IMPACT EVALUATIONS FOR RETURN AND REINTEGRATION PROGRAMMES

IOM
UN MIGRATION

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the International Organization for Migration (IOM). The designations employed and the presentation of material throughout the publication do not imply expression of any opinion whatsoever on the part of IOM concerning the legal status of any country, territory, city or area, or of its authorities, or concerning its frontiers or boundaries.

IOM is committed to the principle that humane and orderly migration benefits migrants and society. As an intergovernmental organization, IOM acts with its partners in the international community to: assist in meeting the operational challenges of migration; advance understanding of migration issues; encourage social and economic development through migration; and uphold the human dignity and well-being of migrants.

This publication was produced with the financial support of the European Union. The contents of this report are the sole responsibility of the authors and do not necessarily reflect the views of the International Organization for Migration (IOM), nor those of the European Union. Neither IOM nor the European Union make claims – expressed or implied – on the completeness, accuracy and suitability of the information provided through this report. Names and boundaries do not imply official endorsement or acceptance by the International Organization for Migration (IOM), nor by the European Union.

Publisher: International Organization for Migration
17 route des Morillons
P.O. Box 17
1211 Geneva 19
Switzerland
Tel.: +41 22 717 9111
Fax: +41 22 798 6150
Email: hq@iom.int
Website: www.iom.int

This publication was issued without formal editing by IOM.

Cover photo: On arrival in Juba, South Sudanese returnees are moved to a transit site established and managed by IOM and UNHCR, in coordination with South Sudan's Ministry of Humanitarian Affairs and Disaster Management and the Relief and Rehabilitation Commission. IOM provided shelter, water and sanitation, lighting and non-food emergency relief items in the transit site. © IOM 2020

Required citation: International Organization for Migration (IOM) (2024). *Impact Evaluations for Return and Reintegration Programmes*. IOM Regional Data Hub for the East and Horn of Africa.

ISBN 978-92-9268-791-5 (PDF)

* https://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode

PUB2023/100/R

IMPACT EVALUATIONS FOR RETURN AND REINTEGRATION PROGRAMMES

Andrew Pinney

Jane Poole

Emily Nevitt

Carlos Barahona

April 2023

Funded by
the European Union

IOM
UN MIGRATION

Stats4SD

FOREWORD

In view of gathering more and better evidence on the impact of return and reintegration assistance, it is essential to build and strengthen capacity to design and conduct impact evaluation studies. To this end, IOM offers this self-paced training on how to design and conduct impact evaluations in the context of return and reintegration programmes.

The training course will prepare you to support the planning and implementation of an Impact Evaluation study for a return and reintegration project or programme. Through seven self-paced modules, you will become familiar with the fundamental concepts and methods of impact evaluations and develop a "real-world" understanding of the challenges and opportunities of evaluating the impact of return and reintegration assistance with robust methodologies.

A note on this adapted format of the course

This publication is an adaptation of the e-course, *Impact Evaluations for Return and Reintegration Programmes*, produced for IOM's e-Campus online training platform. The e-course has been converted into document format for accessibility and for convenient referencing and citation of the course content. This document may be helpful to use alongside the e-course, as a helpful reference to look back at after finishing the e-course, or, if necessary, as a substitute in cases where it is not possible to access the e-course.

You can access the full e-course on IOM's e-Campus online training platform here:

🔗 Web page: *www.ecampus.iom.int/course/view.php?id=729*

The material was originally written to be an interactive multimedia course, and some compromises have had to be made in the process of adapting it into this format. Interactive lessons and activities are now static; animated presentations and video interviews have all been transcribed into lengthy chapters of text. The multimedia formats were an important consideration in the production of the original course, as the topic covered is quite information-dense. This text-based version of the course cannot and will not provide an optimal learning experience.

Although it is not recommended, for those readers that are using this document as a substitute for the e-course, we suggest taking the time to undertake the activities, knowledge checks and quizzes as directed in the text and making note of your responses before checking the correct answers. It would also be beneficial to work through the material a small amount at a time.

AUTHORS

The *Impact Evaluations for Return and Reintegration Programmes* course was produced for IOM by Statistics for Sustainable Development (Stats4SD). Andrew Pinney and Jane Poole provided their expert knowledge to the content development. The final course content was written and edited by Emily Nevitt with support from Carlos Barahona. The e-course multimedia and e-learning content was produced by Emily Nevitt with support from Emma Burrow. This text-based adaptation was designed and produced by Emily Nevitt with support from Emma Burrow.

