implemented sequentially but aims to support target groups that feature comparable characteristics. Another approach that is useful when the treatment decision depends on the beneficiary presenting a threshold level of a specific characteristic (e.g. income or age), called the **regression discontinuity** approach, is discussed in Section 3.1.4. In Section 3.1.5, the potentials of **time-series evaluation designs** are discussed by means of the so-called **panel analysis**. Time-series designs are of particular value for long-term projects and observation periods and when complementary statistical data is accessible. The last design discussed in detail is suitable for projects that aim to make a difference at policy level (type I and II projects) as it can provide empirical evidence about their contribution to changes in a complex environment, taking external influences into account (3.1.6). The design is called **structural equation modelling** and comes from the field of econometric research, where it is commonly used to identify causal linkages between social and economic phenomena such as education and health. The section closes (3.1.7) with a tabular overview of the applicability of the evaluation designs discussed above, their methodological and practical requirements and the validity of their findings (if applied correctly). The purpose of this overview is to enable the reader to choose a particular design bearing in mind the adaptation-specific project characteristics and given the financial, technical and logistical framework conditions.

3.1.1 Experimental and quasi-experimental designs

The following section outlines the main characteristics, potentials and limitations, and methodological prerequisites for applying an experimental or quasi-experimental design during an evaluation. We will also discuss the kind of CCA-related interventions to which such designs can be applied and which methodological skills and (financial, human and time) resources are required.

Generally speaking, experimental and quasi-experimental designs (in the field of applied empirical social sciences) consist of a comparison of two groups at two points in time. One group receives a certain treatment (intervention, measure, activity, etc., in the following named 'treatment group') and the other does not. A design is called experimental, or more precisely a **randomised controlled trial** (RCT), if the members of each group are selected at random. This means that (ex-ante) the probability of each individual being a member of the treatment group is the same.[21] The group that has not received any treatment is then called a 'control group'. If such a random selection cannot be ensured (e.g. due to self-selection bias or conscious selection) then such a design is called **quasi-experimental**. In that case the group without treatment is called a 'comparison group'.[22] However, in order to simplify matters, we will consistently use the term 'comparison group' in the remainder of this document.

> Experimental designs require a **randomised selection** of participants for an intervention.

The purpose of employing such a design is to allow for attributing observable changes in a given population to an intervention. This is done by estimating the so-called counterfactual, i.e. the hypothetical situation of that population without having received the treatment. In other words, we do so by answering the question 'What would have happened if no intervention had taken place?' While estimating the counterfactual is also the foundation for calculating the treatment effect in pre- and post-test analyses and comparative analyses, these designs have considerable weaknesses regarding their underlying assumptions. These weaknesses compromise the validity of their findings in many cases. In a simple pre- and post-test design, the measurement prior to the intervention is taken as the counterfactual, assuming that the situation of the target group would have remained stable without the intervention. However, because there is usually a considerable time span between the measurements, this assumption is not valid, as it cannot be assured that no other external factor would have influenced the situation of the treatment group. Comparative designs estimate the counterfactual by collecting ex-post data from both the treatment group and a group that has not taken part in or benefited from the intervention, i.e. a comparison group. In this case one assumes that the situations of the treatment group and the comparison group were identical before the intervention. However, this assumption cannot be held true either due to the potential selection bias. The selection bias may be based on the intervention design (e.g. focusing on the most needy persons), the individual characteristics of the participants (e.g. self-selection by voluntariness) or on practical issues such as a limited budget, the accessibility of the treatment group or other logistical constraints.

Experimental and quasi-experimental designs tackle the issue of intervening time factors and potential differences between target and comparison groups by combining both a pre- and post-test and a comparative design. This means that data is collected from both the target group and the comparison group before and after the intervention. The net effect of an intervention (i.e. net average treatment effect) is then calculated on the basis of the average difference in the observable changes between the treatment and the comparison group (also called double-difference or difference-in-difference approach; see Annex 5.2 for a further introduction to the calculation of the net average treatment effect using this approach).

Photo: © GIZ/Harald Franzen

[21] It has to be highlighted that the equal probability that individuals of a given basic population will receive a treatment needs to be considered at the project planning stage. That means that an RCT can only be applied if the project design allows for a random selection of the treatment group.

[22] The rationale behind this distinction is that a 'non-intervention' group that has not been selected at random cannot by definition act as a 'control' for any time-variant or invariant variables. Therefore, particularly in the social sciences the term 'comparison group' is frequently used in settings where the groups were not selected at random.

CCA-specific requirements and challenges

Experimental and quasi-experimental designs are applicable to CCA projects whenever project measures aim at generating a **direct impact at individual level**, i.e. if a distinguishable target group (e.g. inhabitants of a particular disaster-prone area, farmers in arid regions) receive a specific benefit (e.g. knowledge of how to seek shelter in case of emergency, technical assistance to install drip-irrigation systems). In order to apply such a design, it must also be possible to establish a treatment and control or comparison group. While in some cases the division of a population into a treatment and control/comparison group may not be appropriate for ethical reasons (i.e. if the denial of support would lead to an immediate deadly peril, e.g. medication in an epidemic), in many other cases financial, logistical or technical constraints inevitably lead to selecting a sub-group of a larger population that in principle shares the same needs.

A further requirement for applying an experimental or quasi-experimental design is that it must be possible to detect an impact within a **manageable time frame**. While it holds true that it may take a long time for CCA project impacts to become visible – which is also the case for impacts in other fields, such as education, financial systems or private sector development –, there may also be some ways to work around the relevant constraints. Referring to the first example mentioned above, for instance, one would not have to wait until another disaster happens to see if the project participants are able to seek shelter faster than those who did not participate. A simple observation with a stopwatch during a test alarm would probably do the trick as well. Concerning the second example, it may not be necessary to wait until the environmental framework conditions have deteriorated in another 20 years to see if the drip-irrigation systems lead to the desired impact. Even without any (further) climate change, one can measure the reduced water consumption and increased yield in comparison to conventional farming methods. However, in the latter case too, a time-series design (cf. Section 3.1.5) may be a viable alternative for identifying the long-term impact of the project, particularly if a comparison group can be established.

It must be mentioned, though, that experimental and quasi-experimental designs also feature a number of challenges, which limit their applicability to some extent. First of all, such designs require considerable **technical, logistical and thus financial resources**. As one can imagine, the collection of data from a large number of people (cf. Section 3.2 for further considerations about required sample sizes), particularly if they are located in difficult-to-reach areas, is costly and time-consuming. Furthermore, the necessity to collect this data also from people who did not receive any benefit may be difficult, as they may be not as willing to contribute to such an analysis as those who did. For example, a farmer who did not receive any technical assistance may not be as willing to provide information about his yields. In such cases, a suitable project design (possibly a pipeline approach, see Section 3.1.3) may increase the willingness to participate.

Finally, the findings of such designs are prone to a certain degree of bias such as **spill-over** and **contagion effects**. A spill-over effect means that the control/comparison group indirectly receives a benefit, e.g. through copying or learning from the treatment group (e.g. if the trained habitants in the disaster-prone area tell others who were not trained how to behave, or if farmers who did not receive any assistance adopt irrigation schemes because they recognise their comparative advantage). Contagion effects may occur if the project under investigation is not the only one to provide support in the target area, i.e. if other DC agencies are active in the same field. In that case it may be difficult to attribute the observable impact to the particular intervention (e.g. if the farmer was able to increase his yield because of the technical support provided by GIZ or because of the micro-credit he received from the ADB). While spill-over effects usually lead to an underestimation of the treatment effect (i.e. as the control/comparison group receives a benefit as well), contagion effects can lead to both, the over- and underestimation of the treatment effect, depending which group has been (more) contaminated by other support measures. Thus, in multi-donor settings contribution analyses such as those conducted as part of a structural equation modelling approach (cf. Section 3.1.6) may be more promising.

For further reading about impact evaluation in general and the above-discussed methodological aspects, the potentials and limitations of experimental and quasi-experimental designs and particularly their practical implementation, the following books, articles and papers are recommended:

General methodological introduction

Bertrand, M., Duflo, E., Mullainathan, S., How Much Should We Trust Differences-in-Differences Estimates?, Working Paper 01 – 34, MIT, Cambridge, 2001.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., Befani, B., Broadening the range of designs and methods for impact evaluations, Working Paper 38, DFID, London, 2012.

White, H., Some Reflections on Current Debates in Impact Evaluation, International Initiative for Impact Evaluation, New Delhi, 2009.

Practical guidelines and examples

Chambers, R., Karlan, D., Ravallion, M., Rogers, P., Designing impact evaluation: different perspectives, International Initiative for Impact Evaluation, New Delhi, 2009.

Khandker, S.R., Koolwal, G.B., Samad, H.A., Handbook on Impact Evaluation. Quantitative Methods and Practices, The World Bank, Washington DC, 2010.

Leeuw, F., Vaessen, J., Impact Evaluations and Development. NONIE Guidance on Impact Evaluation, NONIE, Washington DC, 2009.

Vaessen, J., Todd, D., Methodological challenges of evaluating the impact of the Global Environment Facilitys biodiversity program, Evaluation and Program Planning, 31, 2008, pp. 231 – 240.

3.1.2 Matching techniques

In an experimental design, comparing the differences in the outcomes of the treatment and control group would be sufficient for calculating the net effect of an intervention. However, as treatment groups are rarely selected randomly, often such a design cannot be implemented in practice. To obtain valid evaluation findings all the same, the induced selection bias described above needs to be compensated for. Various approaches are available for this purpose; among these, matching techniques appear to be especially suitable in evaluation studies. The basic idea of matching techniques is to establish for each member of a treatment group a sub-set of possibly similar comparison group members in order to estimate the treatment effect by calculating mean group differences. That means that the counterfactual situation of the treated person (i.e. what would have happened if this person had not participated in the project) is approximated through the assignment of similar persons who have not participated.

Matching techniques aim to **approximate the counterfactual** by comparing the outcomes of individuals from the treatment and the comparison group that have identical or at least similar characteristics.

A common matching approach is to identify **statistical twins** by comparing individual characteristics (i.e. dimensions) that are supposed to have an influence on the outcomes of an intervention. For example, if one suspects that the effectiveness of a training measure (e.g. teaching farmers how to grow drought-resistant crops) also depends on the age and gender of the beneficiary, the treatment and comparison group can be clustered accordingly and the outcomes can be individually compared for each sub-sample. The following table illustrates the principle of the matching approach using the mentioned example:

Table 2 Example table for matching on observables

Group	Treatment group		Comparison group	
Gender	Male	Female	Male	Female
< 40 years old				
≥ 40 years old				

Annotation: Each cell/arrow colour indicates an individual comparison.

As the table shows, each sub-sample of the treatment group (e.g. men of the age of 40 and above) is compared with a sub-sample of the comparison group that has the same characteristics. The problem with this so-called **matching on observables** approach is that the size of comparable sub-samples decreases as the number of relevant dimensions grows. This again limits the possibility of conducting statistical tests (e.g. t-test for comparing mean values of independent samples) that require a minimum cell count of approximately 30 cases in order to reach valid findings.

Another matching technique that has gained attention in the last years is called **propensity score matching** (PSM; cf. Rosenbaum/Rubin, 1985). The advantage of this technique is, that, in contrast to matching on observables, it provides a one-dimensional score for comparison (the so-called propensity score) that indicates the probability of any individual being a member of the treatment group. Since this approach is methodologically quite demanding and thus requires a strong theoretical background, its application will not be discussed here in detail. Instead, a short introduction into PSM and further reading material can be found in Annex 5.3.

CCA-specific requirements and challenges

When we look at how these matching techniques can be applied to CCA projects, the same characteristics apply as mentioned in the preceding section because they are primarily used on an **individual level**. While institutional or even spatial (regional, city-wise) matching can also be found in evaluation practice, the methodological foundations and practical experiences concerning the validity of the findings provided by such matching approaches are still rather limited. Therefore, it is not recommended to pursue such an approach 'above' the individual level. Regarding the variables that are used for matching in the context of CCA projects, it may be possible to work with characteristics such as proximity to hazardous areas, degree to which individuals are affected by climate change effects or affiliation to specific ethnic, social or religious groups, etc.

The challenges presented by matching approaches include two issues in particular: The **necessity of a relatively larger sample** in order to find sufficient matching partners and the question of which **characteristics to choose for the matching**. While the first issue constitutes a technical problem that can be solved by allocating sufficient resources, the second is a theoretical one, which requires comprehensive sectoral and regional expertise. The analysis of matched data can lead to significantly different findings compared with unmatched data, depending on which characteristics are considered relevant. Therefore, it is of the utmost importance to substantiate the selection of the matching variables based on sound theory. The most important questions to be answered in that regard are:

- Could the characteristic influence the effectiveness/outcome of the intervention for the individual?
- Does the basic population comprise sufficient individuals with different characteristics?
- Will it be possible to identify sufficient individuals in the target group with identical characteristics?

We should also point out that matching may not prevent potential bias as described in 3.1.1. However, it may be possible to perform matching using aspects such as 'degree to which individuals are affected by other interventions' in order to at least reduce the risk of contagion effects.

For further reading on the methodological aspects, potentials and limitations of PSM as well as its practical implementation, the following books, articles and papers are recommended:

General methodological literature about PSM

Becker, S., Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. The Stata Journal, 2(4):358 – 377.

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity scores. An introduction and experimental test. Evaluation Review, 29(6):530 – 558.

Morgan, S.L., & Winship, C. (2007). Counterfactuals and causal inference: methods and principles for social research. Cambridge, N.Y.: Cambridge University Press.

Guo, S., & Fraser, M.W. (2010). Propensity score analysis: Statistical methods and applications. Thousand Oaks: Sage.

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. Psychological Methods, 15(3):250 – 267.

Practical guidelines and case-studies

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1):31 – 72.

Cook, T.D., & Steiner, P.M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. Psychological Methods, 15(1):56 – 68.

Diaz, J.J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. Journal of Human Resources, 41(2):319 – 345.

Gaus, H.; Müller, C. E. (2012). Evaluating free-choice climate education interventions applying Propensity Score Matching. Evaluation Review, 35(6):673 – 722.

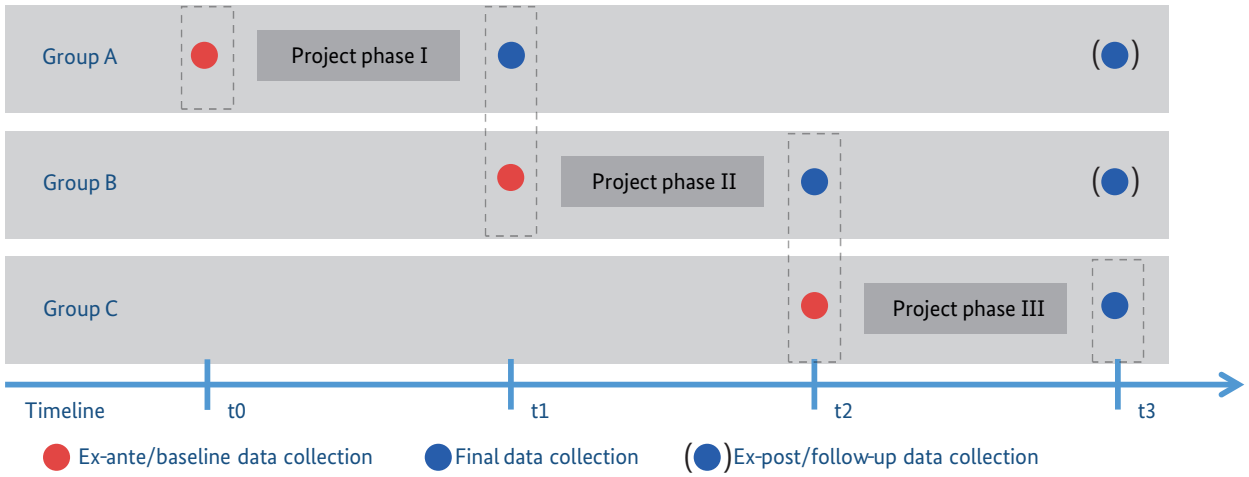
3.1.3 Pipeline approach

Often, projects are not implemented on a broad scale right from the beginning, targeting the entire potentially relevant target group, but sequentially. This may be because of logistical or budgetary reasons. In these cases, a so-called **pipeline approach** can be of particular use. The idea behind the approach is that the section of the target group that receives a treatment (i.e. benefits from or takes part in a project) at a later stage can be used to approximate the counterfactual situation of those who did receive the treatment at the time of data collection. In other words, the part of the target group that is not yet affected serves as a quasi-comparison group for the already affected part of the group. In principle the approach can be applied repeatedly during an intervention, depending on the number of phases in which the treatment (i.e. measures, activities, etc.) is provided.

In a pipeline approach, the part of the treatment group that has not yet received any treatment is used for comparison.

Figure 2 illustrates the basic concept of the approach using a three-phase intervention design:

Figure 2 Illustration of the pipeline approach



As the figure shows, in a three-phase project a total of four data collections (t0 – t3) are necessary, of which at least during two (t1 & t2) data sets have to be collected from two groups (group A & B @ t1 respectively group B & C @ t2). The treatment effects would then be estimated as in an ordinary single-difference design whereby group B would represent the comparison group to group A and group C to group B. Provided the total project duration suffices, ex-post/follow-up data from groups A & B could also be collected at t3, which would probably allow further conclusions to be drawn about the sustainability of the achieved results. Further data collections are conceivable, such as from group C at t0 and t1 in order to estimate spill-over effects from groups A and B or to control for external influences on outcome variables (e.g. contagion effects from other projects). In any case the approach is quite flexible as it can be modified according to the information needs. It can even compensate for a 'forgotten' baseline study to some extent, provided the group characteristics are comparable and remain relatively stable over time, i.e. when the influence of external factors is negligible.

CCA-specific requirements and challenges

With regard to evaluating CCA projects, the pipeline approach is applicable for interventions that are planned to be implemented in phases, for instance in different areas and/or institutions (e.g. development and implementation of environmental information systems for local authorities in different provinces) and that aim to generate impact at **individual or institutional level**. Provided that this is the case, as outlined above, it is still possible to start collecting data after the project has started, i.e. it is **not necessary to have baseline data for the first project phase**. This may be a considerable benefit for all projects that follow a **replication** or **scaling-up design** with a preceding pilot phase in which the (technical, logistical, etc.) feasibility of the measures is tested before they are rolled out in other areas.

The pipeline approach might also be of particular value for CCA projects, which are confronted with the **problem of shifting baselines** (cf. 2.2). The pipeline approach makes it possible to trace developments by comparing the baselines for each group (provided the group characteristics are in principle comparable). It may also be an option for projects that aim at generating **long-term impacts** that might not be observable directly at the end of their implementation.

When applying a pipeline design, however, it should be borne in mind that each data collection phase is **just as labour-intensive and time-consuming as a quasi-experimental design**, given that data needs to be collected for both the treatment and a comparison (i.e. future treatment) group, except for the first and last round. The same quality criteria therefore apply.

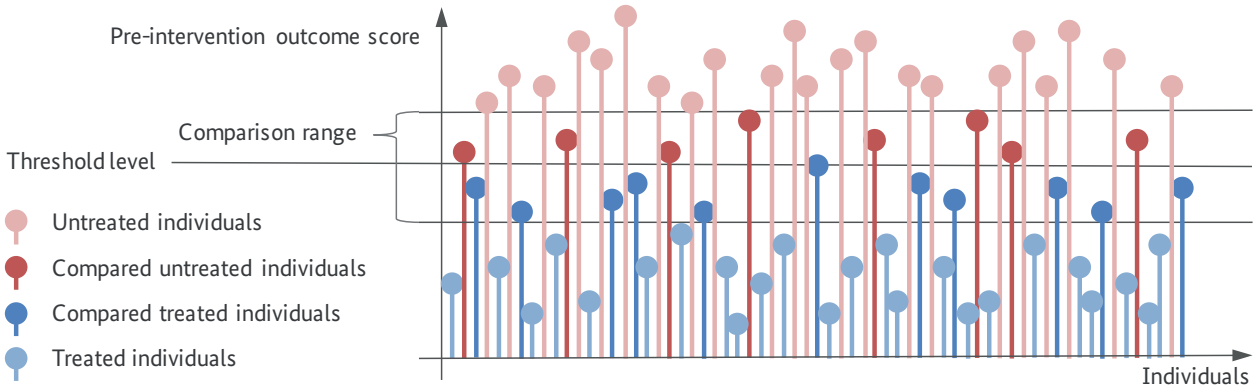
The references to literature on experimental and quasi-experimental designs at the end of Section 3.1.1 are also suitable for further reading on the pipeline approach.

3.1.4 Regression discontinuity design

The approach discussed in this section is of particular use when project participants/beneficiaries are consciously selected on the basis of one specific characteristic that distinguishes them from non-participants/beneficiaries (e.g. income), though they are similar in a number of other respects (e.g. location, income sources, education). If the participation/eligibility of an individual depends on the value of a particular indicator that is relevant to the outcome of an intervention, the treatment effect can be estimated by the so-called **regression discontinuity design (RDD)**. The basic idea behind the approach is to estimate the treatment effect by comparing the pre- and post-treatment indicator values (scores) of only those individuals from the treatment and comparison group that are most similar with regard to their pre-treatment outcome score, i.e. that lie within a comparison range close to the threshold level. Figure 3 shows the sample selection for assessing the treatment effect.

> The regression discontinuity design can be applied if project participants/beneficiaries are **selected based on a special characteristic** that distinguishes them from non-participants/beneficiaries.

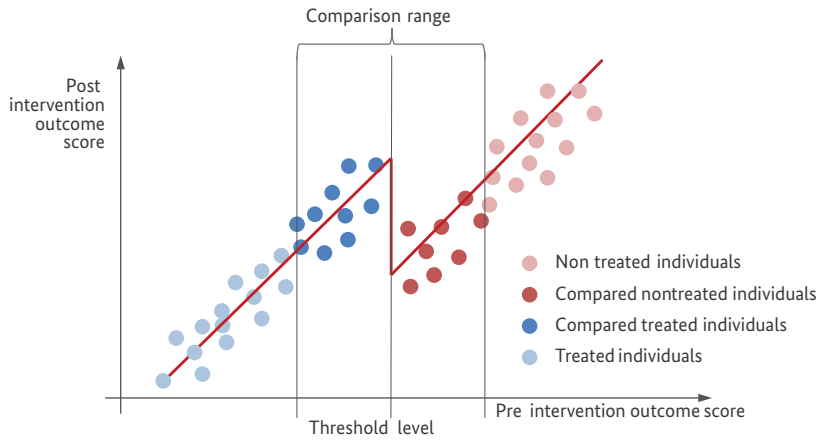
Figure 3 Selection of individuals for comparison



It is mandatory to adhere to the threshold level throughout the selection process as otherwise the regression analysis would lead to less valid findings. Furthermore, it must be assured that the individuals cannot influence the pre-intervention outcome score (e.g. by temporarily lowering their income in order to receive support) as this would also bias the findings.

Once the sample has been identified, the pre- and post-treatment outcome scores are used to calculate regression functions for each group. If the intervention had an effect on the treatment group, the functions should show different values at the threshold level. This difference can be interpreted as the local average treatment effect for individuals that are most similar, i.e. whose outcome values lie very close to the threshold line. The discontinuity of the regression functions of the treatment and comparison group can also be visualised by a Cartesian coordinate system in which the X-axis represents the pre-treatment and the Y-axis the post-treatment outcome score. Figure 4 shows how such a discontinuity may look.

Figure 4 Outcome scores before and after intervention



As the treatment effect estimated with the RDD is based on a comparison of most similar individuals, it can be considered to be widely unbiased. However, a disadvantage of the approach is that the effect may not be generalisable for the entire population because individuals become less comparable the further they are from the threshold level. Another disadvantage is the fact that in comparison to other designs, RDD only uses a small fraction of the entire population to estimate the treatment effect.

Due to its relatively ambitious requirements and methodological reservations, RDD has only played a minor role in impact evaluation studies so far. Nevertheless, it may be applicable particularly for policy-field evaluations with large and clearly defined treatment groups, provided the above-mentioned methodological requirements can be met. In some publications (Khander/Koolwal/Samad 2010), the use of a regression discontinuity approach is also recommended in scenarios where a threshold level is spatially defined (i.e. a geographical, national or administrative border). However, such an application needs to be carefully reviewed as populations in different regions present unobserved covariates (e.g. cultural or religious background) that may limit their comparability.

CCA-specific requirements and challenges

RDD can be of particular use for evaluating CCA projects that focus on creating an impact at **individual level** and are implemented under framework conditions that involve a **high degree of uncertainty**, provided the projects comply with the methodological requirements outlined above. The benefit lies in the comparatively simple data collection setup and clear interpretability of the evaluation findings, which are quite robust against bias at least for those individuals who are very similar.

On the other hand, it has to be added that the findings are only of **limited external validity**. Furthermore, as only very similar individuals are compared (at least with regard to the outcome variable that is of interest), the **sample selection can be less efficient** because only a smaller number of individuals (i.e. those who are very similar) can be used for comparison. Another challenge, which is not restricted to CCA projects, though, is that it requires the project to **adhere to a defined threshold level**, not to 'shift' it or make exceptions, which in practice often proves to be a problem. It is also only suitable for interventions where the impact becomes visible in a manageable time frame, which is not often the case for CCA projects.

Finally, it should be added that besides the outlined basic RDD, there are several further, more sophisticated variations of the approach. These include random assignment at the threshold level or with more than one threshold level (e.g. if the treatment group is split further in sub-groups that receive different kinds of support depending on their pre-treatment outcome score), which cannot be discussed here.

Therefore, the following books, articles and papers are recommended for further reading about methodological and practical requirements as well as further applications of the approach.

Imbens, G.W., & Lemieux, T., Regression discontinuity designs: A guide to practice, Journal of Econometrics, 142(2), 2007, pp. 615 – 635.

Khandker, S.R., Koolwal, G.B., Samad, H.A., Handbook on Impact Evaluation. Quantitative Methods and Practices, The World Bank, Washington DC, 2010.

3.1.5 Time-series designs

While most evaluations focus on a limited time frame defined by the duration of an intervention, sometimes it may be necessary to cover a longer period, e.g. when the effectiveness of a series of past support activities or the development of (political, economic, etc.) framework conditions are to be reviewed. In those cases **time-series designs** can be of particular use. We should add that there is no single time-series design; this is rather an umbrella term for a number of different approaches. What they have in common is that they are based on repeatedly collected data over a longer period. Time-series designs can be distinguished by whether they make use of only descriptive or also inferential statistics, if data from the same analytical units (e.g. individuals, households, organisations, cities, countries) is used or not, and if the data is quantitatively or qualitatively analysed.

Time-series designs are based on frequently repeated data collections and are of use for **measuring long-term impacts and the development of framework conditions.**

While it is simple to analyse the development of individual indicators descriptively by means of a trend analysis, the identification of causal relationships calls for more sophisticated approaches. Such approaches are often based on regression analyses that make it possible to estimate the influence of several independent variables (e.g. resources, contextual factors) on dependent variables (e.g. attitudes, behaviour, well-being). We shall go on to discuss an exemplary design that has been successfully applied in a number of impact evaluation studies, the so-called **panel analysis**.

A panel analysis is based on repeated data collections from identical analytical units. The data needs to be comparable with regard to their information content, i.e. the data collection instruments of each survey have to contain identical question and answer categories. Furthermore, the data sets must be assignable to each respondent throughout the observation period. Panel analyses usually require datasets with a large number of analytical units in order to be able to draw conclusions from the panel sample that apply to a larger basic population (cf. 3.3.1). If these requirements are met, panel analyses make it possible to trace developments not only on average but also on an individual basis and thus to identify variations caused by individual differences (i.e. unobserved heterogeneity). In addition, given a sufficient observation period, panel analysis makes it easier to identify the chronological sequence of changes and thus develop hypotheses about causalities. The treatment effect can be estimated on the basis of various models. The two most common of these, the so-called **fixed-effects model (FEM)** and the **random-effects model (REM)**, are briefly outlined in Annex 5.4.

CCA-specific requirements and challenges

Time-series designs appear to be most suitable for CCA projects as they are in principle applicable at **each impact level** and are of particular benefit when it comes to **measuring long-term impacts**. As they make it possible to control time-variant confounding factors, they are likewise useful for dealing with **uncertain framework conditions** as well as with **shifting baselines**, given a sufficiently large sample size and/or number of repeated data collections. Due to their flexibility they can be applied in projects that focus on the adaptation of particular target groups to changing (climate, environmental) framework conditions, on the improvement of institutional capacities to deal with climate change-rooted environmental, social or economic problems, or on national, regional or global developments such as adaptation strategies in response to particular climate challenges. Thus, time-series designs are in fact more common in policy studies than in project evaluations.

The challenges associated with time-series designs lie not so much in their practical implementation or the required financial resources as in the **expert knowledge** needed to do the 'maths'. They also call for a **strong sectoral background** in order to develop the right cause-and-effect hypotheses and draw the right conclusions from the analysis findings. Very few evaluation experts have the know-how to adequately draft and implement such a design. The **high theoretical and methodological demands** are reflected in a multitude of disputes among experts concerning the credibility of research, which is of course not helpful for clearly communicating evaluation findings.

For further reading on panel data analysis, the following literature is recommended:

Berrington, A., Smith, W.F., Sturgis, P., An Overview of Methods for the Analysis of Panel Data, NCRM Methods Review Papers, NCRM/007, ESRC National Centre for Research Methods, Southampton, 2006.

Hsiao, C., Analysis of Panel Data, Cambridge University Press, Cambridge, 2014.

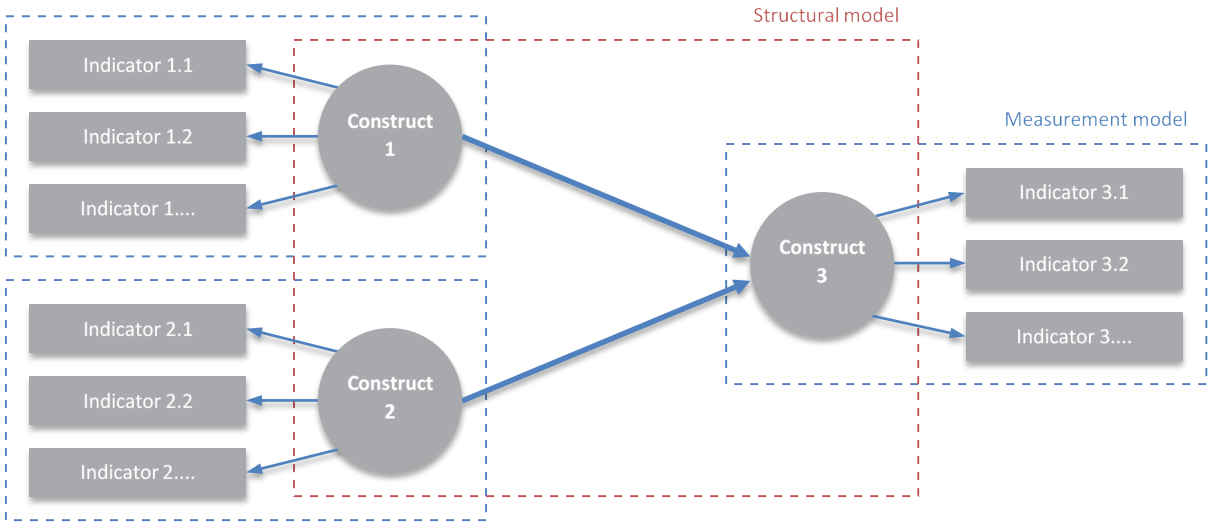
3.1.6 Structural equation modelling

While the approaches discussed in the sections before are suitable for project types that differentiate between a treatment and a comparison group, we will now go on to present an approach that may be applied for evaluating large-scale policy-based programmes that aim to affect an entire (national, regional or global) population or sector (e.g. environmental or education sector, labour market). With the so-called **structural equation modelling (SEM)** approach, it is therefore not possible to attribute an observed effect to a specific intervention but to estimate the statistical relationship between several factors over time, one of which may be an intervention.

Structural equation models make it possible to **estimate statistical relations** between constructs, which are defined by **empirical measurable indicators**.

Basically, SEM is a multivariate statistical approach for testing hypotheses (deductive approach) about causal relationships between two or more latent constructs. Its special feature is that it makes it possible to test multiple statistical relationships (which are assumed to be causal) at the same time and can thus be used in complex environments. In this context, latent construct means that the causes and effects (e.g. health, employability) cannot be measured directly but only by a number of manifest (i.e. empirically measurable) indicators (for health: e.g. blood counts, weight; for employability: education level, qualifications). In principle, a structural equation model consists of two model types: **measurement models** and **structural models**. Measurement models describe the relationship between a latent construct and its indicators; structural models describe the connection between latent constructs. The following figure illustrates the differentiation between the model types according to the graphical representation of the system model described above:

Figure 5 Generic layout of a structural equation model



It is important to note that the indicators that operationalise the constructs and the causal relationships between the constructs have to be specified and theoretically substantiated by the researcher. That means that it needs to be defined in advance which is the independent (i.e. exogenous) and which is the dependent (i.e. endogenous) variable. In other words, the direction of the dependency has to be defined by the researcher. In the figure above, constructs 1 and 2 are exogenous variables (i.e. causes) while construct 3 is an endogenous variable (i.e. effects). A practical example on how to implement a SEM is given in section 4.2.

CCA-specific requirements and challenges

With regard to the application of SEM for evaluating CCA projects, the requirements and challenges summarised in the previous section on time-series designs are widely transferable with the difference that SEM is mainly suitable for projects that aim to create an impact at **institutional and system level** (e.g. political strategies). The practical requirements relating to SEM are not so difficult, though it does take **comprehensive theoretical know-how** to develop the model. As mentioned above, the validity of the models relies on the expert knowledge of the people who create them. Thus, the findings of SEM analysis are as right or wrong as the model they are based upon. Here the particular danger lies in omitting a system-relevant construct (i.e. influential factor) or indicators that define such a construct, which would cause substantially biased findings. What aggravates the problem is that in contrast to the designs discussed above, unfortunately **no quality criteria** are available for assessing the **validity of an SEM model**. Nevertheless, given sufficient expert knowledge, SEM can be of great benefit for providing the 'bigger picture' and thereby revealing the manifold causal linkages between social, political, economic and environmental developments in the long run.