ACKNOWLEDGEMENTS

Stats4SD

Project coordination and e-learning production: Emily Nevitt

Content development and impact evaluation experts: Andrew Pinney, Jane Poole and Carlos Barahona

Additional support: Dave Mills, Alex Riba, Grace Gilbert, Emma Burrow (Emma Burrow Design Limited)

IOM

Project steering and supervision: Davide Bruscoli, Joy Paone

Additional support: Elisabeth du Parc, Fatma Fetki, Lubna Jahaf

External review of course content: Dr Berkay Ozcan, London School of Economics

With contributions from:

Stefanie Barratt, Samuel Hall

Jasper Tjaden, PhD, University of Potsdam

Martin Schmitt, IOM Regional Office, Costa Rica

Professor Papa Sakho, Cheikh Anta Diop University

We wish to thank the European Union for funding the development of this course under the EU-IOM Joint Initiative for Migrant Protection and Reintegration.

CONTENTS

LIST OF FIGURES

IMPACT EVALUATION GLOSSARY

Attrition
Loss of respondents before all observations have been completed.

Baseline
Observations(s) taken before an intervention is implemented, with the aim of repeating the measurements at a later point and comparing the results.

Before/after
The principle of a before/after comparison is that you can measure an impact by collecting information on your chosen indicators before an intervention and then afterwards.

Causal attribution
In an impact evaluation, causal attribution is the act of establishing whether – and to what extent – changes that have been observed were caused by the treatment.

Comparison group
A group of individuals whose characteristics are similar to those of the intervention groups (or programme participants) but who do not receive the intervention.

Contamination
When factors other than the treatment have had an effect on the indicators being measured in an evaluation, making it difficult to understand the impact of the treatment.

Contribution
The evidence from an impact evaluation study shows that an intervention was related to an observed change, but there is insufficient evidence to understand if it was the sole cause.

Control group
A group of individuals excluded at random from receiving the intervention and for whom factors that may affect the measurement of the intervention's impact are the same for both the excluded and non-excluded individuals.

Counterfactual
The counterfactual is the hypothetical situation that would have occurred without the intervention. By comparing the counterfactual to the result with the intervention, we can isolate the effect of the intervention.

Difference in difference
The analytical approach of combining the before/after and with/without dimensions of comparison.

Endline
Measurement of the indicators of interest after the end of the treatment.

Group discussions
An interview-like discussion with a group, rather than one-on-one.

Impact
An impact is the effect caused by an action; this means the impacts are things that happen as the result of an intervention.

Indices
Multiple indicators are combined to create a single "index" or score, allowing for reporting and comparison. These composite indexes are more than a way of combining information – they define what reintegration means for the purpose of the study or studies that use it and the decisions that may be made as a result.

Instrumental variables

A variable that is strongly correlated with an individual's likelihood to receive the treatment, but has no correlation with the outcome of the treatment other than through whether they receive the treatment. The use of instrumental variables is a way to counteract selection bias when estimating the impact of an intervention.

Interviews

A conversation between an interviewer and a respondent, in which the interviewer seeks to obtain information by asking questions. This is a good way to gain an understanding of individuals' opinions, experiences and perspectives.

Logical framework

Outlines the linear hierarchy of inputs, outputs, outcomes and impacts, including specifying the indicators, sources of information and means of verification that can be used to measure progress towards the intended results. (Also called a Results Matrix.)

Longitudinal study

A study in which the same units are used in each round of data collection.

Matching

A strategy in which a comparison group is created by finding units that similar to each unit in the treatment group, according to the characteristics that are understood to be relevant to the effect of the intervention.

Midline

Measurement of the indicators of interest often taken after the treatment has commenced but before the end of the study.

Natural experiments

A scenario where a treatment and comparison group occur "naturally" – by chance and circumstance, rather than having been designed from the outset of the impact evaluation.

Normative thresholds

Some indices include thresholds. These are a bit like targets, or benchmarks, that allow measurements of reintegration to be compared against a theoretically constructed idea of "successful reintegration", or "good" or "poor".

Participant observation

Data are collected by watching people in a relevant setting – such as their regular day-to-day life or at a particular event – and making notes, taking photos, video or audio recordings.

Programme theory

The process of planning an intervention should include the development of a programme theory that maps out the changes, i.e. the outcomes and impacts, that the intervention was expected to produce and how these will come about, i.e. the activities and outputs.

Propensity score matching

A type of matching approach. Propensity score matching focuses on characteristics that make a unit more likely to receive the treatment and combines them to create a "propensity score" for each unit, that summarizes how likely they are to receive the treatment. Units with similar scores are matched, even if their specific characteristics are different.

Quasi-experiments

Methods that try to counter the risk of bias using various strategies to create a comparison group that allows for useful comparisons. Quasi-experimental designs resemble experiments in that they have many of the demands of a full experimental approach, such as management of contamination, but are different because they lack the random selection of the control group.