For further reading on the methodological aspects, potentials and limitations of SEM as well as its practical implementation, we recommend the following books, articles and papers:

Cassel, C., Hackl, P., Anders., H.W. (2010). Robustness of partial least-squares method for estimating latent variable quality structures. Journal of Applied Statistics, 26(4):435 – 446.

Chin, W., Newsted, P.R. (1999). Structural Equation Modeling Analysis with small Samples Using Partial Least Squares. In: Hoyle, R. (ed.) Statistical Strategies for Small Sample Research, Thousand Oaks: Sage.

Diamantopoulos, A., Riefler, P., Roth, K.P. (2008). Advancing formative measurement models. Journal of Business Research, 61(2008):1203 – 1218.

Garthwite, P.H. (1994). An Interpretation of Partial Least Squares. Journal of the American Statistical Association, 89(425):122 – 127.

Haenlein, M., Kaplan, A.M. (2004). A Beginner's Guide to Partial Least Squares Analysis. Understanding Statistics, 3(4):283 – 297.

Temme, D., Kreis, H., Hildebrandt, L. (2006). PLS Path Modeling – A Software Review. SFB 649 Discussion Paper 2006-084. Berlin: Humbold-University.

Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., Lauro, C. (2004). PLS path modeling. Computational Statistics & Data Analysis, 48(2005):159 – 205.

3.1.7 Summary

The prior descriptions show that a number of evaluation designs of sufficient methodological rigour are available that can provide reliable empirical evidence about the impact of CCA projects. This section gives advice on which design(s) to choose, based on the respective project characteristics. Here, it is again important to note that the applicability of a design primarily depends on the aggregation level at which an impact is generated (cf. Section 1). Accordingly, since many CCA projects operate on more than one level, a combination of two or more evaluation designs might be necessary in order to assess the total impact balance of a project. This complies with the need to combine qualitative and quantitative methods, and triangulate data, methods and researcher perspectives.

At the individual level, the decision which design to choose depends firstly on the existence of baseline data. As explained in the previous sections, three designs, namely experimental, quasi-experimental and regression discontinuity designs require such data. This necessarily means that these designs can only be applied if they were drafted during the planning of the project. If this is the case the next question is whether or not the beneficiaries were selected at random. If this condition is also fulfilled, in principle an **experimental design** can be applied, provided sufficient resources are available to do so. As described in 3.1.1, RCTs are quite costly due to the laborious data collection process, which usually calls for the assignment of a large team to collect data. It should also be said, though, that an RCT provides the highest internal validity and allows for clear impact attribution.

If the project design does not allow for a random assignment of the treatment group, a r**egression discontinuity design** may also be feasible. However, such a design requires the treatment group to be selected based on a specific outcome variable of interest (e.g. proximity to a disaster-prone area). If this is not the case, a **quasi-experimental design** may be a feasible alternative. The time and budget requirements of these two designs are comparable with those of an experimental design with the constraint that RDD might require a larger sample in order to identify sufficient cases that are close enough to the 'threshold level' to be compared. With regard to the validity of the findings and their informative value, we should add that while RDD-based evaluation findings are restricted to groups that are quite similar, findings obtained using a quasi-experimental design can almost approximate those of an RCT if the samples are adequately matched (e.g. with **propensity score matching**).

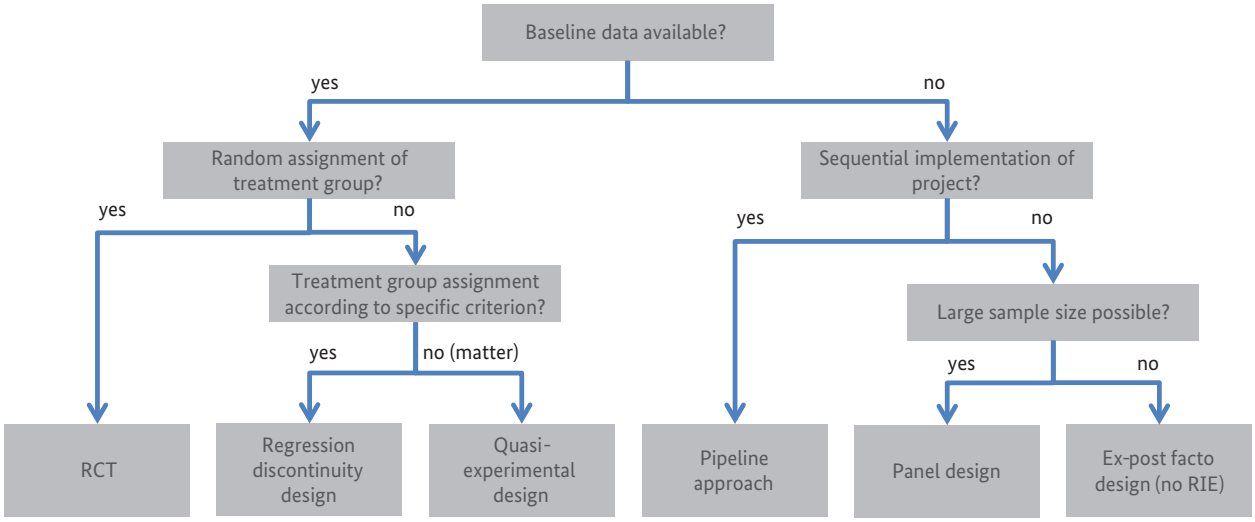
In case no baseline data has been collected beforehand, two further rigorous evaluation designs may be used: the **pipeline approach** and the **panel design**. Both designs have further requirements: the pipeline approach requires that the project be implemented sequentially (e.g. in different regions, with different target groups), and the panel design that a larger sample size be available to enable reliable estimation of the impact. It should also be considered that while a pipeline approach enables the attribution of impacts to an intervention, a Panel Design 'only' allows the contribution of an intervention to an observed change.

If none of the above prerequisites can be satisfied, most likely only a so-called **ex-post facto design**,

with or without (single) comparison with a non-intervention group or with reconstructed baseline data, may be possible. However, such a design is not regarded as 'rigorous' as it cannot provide for an attribution or contribution analysis.

Based on these selection criteria, the following decision tree can be used to identify the 'right' evaluation design for interventions that aim to generate impact at individual level:

Figure 6 Decision tree for selecting an evaluation design for measuring impacts at individual level

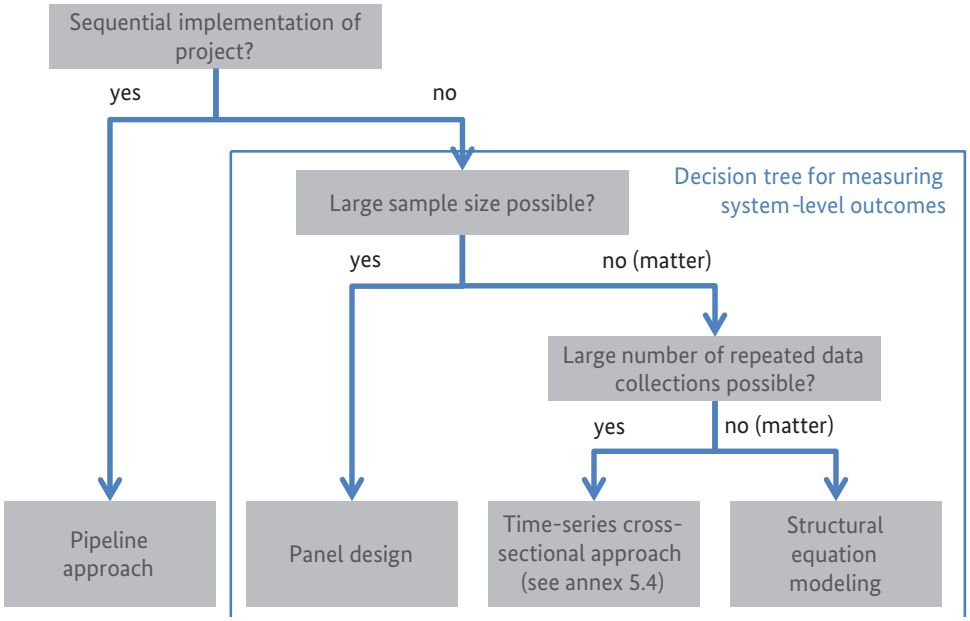


If an intervention aims to generate an impact at institutional level, four rigorous evaluation designs are suitable. Three of them have been previously introduced: pipeline approach, panel design and structural equation modelling (SEM). A potentially suitable design not discussed before is a time-series cross-sectional approach (TSCS). As already discussed, **the pipeline approach** requires a sequential implementation of the project in order to be applicable, leaving three for those that are not. As also discussed, a **panel design** can only be applied if the sample (here: of institutions, e.g. governmental or non-governmental organisations, enterprises) is large enough for statistical calculations (approx. 30). If the intervention only focuses on a few institutions but at least about 10 repeated observations are possible, then a **TSCS** might be the design of choice (see annex 5.4). Finally, if none of these conditions can be fulfilled, **SEM** is the only 'rigorous' option to provide meaningful evidence about the impact of an intervention operating at the institutional level, although it should be added that SEM, like panel and TSCS designs, can only provide for the estimation of its contribution. On the positive side, we should note that the latter three designs can be implemented with considerably less resources, provided that statistical data of sufficient quality are available for quantitative analysis.

With regard to interventions geared to the system level (e.g. SWAPs, government advice), the same selection of designs is basically available, except the pipeline approach (due to the lacking comparability of sectors or countries, for instance). Therefore, in the following figure a joint decision tree for both intervention levels summarises the available choices of RIE approaches for institutional and system-level interventions:

Photo: © GIZ/Ranak Martin

Figure 7 **Decision tree for selecting an evaluation design for measuring impacts at institutional and system level**



Finally, in table 3 on the next page all suggested RIE designs are listed with brief summaries of the impact level they are suitable for, the project characteristics that must be in place, the data, time and budget requirements, the validity of the analysis findings and their informative value (i.e. if they enable attribution or contribution analysis).

3.2 Providing reliable large-scale data

As outlined above, empirical data is the foundation of any evaluation – and more or less of applied social sciences in general, whether they are collected during the research process (i.e 'primary data') or reanalysed on the basis of statistical or documentary data (i.e. 'secondary data'). Moreover, impact evaluations mostly require data collections on a large scale in order to provide empirical evidence about the changes that have occurred in an intervention's sphere of activity. Such data is usually collected by means of a survey. Surveys are used when mostly quantitative data from a considerable number of people, e.g. all or a significant share of the project's beneficiaries, non-beneficiaries or more generally residents of a certain area or particular social strata, needs to be analysed. During the data analysis, statistical parameters such as average means, shares or distribution measures (e.g. standard deviation, quantiles) of a given population (i.e. descriptive statistics) are calculated. Provided that the data is collected from a random sample of sufficient size (see further elaborations below), probabilities regarding the values and distribution of certain parameters in a larger population can also be estimated (i.e. inferential statistics). In the following, the methodological aspects to be considered when collecting quantitative data by means of a survey are discussed, with particular focus on an appropriate sampling strategy.

Surveys often deal with knowledge, opinions, attitudes, beliefs, behaviours, plans, backgrounds, developments or assessments of the respondents. The primary survey tool for data collection is the questionnaire. With a questionnaire (in contrast to an interview guideline), all respondents are asked the same questions in the same order. The information gathered by a survey is often used to identify variations of these parameters between different points in time (i.e. in order to identify developments) or differences between groups (e.g. in order to attribute the outcomes to a certain intervention or to reveal systematic differences due to group characteristics such as gender, age, etc.) or both. With a survey, information is collected in a systematic, structured and mostly standardised way. Structuring the subject of a survey means that not all potentially available information is relevant, but only data that help to describe and/or assess a certain measure, development or framework characteristic. Standardisation means that the respondents can choose from a number of answer options and that these options are grouped according to a predefined categorical system. This means that the data is statistically calculable and, consequently, survey findings are comparable. Such a categorical system can be dichotomous, nominal, ordinal or metric. Table 4 contains an example of each single-choice question.

Table 3 Required project characteristics and data

Evaluation design	Impact level	Required project characteristics	Data requirements	Time and budget requirements	Validity of findings	Explanatory power
Experimental (RCT)/quasi-experimental	Micro (individual) level	• Distinguishable treatment group • RCT requires random assignment to treatment and comparison group	• Ex-ante (baseline) and ex-post data from treatment and control/ comparison group	• Depending on the size and accessibility of the target group • Data collection process often quite costly as a large team of enumerators is needed	• RCT has highest internal validity • Internal validity of quasi-experimental design depends on selection bias • External validity/ transferability depends on comparability of framework conditions	• Enables impact attribution at target group level, given a sufficient sample size
Propensity score matching	Micro level	• Discriminable treatment group • Individual characteristics relevant for the treatment effect (covariates) must be observable	• Ex-ante and ex-post data from treatment and comparison group • Data collection must include covariates	• Increases the costs of quasi-experimental designs as usually larger samples are necessary in order to identify sufficient matches • Comparison group sample must be considerably larger	• Internal validity depends on completeness of covariates • External validity/ transferability depends on comparability of framework conditions	• Same as above • Application in quasi-experimental designs can contribute to improved quality of findings
Pipeline approach	Micro and meso (institutional level)	• Project needs to be implemented in phases • Treatment groups of each phase must be comparable	• Ex-ante and ex-post data from each group	• In principle comparable to quasi-experimental designs, whereby costs increase with every round of data collection	• Internal validity of quasi-experimental design depends on comparability of groups • External validity depends on comparability of framework conditions	• Same as above • Allows for identifying time-variant effects if applied according to quasi-experimental design
Regression Discontinuity Design	Micro level	• Treatment group must be selected according to a specified criterion	• Sufficient number of comparable cases • Larger sample size than for experimental/ quasi-experimental	• Comparable with quasi-experimental design, mostly depending on sample size	• Internal validity restricted to comparable cases	• Enables impact measurement for individuals who have similar characteristics
Time series	All levels	• Panel: focusing on individuals, households, organisations • TSCS: focusing on sectors, countries, regions	• Panel: large sample size, few repeated data collections • TSCS: sample size irrelevant, large number of repeated data collections	• Depending on data availability, individual data collections can be implemented; quite cost-efficient • Costs increase with length of panel and sample size	• Panel: validity depends on compliance of sample size with statistical requirements (cf. 3.3.1) • TSCS: validity is restricted to sample	• Provides reliable assessments of the contribution of an intervention to an observed change • Makes it possible to control time-variant influences
Structural Equation Modelling	Meso and macro (system) level	• Focusing on entire sectors, countries, regions	• Statistical and/ or empirical data for each model- relevant construct	• Depending on data availability, can be implemented cost-efficiently if combined with statistical data	• Validity depends on model fit (i.e. to what extent the endogenous construct is explained by the model)	• Enables contribution assessment and time-variant and invariant influences

Table 4 Example of question types

Question type	Question	Answer options
Dichotomous question	What is your gender?	○ male ○ female
... with nominal scaled answer options	Where do you come from?	○ area 1 ○ area 2 ○ area 3 ○ other area: _____
... with ordinal scaled answer options	How do you assess the development of your livelihood situation in the last year?	○ Deteriorated considerably ○ Deteriorated slightly ○ Did not change at all ○ Improved slightly ○ Improved significantly
... with a metric scaled answer option	What is your monthly income?	____ €/month

It has to be added that for ordinal-scaled answer options that reflect a rating (e.g. satisfaction, quality) it is important to 'balance' the answer options, meaning that an equal number of positive as negative answer options must be provided in order to avoid biased findings.

Besides single-choice questions, there are also multiple-choice questions, for example:

What are your income sources (multiple answers possible):

☐ Farming ☐ Livestock ☐ Fishery
☐ Temporary work ☐ Other: ..... ☐ None

When analysing multiple-choice question data, it has to be taken into account that it might be necessary (e.g. for calculating average means) to transform the data into separate 'dummy variables' with dichotomous scales for each answer option (i.e. yes/no for farming, ... for livestock, etc.).

When a survey is conducted, its objectives and the information required to meet these objectives need to be considered first. For this, it makes sense to ask the following questions:

Questions for identifying the survey objectives

- Which indicators need to be measured?
- Who will use the information and how? Which decisions are depending on the survey findings?
- Are there any particular expectations about the survey findings? When are the findings of the survey considered positive or negative? Or: what would be a good or a bad finding?
- Which information can the respondents provide? How does this match the information needs?
- Which factors could have a negative impact on data collection? Do the respondents have and remember the necessary information? Might they be reluctant to answer? Do they have the (verbal, cognitive) capacities to answer correctly?

Photo: © GIZ/Alan Walsch

Selecting a survey approach

After having defined the survey's objectives and information needs, the survey approach can be specified giving consideration to the available and required resources. The required financial, human and time resources mainly depend on the size of the population that has been defined (as the target or comparison group), its spatial distribution and the logistical framework conditions in the area where the survey is implemented. Furthermore, the length and complexity of the questionnaire considerably influences your resource demands. In order to get a better idea of how much time is needed, it may be helpful to prepare a list of the individual tasks and estimated time needed for each task, such as clarifying the survey objectives and sampling approach, designing and testing the questionnaire, collecting, managing and analysing the data, and reporting/communicating the survey findings. The following table shows what such a to-do list might look like:

Table 5 Exemplary to-do list for preparing a survey

Task	When	Who	Requirements
Clarification of survey objective	First two weeks in January	Project staff	Results framework, sample selection
Development of draft questionnaire	By 31 January	Evaluator in coordination with project staff	Clarification of information needs, development of survey items
Pre-test of questionnaire	1 to 10 February	Evaluator together with enumerator team	Logistical preparation, information of target groups, accessibility of target groups
Finalisation of questionnaire	By 15 February	Evaluator	Consolidated feedback from the pre-test
Implementation of survey	16 to 31 February	Enumerators coordinated by evaluator	Logistical support by partner staff, etc.
Data entry	1 to 5 April	Enumerators coordinated by evaluator	Availability of technical infrastructure (notebooks, software, etc.)
Data quality check	6 April	Evaluator	Full data set
Data analysis	7 to 15 April	Evaluator	Quality-assured data set
Integration of survey findings into evaluation report	16 to 30 April	Evaluator	Data analysis findings
Presentation of findings	1 May	Evaluator and project staff	Technical infrastructure for presentation (notebook, projector, etc.)

On the basis of such a timetable, the costs of the survey can be estimated. These usually comprise logistical costs (e.g. for travel/transportation, accommodation, per diems), expenses for survey materials and the remuneration of the survey team.

Surveys are usually carried out face-to-face or electronically mediated (e.g. by telephone, email or a web-based application). Both face-to-face and electronically mediated surveys can be self-administered (i.e. the respondent fills out the questionnaire) or assisted (i.e. the surveyor fills out the questionnaire or at least supports the respondent in doing so). As all options have their advantages and disadvantages, aspects such as the complexity and comprehensibility of the survey topics, the skills and accessibility of the potential respondents and the available time, budget and human resources need to be considered. Particular attention has to be paid to the aspect of the confidentiality of the information and the anonymity of the respondent when personal data (e.g. age, income, health status) are collected. Ensuring anonymity is often an issue when it is necessary to be able to trace responses to individuals, such as in panel surveys (cf. 3.1.5) or when third parties (e.g. project staff, beneficiaries) are tasked with data collection. In those cases it might be necessary to separate the personal data from the survey data and link the data sets using a list of identification codes, which can only be accessed by the people who are tasked with analysing the data.

Selecting a sample

Due to time and budget constraints, it is often not feasible to conduct a full population survey, i.e. collect data from every potentially relevant respondent (e.g. beneficiary, project participant). In such a case, it is necessary to collect a sample, i.e. to select respondents from the population. However, the findings of a sample population might be different to those from a survey of the entire population. These disparate findings are induced by the so-called selection bias (see also 3.1.1), which can be reduced by various sampling techniques. The common objective of these techniques is to approximate the sample characteristics to the characteristics of the full population relevant for the survey (e.g. the share of male and female respondents, or the share of respondents from rural and urban areas). Doing so is necessary in order to draw conclusions from the survey findings with reference to that population, bearing in mind certain statistical quality criteria as further outlined below.

In order to draw such conclusions, the size and distribution of the population for which the sample should provide representative findings needs to be identified. For instance, if the survey is to be used to identify the impact of a certain intervention (e.g. nationwide information campaign) on the behaviour of target groups (e.g. farmers adopting a new farming method) in different provinces, then it needs to be determined who constitutes the basic population (i.e. all farmers of the country) and how it is distributed among the provinces (e.g. number of farmers in each province).

Once the basic population has been identified, two sampling strategies are available: probability (i.e. randomised) and non-probability (i.e. non-randomised) sampling. Probability sampling means that each individual of a given population has the same known, non-zero probability of being included in the sample, i.e. it is randomly selected. Non-probability sampling means that the chance of being selected for the survey differs among the population (e.g. due to accessibility or ability to participate). In the latter case, findings based on the sample may not be generalised without giving further thought to how to establish representativeness [23] otherwise (see below).

Only **probability sampling** (i.e. random sampling) makes it possible to calculate statistical parameters that inform use of the true value of a parameter in a basic population.

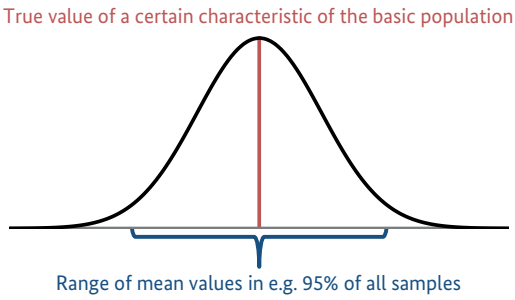
Probability sampling

If probability sampling is applied, it is possible to draw statistically valid conclusions from the survey findings for a population bearing in mind three statistical parameters: the aspired **margin of error**, the **level of confidence** and the **degree of variability.** These criteria are briefly explained below.

The margin of error (also called 'sampling error' or 'level of precision') indicates the maximum difference of the value of a certain characteristic between the sample and the population, i.e. the range in which the true value of the population is estimated. This difference or range is usually defined in percentage points (e.g. 1%, 5%, 10%). For instance if a 5% margin of error is defined and your sample shows that 50% of the respondents adopted a new farming method, the actual proportion of the basic population who did so is somewhere between 45% and 55%. This is subject to a certain level of confidence, see next paragraph.

The level of confidence expresses the probability with which the actual sample value lies within the before- mentioned margin of error. It is based on the assumption that if a sample is repeatedly (randomly) drawn from a given population, the average value of an observed characteristic in this sample approximates to the true value of this characteristic in this population. Furthermore, the sample values are distributed 'normally' around this true value (i.e. their frequency distribution follows a specific, so-called 'Gaussian' function; see the following figure). In practice, confidence levels of between 90% and 99.9% are usually chosen. A confidence level of 95% for instance means that in 95 out of a 100 samples, the sample value would lie within the margin of error as specified before. In other words, there is a 5% chance that the sample value will deviate from the true population value by more than the defined margin of error.

Figure 8 Distribution of sample mean values in relation to the true population mean value



Finally, the degree of variability refers to the heterogeneity of the population with regard to an observed characteristic, i.e. how this characteristic is distributed in the given population. The more heterogeneous the population, the larger the sample size necessary to reach a certain level of precision. In this context, 50% is the highest degree of variability, because it indicates that half of the population has a certain characteristic (e.g. has adopted a new farming technique) while the other half does not. Annex 5.6 contains an introduction on how to calculate the appropriate sample size bearing in mind the above-mentioned statistical parameters.

Non-probability sampling

As probability sampling requires all respondents to be within reach during data collection, regardless of the required resources, it may not always be feasible to apply such a strategy. In that case non-probability sampling may be applied, which requires the identification of characteristics of the basic population that are relevant for analysis in order to establish representativeness. For instance, if the survey is intended to enable conclusions about the extent to which the impact depends on age, gender, ethnic affiliation, etc., information needs to be gathered about the composition of these characteristics in the population. We should mentioned that while with non-probability sampling it may be possible to

[23] It has to be added that 'representativeness' is not a defined statistical term but rather colloquially used to signify that a sample has been composed according to the prevalence of particular characteristics of the basic population.

develop plausible assumptions about a population, the quality of these assumptions cannot be tested statistically and thus rely solely on the expertise of the researcher. For non-probability sampling, three common techniques may help to approximate the sample composition to the composition of the basic population: **systematisation, stratification** and **clustering**.

Non-probability sampling does not make it possible to calculate statistical parameters.

The **representativeness** of a sample relies on the knowledge about the basic population.

Systematisation is an option when sufficient data about the basic population is available so that in principle each potential respondent can be identified. Ideally a comprehensive list is available that is used to draw the sample according to a particular selection scheme based on the individual data. Such a scheme can for instance be oriented towards a targeted sample size (e.g. every 10th person), on logistical considerations (e.g. one household from each street, village, etc.) or a certain threshold level (e.g. only families with more than three children or below a certain income). Which scheme is applied basically depends on the intervention logic (e.g. activities aiming at the poor) and the intended type of representativeness (e.g. equal spatial representation).

A particular way to receive more specific findings is to stratify a sample. Stratification means that a predefined number of individuals from each analytically relevant sub-group of a population (e.g. age groups, residence areas, gender) is included in the sample. Here too, data about the composition of the basic population is necessary, at least on a global level. The intention is usually to match the sample composition with the composition of the population, i.e. ensure that the proportions of each sub-group in the sample and in the population are equal in order to achieve representativeness (i.e. proportional sampling). However, it may also be necessary to draw a disproportional sample, e.g. due to logistical or budget constraints (e.g. 100 households per region). If that is the case, then the cases for each sub-group need to be weighted reciprocally during data analysis. So if for example a certain group represents 20% of a basic population but has only a share of 10% in the sample, its cases need to be weighted by the factor two during the analysis.

Clustering a sample may be helpful if it can be assumed that the sub-groups of a population show a sufficient degree of intra-group homogeneity, i.e. that the individuals in each sub-group are similar with regard to a certain research question, and inter-group heterogeneity, i.e. that the individuals of different sub-groups differ considerably in that regard. If for example a project focuses on both several urban and rural areas and differences of its effectiveness in these areas are to be analysed, it might be sufficient to conduct a full-population survey in one urban and rural area instead of taking samples in each area, provided the residents in the respective areas show some commonalities that are relevant for the analysis (e.g. high incomes in urban areas, low income in rural areas).

Finally, we should note that other techniques exist for non-probability sampling in addition to those outlined above. However, other techniques such as convenience sampling (i.e. selecting individuals who are easiest to reach) or theoretical sampling (mainly used in explorative qualitative research) may not be applicable in rigorous evaluation designs as it is unlikely that they provide representative data. Hence these techniques are not further discussed here.

We recommend the following books, articles and working papers for further reading on methodological requirements to collect quantitative data by means of surveys.

Fink, A. (2003). How to Sample in Surveys. The Survey Kit, 2. Edition. Thousand Oaks: Sage Publications.

Israel, G.D. (2012). Determining Sample Size. Gainesville: University of Florida.

Photo: © GIZ/Harald Franzen

4 Case study: Urban Management of Internal Migration due to Climate Change

In this last chapter, the outlined methodological options and requirements are discussed by means of a practical case study. The case study is based on a GIZ project entitled 'Urban Management of Internal Migration due to Climate Change' that began in January 2015 and is being implemented on behalf of BMZ in two cities in Bangladesh. The objective of the case study is to illustrate the possibilities for providing empirical evidence about project impacts by applying an elaborated evaluation design that satisfies scientific standards. It will also be investigated which organisational and logistical precautions need to be taken to enable efficient data collection and analysis. After a brief description of the project (4.1), two exemplary designs for measuring its impact are presented (4.2), followed by a draft of their practical implementation (4.3).

4.1 Background and project objectives

In Bangladesh, climate change is considered to be one of the most important challenges. Climate-related internal migration is jeopardising the overall social stability of the country. Climatic changes induce extreme weather events such as floods, extreme precipitation and droughts. In Bangladesh, 40 out of 64 districts are affected by the impacts of climate change. According to estimations, six million people have already migrated due to weather and climate stresses, of whom the majority lives in urban slums. Especially in Khulna and Rajshahi, the proportion of migrants is very high (70% of slum dwellers are migrants). For poor and vulnerable households, migration is a crucial diversification and adaptation strategy. However, this can lead to acute vulnerability and have conflict-exacerbating effects, if migrants who live in urban slums with inadequate infrastructure, no access to basic services or income opportunities, receive hardly public support from local government and administrative structures. It is true that the decision to migrate is multi-causal; that means that it is due to political, social, economic, demographic and environmental factors. However, in the two preselected cities, experts believe that climate and weather-related stress factors play a dominant role in migrants' decision to leave their homes. In Rajshahi, rising temperatures and drought, declining groundwater levels and river flooding threaten the population, whereas in Khulna the population faces storms and storm surges, extreme precipitation, water logging and groundwater salinisation. According to estimates, the urban population (today: 34%) will exceed the rural population (today: 66%) by 2040, if economic opportunities do not reach the countryside.

The **core problem** in the current situation is that so far, there are no needs-oriented measures to improve the living conditions of climate migrants [24] in Khulna and Rajshahi. City master plans do not address the influx of migrants. This problem is caused by a lack of capacity of local governments and municipalities to deal with the challenges of climate-related internal migration. Public investment in the development and expansion of a basic infrastructure to mitigate the pressure on resources such as water, energy, food and housing is insufficient. There is a lack of policies and measures to manage settlement development. Nor is there any data or information base on the movement of climate migrants to cities and their (urgent) needs. This situation leads to an explosion of the population in these urban areas, increasing poverty, congested infrastructure and basic social services, shortages of resources, increasing distribution and opportunity conflicts, and eventually reduced economic development opportunities for the entire population. As urban development is a priority of the 6th Five-Year-Plan 2011–2015 (i.e. infrastructure development, promotion of economic activities), national policies indicate that there is potential for change. However, there is a considerable gap between strategy and implementation.

According to this situation, the **objective** of the project is to **improve the living conditions of climate migrants** in preselected settlements in the divisional capitals Khulna and Rajshahi through needs-oriented measures. [25]

[24] The International Organization for Migration (IOM) produced a definition of climate migrants as 'persons or groups of persons who, for compelling reasons of sudden or progressive changes in the environment that adversely affect their lives or living conditions, are obliged to leave their habitual homes, or choose to do so, either temporarily or permanently, and who move either within their country or abroad'.

[25] Since it is an integrative approach, other inhabitants of the settlements will also benefit from improved basic urban services, vocational training and temporary employment opportunities.

Four **indicators** were defined to achieve the project (module) objective:

- In Khulna and Rajshahi, X inhabitants of Y 'hotspots' with a high concentration of climate migrants have improved access to basic urban services (e.g. water, energy or sewage systems through well drilling, solar home systems or the construction of drainage ditches).
- In Khulna and Rajshahi, X inhabitants of Y hotspots with a high concentration of climate migrants have participated in labour-intensive measures for the expansion and building of a climate-resilient basic infrastructure, of whom Z% are women.
- In Khulna and Rajshahi, in Y hotspots with a high concentration of climate migrants, participants in needs-based general education and vocational training courses increased their income by X%.
- In each city, the city corporations developed and tendered X project proposals for needs-based and labour-intensive measures for the expansion and building of climate-resilient basic urban services (e.g. health, education, water, energy supply, waste water/refuse) as well as for the promotion of local economic development with the support of the Ministry of Social Welfare.

In order to achieve this objective, the project focuses on **three intervention areas**, for which particular **measures** are currently being planned:

1. Providing access to social services to the target groups by improving and extending existing services and adapting them to the needs of the migrants (e.g. finding social housing, providing health and sanitation services), and raising the awareness of social workers for the needs of climate migrants.
2. Providing short-term income opportunities through labour-intensive work in and outside slum areas, and providing access to basic infrastructure by its improvement together with city corporations and civil society.
3. Providing long-term income opportunities by developing and implementing skills development measures in cooperation with training centres and formal and informal SMEs in different sectors.

Various Capacity Development measures will be implemented by long-term and short-term experts, and financial support will be provided to upgrade slums and enhance resilience. Particularly in the third intervention area, cooperation is also planned with the local economy and training providers.

It should be highlighted that the project is the first of its kind in German development cooperation to deal with the issue of climate-induced (internal) migration. The project therefore has an innovative and pilot character and aims to showcase a variety of measures to deal with climate-induced internal migration. The integration of climate change adaptation into urban development planning and the continuation of successful measures will be strengthend by other projects implemented by GIZ, KfW and BGR.

Selection of an impact evaluation design

In order to decide which evaluation design(s) could be suitable for measuring the impact of the project, it makes sense to revisit the table at the end of Section 3.1 that summarises the characteristics and requirements of the discussed designs. Looking first at the required project characteristics, we should consider where the project aims to generate tangible results that can be empirically verified. According to the project description, the measures target climate migrants, who constitute a distinguishable target treatment group as they are a sub-population of the entirety of residents in the intervention areas. This treatment group, however, is not randomised, but rather selected based on predefined criteria such as the settlements ('hotspots') and their socio-economic status ('climate migrants'). Hence an experimental design does not come into consideration. As the project description does not indicate a sequential implementation of the measures, a pipeline approach is also out of the question. However, if the project is scaled up at a later stage, such an approach would still be feasible based on the data gathered by the design proposed below.

Another feature of the treatment group is that it has individual characteristics that are likewise relevant for their selection and the treatment effect. As these characteristics are in principle observable (e.g. economic situation, migration background, family status) both propensity score matching and a regression discontinuity design are in principle feasible. With regard to the long-term measurement of treatment effects and their sustainability beyond the project implementation period, a household panel appears to be suitable. However, according to the description of the framework conditions the migrant population fluctuates considerably. Such a fluctuation will probably result in a very high panel mortality, i.e. participants dropping out from the sample, that would undermine the significance and reliability of the findings. Another feature of the project is that it does not focus solely on directly supporting individual climate migrants, but also aims at improving social services and basic infrastructure, which again should contribute indirectly to the improvement of the living conditions of these migrants and the population in the intervention areas at large. Such a 'systemic' approach calls for the use of structural equation

models, which make it possible to measure the influence of changed framework conditions on a population's livelihood.