Randomized controlled trial

The most robust experimental approach, using fully randomized assignment of treatment.

Regression discontinuity design

Instead of random assignment, regression discontinuity designs use some system of ranking or eligibility based on a continuous variable, such as income, and assign a "cut-off" point, beyond which units are not eligible to receive the intervention. Units who are very close to this cut-off point on either side are considered similar enough to compare their outcomes and attribute the difference to the effect of the intervention.

Repeated cross-sectional study

A study in which a new sample of units is used for each round of data collection.

Selection bias

Caused when there are characteristics that affect a unit's probability of receiving the treatment and which also affect the outcome of the treatment.

Stepped-wedge design

A technique to create comparison groups. Rather than having a group that does not receive the treatment, the treatment is rolled out to groups of participants in a staggered manner over time.

Stratification

Stratification is when the units are divided up into groups – or "strata" – for the sake of sampling and/or reporting.

Theory of change

A theory of change (ToC) maps out how, why and for whom change is expected to occur in the intended context, describing the overall theory of how changes happen for an intervention to achieve its intended goal.

Treatment

When discussing evaluation studies, the generic term "treatment" is often used to refer to an activity, such as the intervention, programme or policy (or combination of these), that is being evaluated.

Treatment group

A group that has received the treatment.

Units of analysis

The unit at the level at which the intervention takes place and the analysis is conducted.

Units of observation

The unit or item you actually observe or measure.

Weights

Calculating a reintegration index involves combining information about a number of different factors. If certain factors, such as income, are deemed to be more important than others in terms of determining reintegration success, then the calculation that is done to produce a reintegration index should reflect that by "weighting" that factor so it has more impact over the final score than a less important factor.

With/without

Collect data on subjects who have received an intervention and on subjects who have not (referred to as a comparison group), then compare the results to understand how impacts might have been different without the intervention.

IOM staff conducting dispacement tracking activities in Al Fashir, the Sudan.
© IOM 2020/Muse MOHAMMED

MODULE 1: COURSE INTRODUCTION

MODULE 1: COURSE INTRODUCTION

INTRODUCTION

The aim of this module is to present the course curriculum and to enable trainees to determine whether, and in which ways, the course material is appropriate to their job role and learning needs.

OUTCOMES

At the end of this module, trainees will be able to:

- Identify whether the course is suitable for their job role and learning needs.

COURSE OVERVIEW

INTRODUCTION TO COURSE ON IMPACT EVALUATION FOR RETURN AND REINTEGRATION PROGRAMMES

Introduction by IOM Deputy Director General for Operations, Ugochi Daniels

Welcome to this course on how to design and conduct impact evaluations for return and reintegration programmes.

IOM has been working on return and reintegration for over 40 years and its expertise in this field is at the service of Member States and partners that work towards Objective 21 of the Global Compact for Safe, Orderly and Regular Migration. This objective is about facilitating safe and dignified return and readmission, as well as sustainable reintegration. One key aspect to any national or international agenda for Objective 21 is the ability to gather evidence on the impact of the return and reintegration initiatives that are being implemented.

However, reintegration is a complex process to define and is not easy to measure. A lot of effort has been put towards this in recent years by different organizations. Among these, IOM has invested significantly in the enhancement of monitoring and evaluation frameworks for return and reintegration initiatives, as well as in the development of specific surveys and tools, including its Reintegration Sustainability Index. All this work led to the so-called "IMPACT study": the first impact evaluation conducted on a large-scale reintegration programme implemented by IOM – the EU-IOM Joint Initiative for Migrant Protection and Reintegration.

Moving forward, we need to see more impact evaluations in the field of return and reintegration. This is fundamental to understand what works best and how to maximize benefits for returnees and their communities with the resources available. As we set ourselves to gather more and better evidence on the impact of reintegration assistance and tap into the analytical potential of ongoing initiatives, we must strengthen the capacity of reintegration practitioners to design and conduct impact evaluation studies.

It is against this backdrop this course has been developed, thanks to funding from the European Union. By taking it, you will become familiar with impact evaluation terminology, approaches, methodologies and ethics. Unlike the other courses on impact evaluation that are already available online, this course is entirely centred on the unique challenges of designing and conducting an impact evaluation for return and reintegration programmes. You will learn directly from the people that have been involved in impact evaluation. It is a course for practitioners, by practitioners.

While this course will help you prepare to support an impact evaluation for your return and reintegration initiative, either in a technical, managerial or administrative capacity, I hope that you will also be inspired by it: to introduce innovative features in your programme based on experimental evidence, or propose an entirely new impact evaluation study.