Coming next to the data requirements, the selection of particular settlement areas makes it possible to create comparison groups, i.e. migrants who are not supported by the project. Furthermore, as the evaluation is planned before the project activities start, it should be possible to collect baseline data from both the treatment and comparison groups. Since the project description does not provide any indication that measures in the selected settlements will be restricted to a certain sub-population of the potential target groups (i.e. there is no threshold level according to which the beneficiaries are selected), comparison groups need to be selected in spatially remote areas. This again makes it impossible to apply an RDD approach. Nor would an RDD be recommended if such a restriction were made, because the design of the measures facilitates spill-over effects, which will likely bias the comparison findings.
With regard to the system-oriented measures, the development of SEMs should be feasible, provided that statistical or empirical data can be made available for the required models throughout the project implementation (i.e. by the continuous results-based monitoring system). This is required to describe the relevant elements that constitute the framework conditions of the migrant population. The following table summarises the applicability of the different evaluation designs based on the project characteristics:

Table 6 Applicability of evaluation designs

Evaluation design	Required project characteristics	Full-filled	Data requirements	Full-filled
Experimental (RCT)	Distinguishable treatment group	✓	Ex-ante (baseline) and ex-post data available from treatment and control group	✓
	Random assignment to treatment and comparison group	–		
Quasi-experimental with PSM	Distinguishable treatment group	✓		
	Observable individual characteristics relevant for the treatment effect	✓		
Pipeline approach	Sequenced project implementation	–	Ex-ante and ex-post data available each group	n.a.
	Comparable treatment groups in each phase	n.a.		
RDD	Selection of treatment group according to a specified criterion	✓	Sufficient number of comparable cases (close to a threshold level)	–
Time-series	Panel: focusing on individuals, households, organisations	✓	Panel: large sample size, few repeated data collections	–*
	TSCS: focusing on sectors, countries, regions	–	TSCS: sample size irrelevant, large number of repeated data collections	n.a.
SEM	Focusing on entire sectors, countries, regions	✓	Statistical and/or empirical data for each model-relevant construct	✓

** Unlikely to be achieved due to foreseeable panel mortality.*

Since PSMs and SEMs appear to be the most promising approaches to measure the impact of the project, we will go on to outline how the data collection process could be organised and what the related methodological and practical requirements are.

4.2 Practical implementation

Starting with PSMs, it first has to be clarified who constitutes the treatment and comparison group and how appropriate samples can be selected. In the second step, the hypotheses to be tested and the necessary indicators have to be developed. Then we need to consider which covariates have to be collected in order to enable the matching of cases (e.g. age, gender, place of origin, duration of stay in target area, education level, ethnic affiliation, marital status). Finally, the data collection plan and instruments have to be developed.

Based on the project description, it is easy to identify who should belong to the treatment climate and floating migrants in selected settlements as well as other poorest people in Khulna and Rajshahi. Equally easy to identify is who should belong to the comparison group: climate migrants in settlements in these cities, where no project activities are implemented. If such settlements cannot be identified, it would also

be acceptable to choose climate migrants in other cities, who face the same problem, provided these cities are comparable in terms of the characteristics that were decisive for the selection of the project location. As it can be assumed that the targeted population is rather large and the actual proportion presenting the characteristic of interest (improved living conditions) is not known beforehand, the sample size should be calculated according to the first formula outlined in Section 5.6 (annex). The formula approximates a sample size of 400 for infinitely large populations with a level of confidence of 95% and an aspired margin of error of 5%.[26] Considering the conceivable attrition loss during the data collection process, a target gross sample size of 600 for the treatment group and at least 800 to 1,000 for the comparison group (in order to receive sufficient matches) should be aimed at. Depending how the dissimilarity of the socio-economic framework conditions in Khulna and Rajshahi is assessed, it further has to be decided whether it is sufficient to split one sample across both cities – i.e. if they are generally comparable – or if the sample size needs to be doubled – i.e. if the framework conditions differ considerably with regard to the characteristics of interest.

Next, it has to be decided which hypotheses are to be tested and which data needs to be collected. So far the project objective has been operationalised by four indicators, of which three directly refer to the living conditions of climate migrants: (1) access to 'urban basic services', (2) participation 'in labour-intensive measures' and (3) having an 'increased income'. Leaving aside the question whether these indicators cover the entirety of factors that define the quality of 'living conditions' – which however needs to be revisited during the development of the data collection instruments – it has to be clarified what information is needed to verify or disprove their achievement. With regard to the first indicator, such information would comprise, for instance, data on the types of existing services, their general accessibility and finally the actual use of these services by the target group. Of course, such data would have to be collected for both the treatment and comparison group, before and directly after the project implementation and ideally again another three to five years later, in order to assess the sustainability of the project measures.

[26] It should be noted that the proposed level of confidence and margin of error represent the lowest common accepted parameters for inferential statistics. It is up to the evaluation team to decide on stricter values, which however will result in a need to collect considerably larger samples.

Concerning the second indicator, it may seem easy to provide empirical evidence about its achievement, i.e. by comparing the factual number of participants with the targeted number. It is, however, questionable whether the indicator actually makes it possible to measure an impact, since participation in a labour-intensive measure does not say anything about the living conditions of the migrants. Since the outcome of their participation (e.g. the generated income and created infrastructure) can rather be considered to influence their living conditions, it is recommended to revise this indicator.

While the relevance of the third indicator for the impact to be measured is again clear, it must be borne in mind that the findings are likely to be biased by contagion effects, i.e. that income has been increased by sources other than the ones created by the project. Therefore, further qualitative data needs to be collected on the income sources and how and why they were obtained. Furthermore, it has to be remembered that providing information about one's own income may be a sensitive issue. Hence, data on proxy indicators such as private properties, number of meals per day or school attendance of children (etc.) has to be collected.

In order to obtain a comprehensive picture about the net impact of the project, it is further necessary to also gather information about its unintended positive and negative impacts. Therefore, it must be considered who else, apart from the migrants, might be affected by the project activities, which secondary effects could be caused or how the project objectives interfere with partner strategies and/or the activities of other donors. For assessing the effectiveness of the particular project measures and helping to attribute the observed effects, it is also necessary to collect data on the participation of the target group (e.g. in labour-intensive or training measures).

Finally, to enable matching of the treatment and comparison group data, covariates need to be collected that have to comply with the 'stability criterion' outlined in Section 5.3. This could be, for instance, the origin of the migrant, his or her ethic and religious affiliation, age, gender and family status, etc. However, we also have to bear in mind not to use variables such as profession, socio-economic status, type of migrant and so on, as the project measures (e.g. further training, income generation) may influence these characteristics.

The following table summarises the indicators, data requirements, data collection instruments and analysis methods discussed above that would be part of the data collection plan for the migrant survey:

Table 7 Draft (simplified) data collection plan

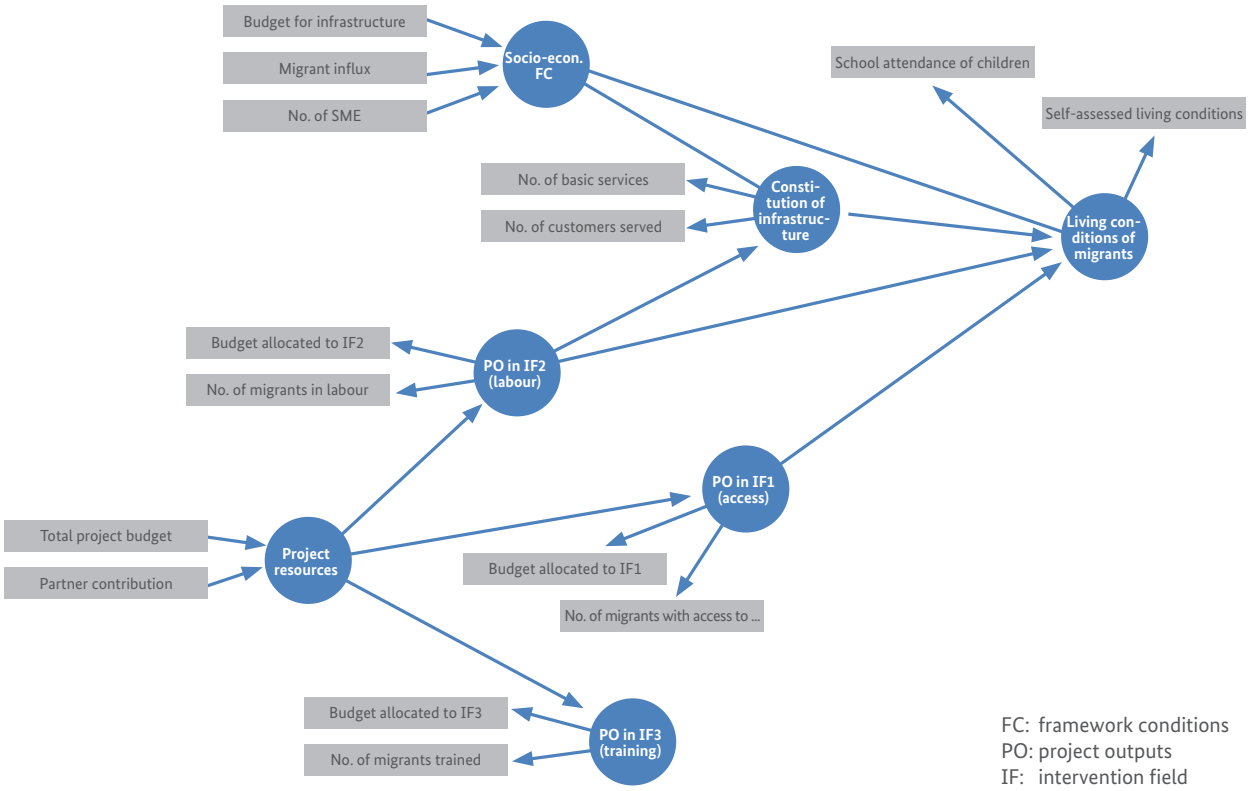
Indicator	Required data	Data collection instrument	Data analysis method
Improved access to basic urban services	Types of services provided	Document analysis, observation (for empirical verification)	Qualitative comparative analysis
	General accessibility of service A, B, C … further operationalized e.g. by: number and distribution of access points, costs per service use, capacities of service in terms of no. of customers per day service can be provided to	Document analysis, observation, migrant survey	Qualitative and quantitative (double-difference) comparative analysis
	Actual use of services by migrants disaggregated by e.g.: types used, frequency of use by type, costs of use in relation to income	Migrant survey, focus group discussion	Quantitative comparative analysis
Increased income	Current income and income sources of migrants (operationalized by proxy indicators)	Migrant survey	Quantitative comparative analysis
➔ Further indicators for measuring the living conditions of the migrants			
➔ Further indicators for measuring the unintended effects of the projects			
➔ Further indicators about participation in project measures			
Covariates	e.g.: origin, ethnic and religious affiliation, age, gender, family status	Migrant survey	Used for calculating the propensity score

The draft shows that there are a number of blank spots that need to be looked into further regarding the operationalisation of impacts, the project activities and the covariates required for matching.

Coming now to the structural equation modelling (SEM), the first step is to identify the constructs that are relevant for the model. Without further knowledge about the framework conditions and project measures, such a model can only be roughly sketched. However, based on the project design, its outputs in the three intervention areas (in terms of the capacities created for basic urban services, training as well as short-term and long-term income opportunities) can be considered as exogenous constructs that aim to influence the living conditions of the migrants. A further exogenous factor would be the general framework conditions, as determined by political strategies (e.g. government budget allocation to infrastructure development) or economic, societal and environmental developments (e.g. regional economic power, number of small and medium enterprises, migrant influx, annual precipitation). Here we should consider whether it makes sense to further split up this construct due to its heterogeneity. The living conditions of the migrants would represent the endogenous construct, whereby the constitution of basic service infrastructure in the settlements could serve as another mediating construct. This is on the one hand influenced by the framework conditions and the project activities and on the other hand has an influence on the living conditions of the migrants. The following figure shows a (highly) simplified illustration of the SEM based on this outline with a few exemplary indicators for each construct:

Photo: © GIZ/Elmar Foellmi

Figure 9 Simplified SEM for the GIZ project [33]



The figure shows that some of the measurement models are formative (project resources, socio-economic framework conditions), while others are reflective (constitution of infrastructure, achievement of project outputs in the fields of intervention and living conditions of the migrants). Of course further constructs, measurement models and structural elements and relationships are conceivable. There is, however, a need to define which elements the model requires in order to adequately reflect the 'real' project situation and to enable measurement of the changes in its fields of intervention. The application of the SEM requires the regular collection of the respective data – at least on an annual basis – to be able to calculate the parameters for assessing the project contributions. Accordingly, the indicators have to be included in the project monitoring system, which should be developed in the course of the baseline study. The latter is the first step of the impact evaluation.

According to the outlined design, the practical implementation of the impact evaluation should start with an in- depth document analysis and the elaboration of a detailed work plan specifying the individual steps of data collection and analysis, and the respective responsibilities. Ideally before starting with the baseline study, an exploratory preliminary mission should be conducted in order to gain further information for complementing the data collection plan (i.e. on the availability of data sources, further needs of the target groups, and other active actors in the intervention area). Furthermore, the availability of secondary data and the accessibility of the migrants for data collection, the logistical requirements and general feasibility of the evaluation design should be investigated.

Once the data collection plan is finalised, the data collection instruments (i.e. interview guidelines, guidelines for focus group discussions, questionnaires) can be developed and tested. In so doing, particular focus has to be placed on the socio-cultural adequacy of the instruments (e.g. appropriate and easy-to-understand language, clear commonly understood categories in closed questions) and – as the survey will most likely be administered – the required qualifications of the interviewers (e.g. use of questionnaire, particularly following the filtering procedure, coding of answers, ability to explain questions if not understood by interviewees, social skills).

[27] The figure was drawn using SmartPLS®, a software for developing variance-based SEMs.

After successful testing and finalisation of the instruments and interviewer training, a sampling procedure should be identified that allows for randomisation, e.g. by random walks through the selected settlements or by randomly selected spots (if reliable maps are available). Data collection should then start, ideally simultaneously with the treatment and comparison group in order to minimise time-induced bias. Provided that the data has been collected according to the previously defined methodological standards, the analysis will provide a picture of the migrants' current living conditions. This will be based not only on the assessment of a few representatives and/or experts, but on a sound statistical foundation that makes it possible to transfer the sample findings to the entire target population. This baseline data will also serve as the comparison base at a later stage for calculating the project impact.

At the end of the baseline study, a framework for the continuous results-based monitoring system should be developed. As the monitoring system will not only be used for continuous reporting on project progress but also for the SEM analysis after project completion, the indicators used must satisfy the information needs of the model and thus not only focus on the project outcomes and impacts but also include data about the framework conditions. The data collection instruments have to be adapted to suit these requirements. Furthermore, the timing of the continuous data collection and the responsibilities have to be coordinated. Given the three-year project term, it may make sense to agree on a six-monthly or even quarterly collection in order to gather sufficient data (about eight to 10 cycles) for robust model calculation. If the measures need to be adapted during project implementation, the indicators and maybe even the model need to be revised likewise.

At the end of the project, a second migrant survey has to be implemented according to the same methodological standards as for the baseline study. In combination with the baseline data, the findings of that survey will be used for impact measurement and attribution as described in Section 3.1.1. Provided that the evaluation team has managed to gather the required data in line with the discussed requirements, the findings should enable valid and reliable attribution of the observed changes in the living conditions of the migrants to the project measures. Due to the broad analytical perspective of the evaluation, it should also be possible to document possible unintended effects as well as the influence of external (confounding) factors.

However, the project measures are not solely designed to improve living conditions in the short term. Long-term effects and sustainability of the project impacts (e.g. climate resilience of built infrastructure) are also of interest. In order to assess such long-term impacts, another evaluation needs to be conducted. It is recommended to conduct such an ex-post evaluation about three to five years after the project has been completed.

As such an ex-post evaluation poses particular practical challenges, particularly regarding the retrieval of the former beneficiaries, access to public authorities or logistical support on site (in case the implementing agencies are not active there anymore), it is of utmost importance to announce such an investigation during the project implementation phase in order to get sufficient support at a later stage. Therefore, all partners and further stakeholders should be made aware of its importance and supported in establishing adequate structures for future data collections. Ideally a Memorandum of Understanding (MoU) could be agreed upon in which the responsibilities are clarified. In return, the partners could be supported in developing their own monitoring system, which again might contribute to their ownership and understanding of the usefulness of collecting such data.

Photo: © GIZ/Ranak Martin

We will go on to present the individual tasks according to the proposed practical implementation of the impact evaluation in chronological order, including a rough estimation of the required time resources for each step:

Figure 10 Draft schedule for impact monitoring and evaluation framework

Winding up, we should point out that the above design and suggestions for its practical implementation can only serve as a rough sketch, which of course needs to be further elaborated during the inception phase of the research. For the sake of convenience, gender-related aspects or further aspects related to the living conditions of the migrants, such as access to health services or education, have not been discussed. Also, the factually available resources for the evaluation and the logistical and practical feasibility of the data collection cannot be assessed without further information about the project concept and the framework conditions. All the more important, then, to start the evaluation with an in-depth scoping phase to ensure appropriate data collection and analysis throughout the evaluation. Valid and reliable evaluation findings can only be produced with high-quality data. In other words: If the data is poor, no design will save the day!

5 Annex

5.1 Overview of current CCA related-project evaluations

N°	Project name	Inst.	Country	Design	Data collection instruments	Methodology
1	Noakhali Rural Development Project	Danida	Bangla-desh	Ex-post facto	Quantitative analysis of project monitoring and contextual data; Qualitative analysis: documentary study, archival work, questionnaire surveys, stakeholder and informant interviews, representative surveys of project components, assessment of buildings, roads and irrigation canals, village surveys and interviews, observation, focus group discussion, case studies	Mixed-method approach
2	Climate Finance Readiness Programme	BMZ GIZ KfW	Global	*		
3	National Adaptation Plan Global Support Programme	UNDP UNEP GEF	Global	*		
4	Africa Adaptation Programme	UNDP WFP UNIDO UNICEF	Africa-wide	*		
5	Strategic Initiative to Address Climate Change in LDCs	UNDP	Global	*		
6	Climate Support Programme (CSP)	GIZ BMU DEA	South Africa	N/A – still ongoing	*	*
7	Public Investment and Climate Change Adaptation (IPACC)	BMUB GIZ APCI Ministries; regional governments of Cusco & Piura	Peru (Cusco and Piura)	N/A – still ongoing	*	*
8	Climate Change Adaptation in Rural Areas of India (CCA RAI)	GIZ, BMZ, ministries, Government of India, governments of federal states	India (Madhya Pradesh, Rajasthan, Tamil Nadu and West Bengal)	*	*	*
9	Sustainable agricultural development (PROAGRO I and II)	BMZ, GIZ SIDA, ministries, VIPFE	Bolivia (Chaco, Northern Potosi, Southern Cochabamba, Valles)	Mid-Term Evaluation	Mainly qualitative evaluation design: analysis of primary and secondary data; semi-structured interviews and surveys with national counterparts, strategic partners and target groups; direct observation; focus group discussions	Mainly qualitative evaluation design

N°	Project name	Inst.	Country	Design	Data collection instruments	Methodology
10	Transboundary water management with the Mekong River Commission	BMZ, GIZ BMUB Mekong River Commission	Cambodia, Laos, Thailand, Vietnam	N/A – still ongoing	*	*
11	Rural Finance and Community Initiatives Project	IFAD	The Gambia	Quasi-experimental	Field-level questionnaires, focus group discussions, interviews with key informants and case studies of the various community-based organisations and their members	For impact assessment, the mission relied on a quantitative survey and participatory rural appraisal (PRA) exercise.
12	Ghana Water Programme	CIDA	Ghana	Before-after	Document review; key-informant interviews; internal benchmarking data from project reviews used for before vs. after; limited resources did not enable ex-post surveys.	
13	Chronic vulnerability to food insecurity	WFP	Kenya	Experimental	Literature review/secondary data analysis; Key Informant Discussions; Participatory Vulnerability Profiles; Field visits; Focus Group Discussions	Study triangulates data from qualitative and quantitative sources. It uses qualitative data from focus group discussions with community members and key-informant interviews with community opinion leaders, and quantitative data involving about 3,000 randomly sampled households.
14	Food and Cash for Assets (FCFA) on Livelihood Resilience in Bangladesh	WFP	Bangladesh	Before-after approach	Document review; On the quantitative front, a household survey was conducted that covered 1,500 households of participants, non-participants and the comparison group. Qualitative data were collected through focus group discussions, asset assessments, key-informant interviews, semi-structured interviews.	Mixed-method approach was adopted. Participatory rural appraisal (PRA)
15	Impact of Food for Assets on Livelihood Resilience in Guatemala	WFP	Guatemala	Quasi-Experimental	Document review, secondary data analysis and institutional Analysis, community profile, asset assessment, institutional analysis, household survey, focus group discussion, semi-structured interviews.	Mixed-method approach. Quantitative survey at the household level; qualitative assessment of impacts at household and community levels; assessment of technical and biophysical assets in each community; social and institutional analysis of networks and linkages at different levels, especially communities
16	Impact of Food for Assets on Livelihood Resilience in Senegal	WFP	Senegal	Quasi-experimental	Document review; household survey; observation; village profiles; gender-disaggregated focus group discussions; semi-structured interviews with major stakeholders; asset assessments	Mixed-method Approach; data triangulation
17	Effects of Climate Variability and Change on Household Food Sufficiency among Small-Scale Farmers of Yatta District	Journal article	Kenya	Experimental	Desk research; use of questionnaires; interviews; focus group discussions; observations; participatory vulnerability profiles; development of indigenous knowledge systems.	Mixed-method approach. Crop production data using Krejcie & Morgan formula commonly used to calculate a sample size (random sampling procedure); coefficients of variation were computed for annual precipitation and then correlated to crop production using Pearson correlation coefficient.

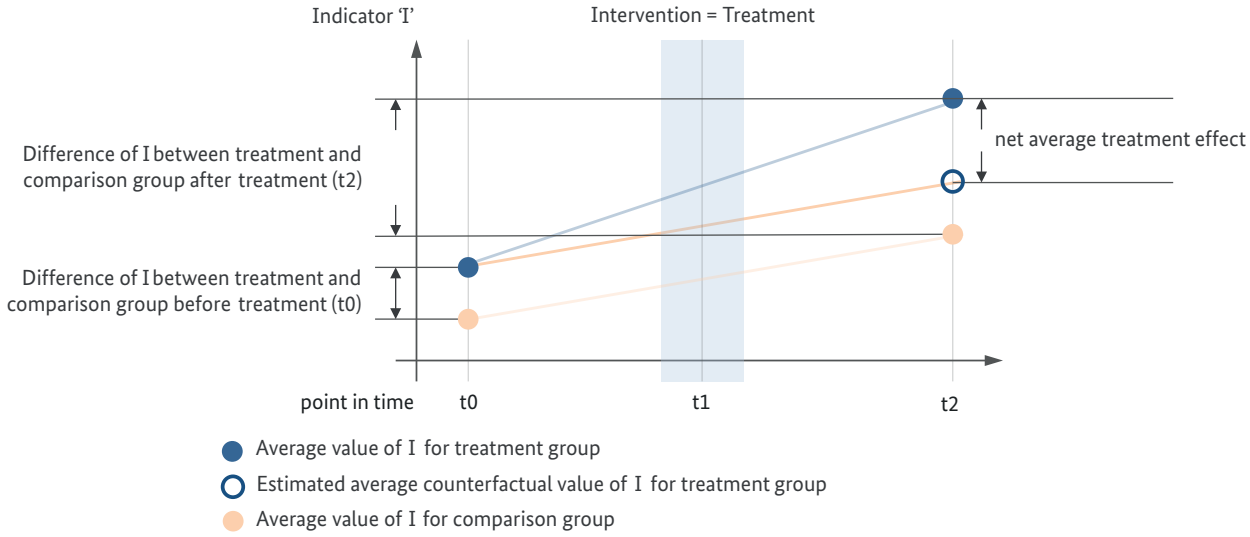
N°	Project name	Inst.	Country	Design	Data collection instruments	Methodology
18	Analysis of vulnerability and resilience to climate change-induced shocks in North Shewa	Journal article	Ethiopia	Experimental	Secondary data analysis surveys; structured questionnaire; interviews; focus group discussion; vulnerability analysis; expert judgment.	Mixed-method approach; principal component analysis correlation with past disaster events; ordered probit regression model; use of fuzzy logic; random sampling procedure
19	Vulnerability of Farming Households to Environmental Degradation in Developing Countries: Evidence from North Central Nigeria	Journal article	Nigeria	Experimental	Survey was used to generate household level data.	Principal component analysis was used to develop vulnerability index for individual household so as to classify households depending on their level of vulnerability to environmental degradation impacts; use of multi-stage random sampling technique; cluster analysis; ordered logit regression model
20	An Index to Determine Vulnerability of Communities in a Coastal Zone: A Case Study of Baler, Aurora	Journal article	Philippines	Experimental	Secondary data analysis; survey (face-to-face and random); questionnaire; observations; mapping/GIS resource mapping activity; workshops	Indicator method; index computations following a balanced weighted average approach; Pearson's correlation coefficient of determination
21	Assessing Household Vulnerability to Climate Change The Case Of Farmers In The Nile Basin	IFPRI	Ethiopia	Experimental	Expert knowledge; use of household-level socio-economic survey; interviews	Mixed-method approach. Indicator and econometric method. Principal component analysis, correlation with past disaster events
22	Cambodia Community-Based Adaptation Programme	UNDP	Cambodia	Experimental	Secondary data analysis; primary data: key-informant interviews, un-structured individual interviews, site visits, observations, semi-structured discussions, structured Focus Group Discussions;	Mixed-method; data triangulation; inductive and deductive approaches using quantitative and qualitative data
23	Strengthening Adaptation Capacities and Reducing the Vulnerability to Climate Change in Burkina Faso	UNDP	Burkina Faso	*	Document review; individual interviews; group discussions; field visits; observation	*
24	National Programme for Managing Climate Change in Malawi and the Malawi Africa Adaptation Programme	UNDP	Malawi	*	Desk review of project; semi-structured interviews; field visits; workshops	Mixed-method approach; data triangulation
25	Adaptation to Climate Change in the Nile Delta through Integrated Coastal Zone Management	UNDP	Egypt	*	Document review and analysis (desktop study); interviews and meetings (face-to-face and by telephone/Skype and email/; questionnaires; data collection in the field (interviews, direct observations)	*
26	The Hill Maize Research Project	USAID	Nepal	Experimental	Secondary data: published and unpublished documents and reports: quantitative and qualitative tools; Primary data: structured and pre-tested questionnaire (household level); semi-structured checklist with a number of open-ended questions (community level); focus group discussion; key-informant interviews using similar semi-structured checklist with open-ended questions; data collected from key informants and focus group surveys were used for descriptive and qualitative analysis, whereas those collected through household survey were used for quantitative analysis.	Mixed-method approach; randomised controlled trial; propensity score matching (PSM).

N°	Project name	Inst.	Country	Design	Data collection instruments	Methodology
27	Small-Scale Irrigation Management Project	JICA	Indonesia	Experimental	Household surveys; direct interviews; Focus Groups Discussions	Mixed-method approach; propensity score matching; regression discontinuity design; estimation strategy; T-test; estimation of average treatment effect; balancing test
28	Vulnerability to tropical storm impacts in coastal areas of the Red River Delta in Viet Nam and the adjacent region	Journal article	Vietnam	*	Quantitative and qualitative surveys; interviews with key informants; semi-structured interviews; discussions; mangrove utilisation survey (quantitative model); focused interviews; visualisation of storm track for various typhoon seasons	Mixed-method approach; participatory rural appraisal
29	Sustainable Soil Management Program	Helvetas Swiss Inter-Cooperation & SDC	Nepal	*	Desk review of relevant documents; field visits; semi-structured interviews; Focus group discussions; observations; separate focus group discussions for discussing issues related to labour migration; appreciative inquiry approach using open-ended questions, storytelling about successes and challenges as well as visioning	*
30	Dry Zone Livelihood Support and Partnership Programme (DZLISPP)	IFAD	Sri Lanka	Quasi-experimental	Qualitative survey (key informant interviews with project staff and government officials; focus group discussions with beneficiaries) and quantitative survey of 2,560 households (both treatment and non-treatment group).	Mixed-method approach
31	Implementing sustainable water resources and wastewater management in Pacific island countries	UNDP	Pacific island countries	*	Evaluation involved an orientation meeting in Nadi, Fiji with Project's Regional Steering Committee, two missions to visit six of the PICs, and Skype and phone interviews with representatives of the remaining PICs and the implementing agencies; construction of theory of change to evaluate the pathway of project's success	Mostly qualitative in nature
32	Supporting Integrated and Comprehensive Approaches to Climate Change Adaptation in Africa - Mainstreaming CCA in the National Sectoral Policies of Tanzania	UNDP	Tanzania	*	Primary and secondary data were collected and analysed; use of semi-structured individual and group interviews	No specific reference but seems to be highly qualitative
33	Climate Change Adaptation Action and Mainstreaming in Mozambique	UNDO	Mozambique	*	Primary and secondary data were collected and analysed; use of semi-structured individual and group interviews	No specific reference but seems to be highly qualitative
34	Integrating Disaster Risk Reduction and Climate Change Adaptation (DRR/CCA) in Local Development Planning and Decision-making Processes	UNDP	Philippines	*	Desk review; key-informant interview (KII) using a questionnaire guide that focuses on particular agencies/institutions vis-à-vis outcome/output; focus group discussion using a questionnaire guide, and study or field visits where observation and interview methods are utilised for more insights.	Mid-term review relies on qualitative information generated through primary data gathering. Opportunities for triangulation were sought. However, not all data/information can be treated in this manner especially as certain key players have not been interviewed despite several attempts.
35	Africa Adaptation Project Namibia Building the Foundation for a National Approach to Climate Change Adaptation in Namibia	UNDP	Namibia	*	Interviews, group interviews and document reviews. Open-ended questionnaires were employed to facilitate interviews. Literature and documents were reviewed to acquire relevant and appropriate information required to answer key evaluation questions and objectives. Observations and case study approaches were used during field visits.	Evaluation applied a mixed methodology of quantitative and qualitative research approaches.

** = No reference | N/A = Not applicable*

5.2 Calculation of net average treatment effect with double-difference approach

Figure 11 Calculation of the net treatment effect in an experimental or quasi-experimental design



The figure shows that the counterfactual value of the indicator for the treatment group is estimated on the basis of its (observed) value for the comparison group and the difference of its values between the groups before the treatment. In other words, it is assumed that without the treatment, the treatment group would have developed in the same way as the comparison group. Accordingly, the net average treatment effect (τ) can be calculated very easily by using the following formula:

$$\tau = (Y^1_{t2} - Y^1_{t0}) - (Y^0_{t0} - Y^0_{t2})$$

Whereby Y represents the value for the indicator of interest for the treatment (exponent = 1) and comparison (exponent = 0) group at the points in time ($t0$ = before and $t2$ = after treatment) when data is being collected, i.e. before and after the treatment. Simply speaking, the difference in the indicator mean value between the treatment and comparison group **before the treatment** is subtracted from its difference between the treatment and comparison group **after the treatment**. Therefore, this approach is also called 'difference-in-difference' or 'double-difference' approach.

5.3 Propensity score matching (PSM)

The propensity score is an index that indicates the conditional probability of an individual's being part of the treatment group. This probability is calculated based on a set of variables that are relevant for both selection for the treatment and the observed treatment effect. The advantage of PSM is that in contrast to other matching techniques, it provides a one-dimensional score for the matching procedure and does not require any distribution assumptions regarding the basic population (cf. 3.3.1). It does, however, require a set of assumptions to be fulfilled, which limit its applicability to some extent. These are:

- Conditional independence assumption (CIA): The treatment effect (i.e. outcome differences between the treatment and the comparison group) has to be independent from the selection, i.e. the effects may only be explained by the treatment; confounding factors have to be excluded. Example: If an intervention aims at increasing the income of a group of people, project beneficiaries need to be subject to the same external economic framework conditions that influence their income as non-beneficiaries. The assumption would be violated, for instance, if a certain occupational group were selected (e.g. only farmers) and compared with a basic population that consists of various occupational groups.
- Stable unit treatment value assumption (SUTVA): The treatment effect of an intervention on one person may not be influenced by the participation of another person in the same intervention. Example: If an intervention aims at providing financial support to a group of people, e.g. by micro-credits, the potentially available amount for each beneficiary must be equal, regardless of the amounts already spent. In other words, the project budget needs to be at least as big as the maximum budget needed if all beneficiaries were to request the maximum individual amount.

- Stability of covariates against the treatment: The covariates that are used to calculate the propensity score must not be influenced by the treatment.
  Example: If an intervention aims at increasing the income of a group of people, then the income of the potential beneficiaries must not be used as a covariate (as it is an intended outcome of the project).
- Common support condition: In order to be suitable for matching, both the treatment and comparison group need to contain individuals with the same or at least a similar propensity score.
- Balancing property: Individuals who have the same propensity score need to be comparable regarding the particular characteristics that were chosen as covariates.

If the project characteristics meet the above requirements, PSM can be applied according to the following sequence [28]:

Figure 12 PSM sequence

Estimating the propensity score

The estimation of the propensity score is based on the selection of the covariates and the estimation procedure. The estimation procedure again depends on the number of groups that need to be included in the analysis. As in evaluations, usually only two groups – the treatment and comparison group – are compared, and only discrete choice models, usually logit or probit regressions [29], are used. While the type of regression only plays a minor role, as both provide very similar findings, the variables that are used to calculate the propensity score are by far more important as they need to fulfil the conditional independence assumption (CIA) in order to provide valid findings. Furthermore, as the propensity score indicates the conditional probability that an individual will receive a treatment, only those variables must be used that reflect this probability. Therefore, it is not correct to generally refer to 'socio-economic covariates', as is still sometimes found in the literature (cf. e.g. OECD 2014). While socio-economic variables may be appropriate for interventions that focus on the poor (i.e. the poorer a person, the more likely he or she is to participate), they may not make any sense for an intervention that focuses on vulnerability (i.e. the more vulnerable a person is, the more likely he or she is to participate). Accordingly, in the latter case variables that reflect the vulnerability of a person (e.g. proximity to a disaster-prone area) need to be taken as covariates. In any case, the selection of covariates needs to be theoretically substantiated.

Once the covariates are selected, the propensity score is calculated by a simple (logit or probit) regression. Here, it is important to note that only metric (i.e. continuous; e.g. income, age) and dichotomous variables (i.e. variables that can have two values, e.g. yes/no, 1/0) can be included. Nominal (i.e. variables that can only have discrete values/categories without ranking, e.g. religion, marital status) and ordinal (i.e. variables with discrete ranked values, e.g. educational level) variables need to be recoded in dichotomous variables for each category before the analysis.

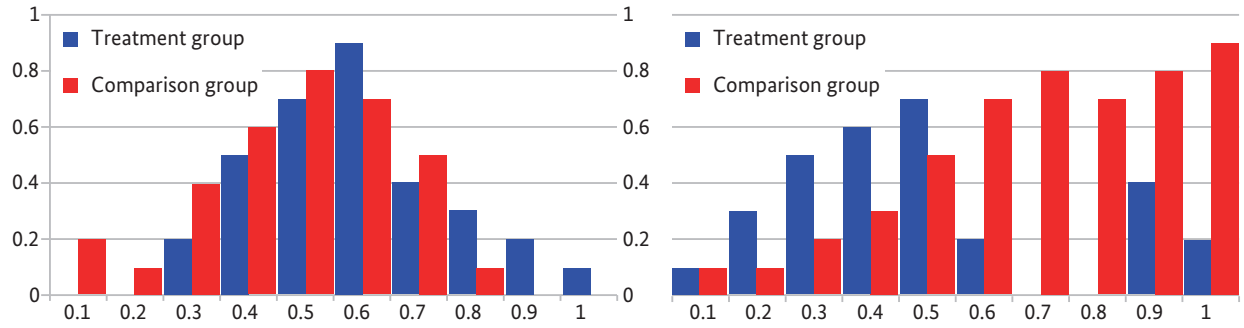
Testing the common support assumption

After the propensity score has been calculated for each case (i.e. person, individual) of the treatment and control group, the so-called **area of common support** (ACS) has to be identified. The ACS comprises the range in which the treatment and comparison group show a similar density of propensity scores. Only cases that lie within this area can be used for the subsequent matching procedure. The ACS can either be defined by cutting off the range at the lowest and highest propensity scores available for both groups or by trimming, i.e. setting a certain density threshold level beneath which cases are discarded. In order to decide how to define the ACS, it makes sense to visualise the distribution of the propensity scores, as the following figure illustrates:

[28] The following descriptions are based on a CEval working paper (no. 19) written by Müller, Christoph 2012.

[29] Regressions are used in statistics for identifying the relationships between one dependent and several independent variables (e.g. income depending on education, gender and age). Logit and probit regressions are usually used when a variable can only have two values (e.g. yes and no, participant and non-participant).

Figure 13 Exemplary illustrations of propensity scores densities



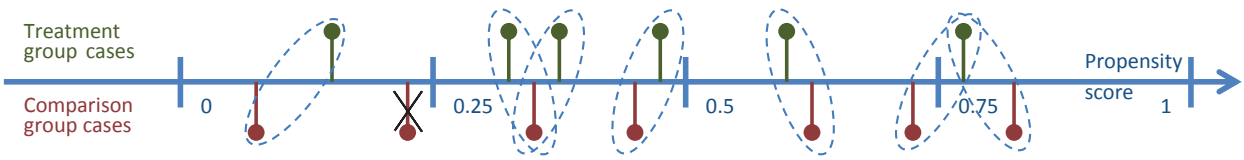
As the illustration in the example on the left shows, it would be appropriate to cut the ACS below 0.3 as there are no cases in the treatment group below this value, and above 0.8 as there are no cases in the comparison group above that value. On the right, both groups feature values in the entire range between 0 and 1. However, the treatment group contains no cases with propensity scores between 0.7 and 0.8, which would make it impossible to create matches in that area. Accordingly, it would make sense to set a threshold level say at about 0.15 beneath which cases are excluded. In that case the ACS would lie between 0.3 and 0.6, and between 0.9 and 1. Note: If no ACS can be identified, PSM cannot be applied!

Selecting a matching algorithm

Once the ACS has been defined, matching then starts with the selection of the matching algorithm. As there are a number of matching algorithms that differ depending on how cases from the comparison group are assigned to cases of the treatment group and the type of weighting, several algorithms are usually applied in order to select the one that is most sensitive for identifying the treatment effect. Three popular matching algorithms are briefly outlined below:

A very popular and fairly simple matching algorithm is the so-called **nearest neighbour matching** (NNM). With NNM, the case from the comparison group that features the most similar propensity score to each case or group of cases of the treatment group is assigned, i.e. the one with the most similar probability of being a member of the treatment group. The approach prevents so-called bad matches, i.e. that cases which are very different – in terms of their characteristics as represented by the covariates – to each other, are compared. Cases from the comparison group can then usually be used repeatedly as a matching partner for cases from the treatment group. It is also recommended to match several cases from the comparison group with each case from the treatment group in order to receive more robust findings (i.e. with less variance). Thus, the comparison group should be considerably larger than the treatment group (cf. 3.3.1). It should be noted, though, that the decision to use cases more than once in order to minimise the variance has the trade-off of a potentially higher bias, as inferior matches may also be used. Figure 14 provides a simplified illustration of how the matches can be identified graphically:

Figure 14 Graphical illustration of nearest neighbour matching



As figure 14 shows, in this example both treatment group and comparison group cases are used repeatedly for matching, and cases that are not the nearest neighbour to any of the cases of the respective other group are discarded.

Another matching algorithm that is frequently used is so-called **kernel matching.** The algorithm uses the weighted means of preferably all individuals of the comparison group to estimate the counterfactual situation of an individual from the treatment group. To apply this matching algorithm, it has to be decided in advance how to weight the individual cases during the matching, i.e. define the probability density function (e.g. a Gaussian function) and how steep the function should be (i.e. the slew rate). Again, here is a graphic to illustrate the matching process:

Figure 15 Graphical illustration of kernel matching

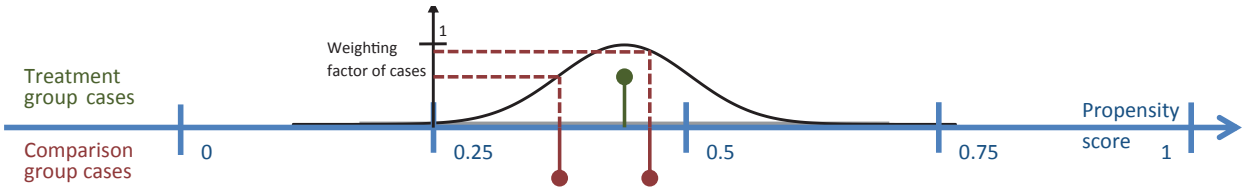


Figure 15 shows that comparison group cases are weighted according to the proximity of their propensity score to the score of the treatment group case they should be matched with.

A matching algorithm that offers a kind of compromise is the so-called **radius matching** for which only those cases of the comparison group are used for the matching that are within a certain predefined range (also called 'caliper') to the treatment group case. The advantage of this algorithm is that the range can be adjusted according to the number of available cases in the comparison group so that it can provide more robust findings, provided sufficient cases are available.

Figure 16 Graphical illustration of radius matching

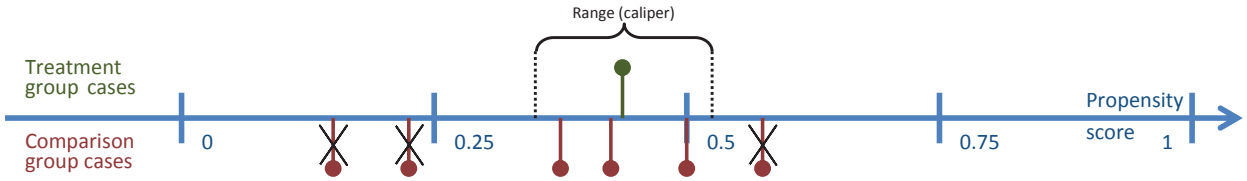


Figure 16 shows that cases that lie outside a predefined range are discarded. Since there are no general rules about how to define this range, we should note that a wider range leads on the one hand to a reduced variance of the findings (i.e. increased robustness); on the other it also introduces more bias as 'not so good' matches are also used. Conversely, a narrower range reduces the bias but increases the variance. Thus, there is always a trade-off between the robustness and bias of the findings.

Testing the quality of matching and calculating the treatment effects

When the matching procedures have been completed, it needs to be verified once again whether the above-mentioned balancing property is fulfilled. A common method for doing this is to check if the difference in the covariates between the treatment and comparison group has been substantially reduced, which can be (amongst others) done by calculating 'Cohens d', an index for comparing differences in mean values, before and after matching. However, further methods are available for verifying the balancing property that cannot be discussed in detail

here but can be found in the recommended literature at the end of this section.

After all these steps have been taken, finally the treatment effects can be assessed in different ways. First of all, of course, the sign and the magnitude of the mean differences in the outcome variables after matching provide information about the direction and size of the relevant effect. Furthermore, one-sided T-tests can be performed for checking if the treatment effects are significant.

Sensitivity analysis

The last step to provide for a distinct attribution of the observed effects to an intervention is to assess the robustness of the findings against unobserved confounding factors. To do this, a sensitivity analysis needs to be conducted, e.g. based on the so-called **Rosenbaum Bounds** (RBs) that make it possible to estimate how strongly a confounding factor would need to bias the selection process in order to compromise the robustness of the estimated treatment effect. RBs represent the limits within which treatment effects are considered significant. Estimation of the RBs is based on the assumption that with an increasing positive or negative selection, treatment effects are respectively over- or underrated. This also means that the probability of a positive treatment effect decreases with positive selection and increases with a negative selection. There are further indexes such as the Hodges-Lehmann estimator (i.e. for estimating the minimum treatment effect to be considered as significant) that can be used to further hedge the findings of the sensitivity analysis.

Unfortunately, since the calculation of these estimators requires a substantial statistical background, they cannot be discussed here in detail. However, statistical programs such as SPSS® or STATA® provide particular modules for these estimators, which make it possible also for laypeople to perform the required calculations (provided they have understood the sense and benefit of such an analysis).

5.4 Fixed-effects, random-effects models and time-series cross-section analysis

Fixed-effects models (FEM) are used to estimate the impact of a particular independent variable (e.g. treatment, framework condition) that varies over time on individual outcomes (e.g. income, health status). It is based on the assumption that these outcomes correlate with that independent variable. Thus FEM can reveal the relationship between predictor and outcome variables within an analytical entity, giving consideration to time-invariant individual characteristics that potentially affect the impact of these predictor variables (e.g. profession on income or gender on nutrition). These characteristics are considered to be time-invariant, unique to the analytical entity and not to be correlated with other individual characteristics. In that case the treatment effect on the outcome variable of interest can be estimated by regressing the individual outcomes as calculated with the following formula:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$$

Whereby y_{it} stands for the dependent outcome variable of interest with i representing its analytical entity and t the point in time when the data is collected. x'_{it} is the value of the independent variable for this entity at that time, β the coefficient for the independent variable, α_i the invariant individual characteristic that affects the outcome and ε_{it} the error term for other unobserved factors that vary over time. With that model, all time- invariant differences between the analytical entities are controlled so that it can be used to identify the causes of observed changes of an analytical entity over time.

In contrast to FEM, the random-effects model (REM) is based on the assumption that variations in the individual outcomes of the entities are random and uncorrelated with the predictor variables. Accordingly, the REM should be used if it is assumed that differences across entities influence the dependent outcome variable. Thus the formula has to be amended by adding a second error term, u_{it} that expresses this difference:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it} + u_{it}$$

As the application of REM requires that the second error term does not correlate with the predictor variable, individual characteristics that potentially have an influence on the predictor variables need to be specified. Accordingly omitting relevant variables in the model may lead to biased findings.

A number of tests are available for deciding which model to use in a panel analysis, such as the so-called Hausman test, which tests whether correlates with the regressors, the Pasaran cross-sectional dependence test or serial correlation tests. As the test procedures cannot be discussed in detail in this Guidebook, further references are provided in the literature list at the end of this section. Furthermore, it has to be added that for both STATA® and SPSS®, do-file respectively syntax templates are available that can perform such tests on suitable data sets.

Another 'panel-like' approach for long-term impact assessment is the so-called **time-series cross-section (TSCS) analysis.** Similar to the panel analysis, the TSCS analysis is based on data being collected over time from the same analytical units. However, in contrast to the former approach, conclusions made with TSCS analysis are limited to the sample, i.e. are not transferable to another population without further ado. Another difference is that while the panel analysis requires a large sample but also works with a few observations (at least three), TSCS Analysis works with a small number of analytical entities but requires more observations (at least about 10) in order to provide robust findings. Due to these features, TSCS analysis is mainly used for policy studies with macro data and analytical entities such as sectors, countries or regions. As one the one hand the statistical fundamentals of TSCS analysis are quite complex and on the other hand, the approach is not (yet) very popular for project evaluations, it will not be discussed here in detail.

However, the following literature reference list gives some recommendations for further reading on this subject too:

Beck, N. (2006). Time-Series–Cross-Section Methods. New York: New York University.

Beck, N. (2001). Time-Series–Cross-Section Data: What Have We Learned in the Last Few Years? San Diego: University of California.

Beck, N., Katz, J.N. (1995). What to do (and not to do) with Time-Series Cross-Section Data. The American Political Science Review, 89(3):634 – 647.

Girosi, F., King, G. (2001). Time Series Cross-Sectional Analyses with Different Explanatory Variables in Each Cross-Section. Cambridge: Harvard University.

Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data. Cambridge: The MIT Press.

5.5 Structural equation modelling

As outlined in Section 3.1.6, structural equation models consist of latent constructs and empirically measurable indicators. With regard to the operationalisation of the constructs/variables, two types of measurement models can be distinguished: models based on reflective indicators and those based on formative indicators. While reflective indicators are defined by their construct (e.g. construct: sportiness → indicator: time needed to run 100 metres), formative indicators define the construct (e.g. indicator: number of training modules per month → construct: sportiness). Accordingly, reflective indicators have to correlate with each other while formative ones do not necessarily have to do so (e.g. time needed to run 100; 500; 1,000 metres vs. number of running modules, number of swimming modules, etc.). Another difference between reflective and formative constructs is that reflective indicators are replaceable (e.g. time taken to run 100 metres, time taken to cycle 1 km), while an exchange of formative indicators would result in a modification of the construct (e.g. number of maths lessons instead of training modules).

Before the causal relationship between the constructs can be analysed, the validity of the measurement models, i.e. how well the indicators operationalise a construct, needs to be checked. Due to the above-mentioned differences, the analysis of reflective and formative indicators differs. As reflective measurement models correspond to an explorative factor analysis (main component analysis), their validity is assessed on the basis of the factor loadings (should be ≥0.707), their significance, its reliability (composite reliability r should be >0.7) and the average explained variance of the construct (should be >0.5). Furthermore, a reflective measurement model is considered valid if the explained variance of a construct is higher than each squared correlation between this and any other construct ('Fornell-Larker-Criterion', i.e. discriminatory validity). Formative measurement models correspond to a multiple regression. Accordingly, their quality is assessed on the basis of the factor weights, their significance and the co-linearity of the indicators (i.e. the so-called variance inflation factor should be <10).

Concerning the number of indicators required to describe a construct, it has to be considered that there is a trade-off between the reliability of the measurement and the danger of creating artefacts, which biases the construct (e.g. overrepresentation of physical indicators when measuring the health status). Furthermore, the sample size should be considerably larger than the number of indicators.

While the measurement model specifies the relationship between the latent constructs and their indicators, the structural model describes the influence of the constructs on each other. These influences are graphically marked by arrows, which represent assumed causal relationships. The direction of the arrow indicates the direction of causality, i.e. defines what is the cause and what is the effect. In principle, both the number of constructs and the number of (estimated) causal relations are unlimited. This also means that a construct can be both affected

by one or more constructs and affect another one or more others.

The relationship between two constructs is expressed by the so-called path coefficient β, which ranges from -1 (perfect negative relationship) to 1 (perfect positive relationship). A value of 0 may indicate that there is no statistical relationship between two constructs. While a high value of β is preferable, in practice values higher than 0.4 (respectively <-0.4) can be considered as very high. It has to be added that SEM makes it possible to calculate not only direct effects but also indirect and total effects.

The path coefficient can be calculated in two different ways, i.e. by means of co-variance analysis or variance analysis, each having some advantages and disadvantages. Although the theoretical foundations of the two methods cannot be discussed here in detail, the main aspects that help decide which one to choose are summarised in the following table:

Table 8 Decisive aspects for choosing co-variance or variance analysis [30]

Co-variance analysis	Variance analysis
Preferred method for analysing established theories and testing hypotheses	Preferred method in explorative studies, i.e. if the causal relations between the constructs are not yet substantiated
Works best with large sample sizes (n >100)	Also works with smaller sample size (n >30)
Method of choice if statistical quality criteria1 are preferred for assessing the validity of the model	Method of choice if the model is validated nomologically
Available software: e.g. AMOS®, LISREL®	Available software: e.g. SmartPLS®

As the table shows, variance analysis might be the preferred choice for evaluations since they are usually explorative in nature and constructs that characterise the objectives of an intervention are more likely to be based on formative measurement models. The smaller required sample size also speaks for the use of variance analysis. However, the fact that no statistical indexes are available to assess the validity of variance analysis-based structural models means that the researcher applying this method needs to have comprehensive knowledge about the system conditions and the causal linkages between its elements.

Nevertheless, variance analysis also provides indices for assessing the contribution of an intervention to an observed effect such as the above-mentioned path coefficient (should be >0.2), its significance and the effect strength (f^2; weak: >0.02; average: >0.15; strong: >0.35). Furthermore, the model fit (R^2; small: >0.2; average: >0.33; substantial: >0.67) indicates the informative value of the model (i.e. how far the independent constructs explain the variance of the individual constructs).

As also shown in the table, meanwhile a number of user-friendly computer programs are available for calculating structural equation models at least on the basis of variance analysis. While in principle even people with a limited methodological background can operate such software (at least it provides plausible findings), it is not recommended to do so because incorrect theoretical assumptions or setting wrong parameters for calculating the quality criteria can lead to biased or even completely false findings.

5.6 Calculating sample sizes for probability sampling

This section outlines three different ways to calculate the sample size depending on the available information about the basic population. In the first example, the size of the basic population is known but the true value of a certain dichotomous characteristic (i.e. can have two values, e.g. whether or not a beneficiary applied an improved farming technique) is not. In this case the required sample size can be calculated simply according to the formula:

$$n = \frac{N}{1 + (N \cdot e^2)}$$

Whereby n is the required sample size, N is the size of the basic population and e^2 is the desired margin of error. So assuming the basic population amounts to 1,000 persons, and the desired margin of error is 0.05, then the minimum sample size would be:

$$\frac{1{,}000}{1 + (1{,}000 \cdot 0.05^2)} = \frac{1{,}000}{1 + 2.5} \approx 286$$

This means data from a minimum of 286 randomly selected respondents would be needed for the survey.

[30] Adapted from Jahn, 2007:16.

An alternative way to calculate the minimum sample size for large populations – even if the exact size of the basic population is not known – is illustrated next. In this case, it is advantageous to have a good estimate of the true value of an observed characteristic of interest (e.g. gender distribution):

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{e^2}$$

Thereby Z is the area under the normal curve corresponding to the defined level of confidence and p is the true share of the population that displays a certain characteristic (e.g. is female).

For most common levels of confidence Z equals as follows:

90%: Z = 1.645
95%: Z = 1.960
99%: Z = 2.575
99.9%: Z = 3.290

e.g. for a population in which 48% are female and the desired margin of error is 0.05, the minimum sample size would be:

$$n = \frac{1.96^2 \cdot 0.48 \cdot (1 - 0.48)}{0.05^2} = \frac{0.9589}{0.0025} \approx 384$$

However, if the actual share of the basic population that features a certain characteristic of interest (e.g. percentage of beneficiaries who have adopted a new farming technique or not) is not known, it is always safe to assume a 50% share as then the term $p \cdot (1 - p)$ and accordingly the entire formula reaches its maximum.

If the size of the basic population is known and if it is smaller than approximately 50,000 one can still use the latter formula in a first step, but reduce the sample size in a second step by applying the following correction term:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Whereby n_0 is the required sample size as calculated above. Assuming that the basic population only comprises 1,000 persons, then the minimum sample size would be:

$$n = \frac{384}{1 + \frac{(384 - 1)}{1{,}000}} = \frac{384}{1.383} \approx 278$$

So according to this correction term, it is acceptable to reduce the sample size by 106 respondents, which equals a reduction of more than a quarter.

It must, however, be kept in mind that the sample sizes provided in these formulas corresponds to the actual number of necessary completed questionnaires. Accordingly, when planning a survey, the non-response rate has to be considered as well. The rate indicates the share of those who do not participate in the survey despite being selected for the sample. This needs to be compensated to assure that the sample remains large enough to achieve the desired level of confidence and margin of error. So if for example a sample size of 500 is required and it is assumed that only 80% might answer, the number of respondents needs to be increased to at least 625.

Furthermore, it has to be considered that if the survey is part of an experimental or quasi-experimental design, data need to be collected not only for the treatment group but also for a comparison group. Therefore, when applying PSM, the comparison group sample needs to be considerably bigger than the treatment group sample – ideally about two to three times the size – in order to assure that sufficient cases are available for matching during data analysis (cf. 3.1.2).

5.7 Literature

Bours, D. et al. (2014). Guidance Note 1: Twelve Reasons why Climate Change Adaptation M&E Is Challenging. http://www.seachangecop.org/sites/default/files/documents/2014%2001%20SEA%20Change%20UKCIP%20GN1%2012%20Reasons%20why%20CCA%20MandE%20is%20challenging.pdf.

Bours, D. et al. (2014). Guidance Note 2: Selecting Indicators for Climate Change Adaptation Programming. http://www.seachangecop.org/sites/default/files/documents/2014%2001%20SEA%20Change%20UKCIP%20GN2%20Selecting%20indicators%20for%20CCA_0.pdf.

Bours, D. et al. (2014). Guidance Note 3: Theory of Change Approach to Climate Change Adaptation Programming. http://www.seachangecop.org/sites/default/files/documents/2014%2002%20SEA%20Change%20UKCIP%20GN3%20ToC%20approach%20to%20CCA%20programming.pdf.

Bours, D. et al. (2014). Monitoring and Evaluation for Climate Change Adaptation and Resilience: A Synthesis of Tools, Frameworks and Approaches. Second Edition. http://www.seachangecop.org/sites/default/files/documents/2014%2005%2015%20SEA%20Change%20UKCIP%20Synthesis%20Report%202nd%20edition.pdf.

Bours, D. et al. (2014). Design, Monitoring, and Evaluation in a Changing Climate: Lessons Learned from Agriculture and Food Security Programme Evaluations in Asia. http://www.seachangecop.org/sites/default/files/documents/2014%2005%20SEA%20Change%20UKCIP%20ER1%20Agriculture%20and%20food%20security_0.pdf.

Bours, D. et al. (2014). International and donor agency portfolio evaluations: trends in monitoring and evaluation of climate change adaptation programs. http://www.seachangecop.org/sites/default/files/documents/2014%2006%20SEA%20Change%20UKCIP%20ER2%20CCA%20MandE%20trends%20from%20porfolio%20evaluations_0.pdf.

Conway, D. & Mustelin, J. (2014). Strategies for improving adaptation practice in developing countries. Nature Climate Change, 4:339-342.

Cordero, C. (2014). Case Study: Working with the Vulnerability Sourcebook in Chullcu Mayu, Bolivia. News. GIZ. http://www.giz.de/expertise/html/15486.html.

Dinshaw, A. et al. (2014). Monitoring and Evaluation of Climate Change Adaptation: Methodological Approaches. OECD Environment Working Paper, No. 74, OECD Publishing.

Ford, J.D. et al. (2013). How to track Adaptation to Climate Change: A typology of Approaches for National-Level Application. Ecology and Society, 18(3):40.

Ford, J. D. & Berrang-Ford, L. (2015). The 4Cs of adaptation tracking: consistency, comparability, comprehensiveness, coherency. Mitigation and Adaptation Strategies for Global Change., 1-21.

Fischerova, G. (no date). Strategic Initiatives to Address Climate Change in LDCs (Boots on the Ground). Integrating Climate Change Risks into Development Planning and Programming. UNDP. Bratislava Regional Centre (PowerPoint Presentation).

Förch, W. et al. (2014). Back to Baselines. Measuring Change and Sharing Data. Agriculture and Food Security 2014, 3:13. http://www.agricultureandfoodsecurity.com/content/pdf/2048-7010-3-13.pdf.

Garcia, J.R. & Zazueta, A. (2015). Going Beyond Mixed Methods: A Systems Perspective for Asking the Right Questions. Institute of Development Studies. http://onlinelibrary.wiley.com/doi/10.1111/1759-5436.12119/epdf.

German Federal Ministry for Economic Cooperation and Development (BMZ) (2013). Climate Finance Readiness Programme. Early Action for Ambitious Goals. http://www.giz.de/expertise/downloads/giz2013-en-climate-finance-readiness-flyer.pdf.

GIZ (2012). Disaster Risk Management and adaptation to climate change. Experience from German Development Cooperation. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. http://www.ipacc.pe/doc/en-disaster-risk-management-climate-change.pdf.

GIZ (2013). GIZ's global GCF Readiness Programme. Competence Centre for Climate Change Environment and Climate Change Division. http://www.giz.de/expertise/downloads/giz2013-en-climate-finance-gcfit.pdf.

GIZ (2013). GIZ in South Africa. Programs and Projects. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. https://www.giz.de/de/downloads/giz2013-en-giz-programs-projects-south-africa.pdf.

GIZ (2013). Guidelines on designing and using a results-based monitoring system (RBM system). https://www.giz2013-0110en-results-based-monitoring-system.pdf.

GIZ (2014). Transboundary Water Management. Support to Cooperation on Shared Waters. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. http://www.giz.de/expertise/downloads/giz2014-en-mappe-transboundary-water-management.pdf.

GIZ (2015a), The Sustainable Agricultural Programme (PROAGRO I & II) http://proagro-bolivia.org (accessed on 22.10.2015).

GIZ (2015b), Public Investment and Climate Change Adaptation (IPACC) project, http://www.giz.de/en/worldwide/13314.html (accessed on 22.10.2015).

GIZ (2015c), Climate Support Programme (CSP), http://www.giz.de/en/worldwide/17807.html (accessed on 22.10.2015).

Hammill, A. et al. (2013). Comparative Analysis of climate change vulnerability assessments: Lessons from Tunisia and Indonesia. Competence Centre for Climate Change. Deutsche Gesellschaft für Internationale Zusammenarbeit. https://gc21.giz.de/ibt/var/app/wp342deP/1443/wp-content/uploads/filebase/va/vulnerability-guides-manuals-reports/Comperative-analysis-of-climate-change-vulnerability-assessments.pdf.

Hedger, M.M. (2008). Desk Review: Evaluation of Adaptation to Climate Change from a Development Perspective. Institute of Development Studies. https://www.climate-eval.org/sites/default/files/events/alexandria/IDS_Report_on_Evaluating_Adaptation_for_GE_publication_version.pdf.

IIED (2015), Capacity Strengthening in Least Developed Countries (LDCs) for Adaptation to Climate Change (CLACC), http://www.iied.org/capacity-strengthening-least-developed-countries-for-adaptation-climate-change-clacc (accessed on 22.10.2015).

Karkoschka, O. et al. (2013). Guidance for Integrating Monitoring and Evaluation of Climate Change Adaptation into Strategies in Mongolia. Deutsche Gesellschaft für Internationale Zusammenarbeit. http://www.seachangecop.org/sites/default/files/documents/2013%2007%20GIZ%20-%20Guidance%20for%20Integrating%20M%26E%20of%20CCA%20into%20Strategies.pdf.

Lamhauge, N. et al. (2012). Monitoring and Evaluation for Adaptation: Lessons from Development Cooperation Agencies. OECD Environment Working Paper, No. 38. OECD Publishing. http://www.oecd-ilibrary.org /docserver/download/5kg20mj6c2bw.pdf?expires=1425238023&id=id&accname=guest&checksum=77BA5587F2D7AD9EF6C22B614A36C4D4.

Leighton, M. et al. (2011). Climate Change and Migration. Rethinking Policies for Adaptation and Disaster Risk Reduction. No. 15/2011. United National University - Institute for Environment and Human Security. https://www.ehs.unu.edu/file/get/8468.

Leiter, T. (2015). Linking Monitoring and Evaluation of Adaptation to Climate Change across Scales: Avenues and Practical Approaches. In: D. Bours, P. Pringle & C. McGinn (Eds.), Monitoring and Evaluation of Climate Change Adaptation: A review of the landscape. New Directions for Evaluation, 147, 117-127.

OECD (2011). Handbook on the OECD-DAC Climate Markers. http://www.oecd.org/dac/stats/48785310.pdf.

OECD (2014). OECD DAC Statistics. Aid to Climate Change Adaptation. March 2014, Version 2. http://www.oecd.org/dac/environment-development/Adaptation-related%20Aid%20Flyer%20-%20March%202014%20v2.pdf.

Olivier, J. et al. (2013). Adaptation made to measure. A Guidebook to the Design and result-based monitoring of climate change adaptation projects. 2nd Edition, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. https://gc21.giz.de/ibt/var/app/wp342deP/1443/wp-content/uploads/filebase/me/me-guides-manuals-reports/GIZ-2013_Adaptation_made_to_measure_second_edition.pdf.

Prowse, M. et al. (2010). Impact Evaluation and Intervention to address Climate Change: A Scoping Study. International Initiative for Impact Evaluation, Working Paper 7. https://www.climate-eval.org/sites/default/files/evaluations/530%20Impact%20Evaluation%20and%20Interventions%20to%20Address%20Climate%20Change%20-%20A%20Scoping%20Study.pdf.

Sanahuja, H. E. (2011). A Framework for Monitoring and Evaluating Adaptation to Climate Change. Community of Practice. Global Environment Facility. https://www.climate-eval.org/sites/default/files/studies/Climate-Eval%20Framework%20for%20Monitoring%20and%20Evaluation%20of%20Adaptation%20to%20Climate%20Change.pdf.

Sara Orleans Reed et al. (2014). Resilience projects as experiments: implementing climate change resilience in Asian cities, climate and development. Published by Taylor & Francis. http://www.tandfonline.com/doi/full/10.1080/17565529.2014.989190#.VPZRBvmG-NM.

Scholze, M. (2013). Country Experiences with Mainstreaming Adaptation – from Programme to National Level. Webinar. https://gc21.giz.de/ibt/var/app/wp342deP/1443/wp-content/uploads/filebase/ms/ mainstreaming_webinare/Country_experiences_with_mainstreaming_adaptation_national_level.pdf.

Schwartzendruber, F. (2014). Evaluation of NRM interventions linked to climate change: A scoping study. Draft version. https://www.climate-eval.org/sites/default/files/studies/NRM-Study-Draft.pdf.

Smit, B. & Pilifosova, O. (2010). Adaptation to Climate Change in the Context of Sustainable Development and Equity. https://www.vie.unu.edu/file/get/9995.pdf.

Smith, J.B., Vogel, J. M. and Cromwell, J. E. (2009). An architecture for government action on adaptation to climate change. An editorial comment. Climatic Change 95:53-61.

UNDP (2009). Thematic Area: Scaling up local and community-based actions UNDP: Community Based Adaptation. https://unfccc.int/files/adaptation/application/pdf/undp_ap_update_sep_09_cba_1_sp.pdf.

UNDP (2011). Achievements in 2011: Strategic Initiative to Address Climate Change in Least Developed Countries. Boots on the Ground Report. http://www.undp.org/content/dam/undp/library/Environment%20and%20Energy/Climate%20Change/2011-Boots-on-the-Ground-Report.pdf.

UNDP (2011). Brochure on Strategic Initiative to Address Climate Change in LDCs. http://www.undp.org/content/dam/undp/library/Environment%20and%20Energy/Climate%20Change/Capacity%20Development/Flyer%20for%20Boots%20on%20the%20Ground_Final_rev%20rc.pdf.

UNDP (2013). UNDP support to LDCs to Access Finance for Adaptation. Basic Facts and Figures. http://undp-alm.org/sites/default/files/downloads/overview_of_undp_support_to_ldcs_to_access_finance_for_adaptation_-_nov_2013.pdf.

UNDP (2015a). Community-Based Wetland Management Project (BIRAM), http://www.undp-alm.org/projects/spa-cba-bangladesh-community-based-wetland-management-project-biram (accessed on 22.10.2015).

UNDP (2015b), National Adaptation Plan Global Support Programme (NAP-GSP), http://www.undp-alm.org/projects/naps-ldcs (accessed on 22.10.2015).

UNDP (2015c), Pacific Integrated Water Resource Management (IWRM) project, http://www.tandfonline.com/doi/pdf/10.1080/19439341003786729 (accessed on 22.10.2015).

Villanueva, P. S. (2011). Learning to Adapt: Monitoring and Evaluation Approaches in Climate Change Adaptation and Disaster Risk Reduction – Challenges, Gaps and Ways Forward. http://r4d.dfid.gov.uk/PDF/Outputs/ClimateChange/SCR-DiscussionPaper9--Learning-to-ADAPT.pdf.

Würtenberger, L. (2013). GCFit. GIZ's Global GCF Readiness Programme. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. http://www.caribank.org/uploads/2013/05/GIZ%C2%B4s-Global-GCF-Readiness-Programm.pdf.

Notes

Published by
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn, Germany

Global Project Effective Adaptation Finance (M&E Adapt)
Friedrich-Ebert-Allee 36 + 40
53113 Bonn
Germany
Tel. +49 (0) 228 44 60 - 0
Fax +49 (0) 228 44 60 - 1766

Dag-Hammarskjöld-Weg 1-5
65760 Eschborn
Germany
Tel. +49 (0) 6196 79 - 0
Fax +49 (0) 61 96 79 - 1115

climate@giz.de
www.giz.de/climate

In cooperation with
UNDP; Center for Evaluation CEVAL

Autors
Stefan Silvestrini, Ines Bellino, Susanne Väth

Design and layout
Ira Olaleye, Eschborn, Germany

Photo credits
Title page: © Ranak Martin

As at
September 2015

On behalf of
Federal Ministry for Economic Cooperation and Development (BMZ)
Special Unit ‘Climate’

Addresses of the BMZ offices
BMZ Bonn
Dahlmannstraße 4
53113 Bonn, Germany
Tel. +49 (0) 228 99 535 - 0
Fax +49 (0) 228 99 535 - 3500

BMZ Berlin
Stresemannstraße 94
10963 Berlin, Germany
Tel. +49 (0) 30 18 535 - 0
Fax +49 (0) 30 18 535 - 2501

poststelle@bmz.bund.de
www.bmz.de