In conclusion, I want to thank again the European Union and the EU-IOM Joint Initiative programme for funding the development of this course. I wish all of you, a great experience and the opportunity to put in practice what you will learn.

Have fun!

WHY TAKE THIS COURSE?

Impact evaluations are becoming more common in humanitarian and development work in general, including in the return and reintegration context.

Organizations working in these areas, such as IOM, are interested in increasing the use of impact evaluations to assess the effects of their programme interventions, provide evidence of these effects and learn from the evaluation outputs.

Depending on your job role, it may therefore be useful or even essential to gain an understanding of the terminology, approaches and requirements involved in impact evaluations.

This course aims to familiarize you with impact evaluations and prepare you to support their planning and implementation for return and reintegration programmes. You will also benefit from expert perspectives on the unique challenges of conducting an impact evaluation in the context of return and reintegration programmes.

TARGET AUDIENCE

This course is aimed at reintegration practitioners and monitoring and evaluation (M&E) staff working in a return and reintegration context. This includes:

- Staff with monitoring and evaluation responsibilities;
- Users of monitoring and evaluation information products;
- Potential commissioners or users of impact evaluations;
- Potential managers of impact evaluations;
- Individuals who are already involved with the implementation of impact evaluations, who wish to understand the specifics of working within the return and reintegration context.

COURSE PREREQUISITE

This course specifically covers impact evaluations in the context of return and reintegration programmes. We assume that users already have an understanding of basic monitoring and evaluation concepts and of the fundamentals of return and reintegration assistance programmes.

Before beginning this course, we strongly recommend:

- Completing IOM's self-paced course on Monitoring and Evaluating Return and Reintegration Programmes;
- Reading/completing Module 5 of The IOM Reintegration Handbook and/or its associated course;

It would also be beneficial to familiarize yourself with the IOM Monitoring and Evaluation Guidelines document, which may be helpful as a reference throughout this course.

IOM self-paced course on Monitoring and Evaluating Return and Reintegration Programmes

Web page: *www.ecampus.iom.int/enrol/index.php?id=648*

Reintegration Handbook - Practical guidance on the design, implementation and monitoring of reintegration assistance

PDF: *https://publications.iom.int/system/files/pdf/iom_reintegration_handbook.pdf*

Reintegration Handbook Online Course

Web page: *www.ecampus.iom.int/enrol/index.php?id=84*

IOM Monitoring and Evaluation Guidelines

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf*

At various points, we may link to or mention content covered in the above materials, and we recommend you review any unfamiliar concepts you come across.

COURSE CONTENT

Outcomes

At the end of this course, trainees will be able to:

- Explain what an impact evaluation is, including how it differs from other types of evaluation;
- Explain the importance and benefits of conducting an impact evaluation for return and reintegration programmes;
- Identify when and why an impact evaluation should be conducted in the return and reintegration context;
- Define and use appropriate terminology in discussion and documentation concerning impact evaluations;
- Describe the methodology used in an impact evaluation and the implications this has for resource requirements, such as budget and technical expertise;
- Assess the characteristics and identify potential strengths and weaknesses of a proposed impact evaluation;
- Demonstrate an understanding of the criticisms of the "standard" before/after with/without impact evaluation design and the range of alternative designs/approaches;
- Participate constructively, and from an informed perspective, throughout the impact evaluation process, from setting out technical requirements to reading reports;
- Make informed decisions regarding the design and implementation of an impact evaluation;
- Continue learning autonomously about impact evaluation.

Modules

Module 1: Course introduction

Presents the course curriculum, provides information to enable trainees to decide if this course is appropriate to their job role and learning needs.

Module 2: Impact evaluation basics

Provides an overall understanding of what an impact evaluation is and why they are used and acquaints trainees with key concepts and terminology.

Module 3: Impact evaluation designs using quantitative methods

Outlines common impact evaluation methods, summarizing their main features.

Module 4: Impact evaluations in the context of reintegration programmes

Elaborates on the challenges, opportunities and possibilities particular to conducting impact evaluations in the context of return and reintegration programmes and presents expert interviews and real-life case studies.

Module 5: The role of qualitative methods in impact evaluations

Provides an overview of qualitative impact evaluation methods and explains how quantitative and qualitative methods can be combined to increase the robustness of the evaluation.

Module 6: Which impact evaluation design, when?

Presents decision trees to select appropriate design options for a quantitative impact evaluation, providing an overview or checklist.

Module 7: Expansions (from theory to quality implementation)
Introduces suggested "extension" topics relevant to the implementation of impact evaluation studies in a reintegration context, provides a basic overview and directs trainees to resources that will enable them to continue learning on these topics autonomously.

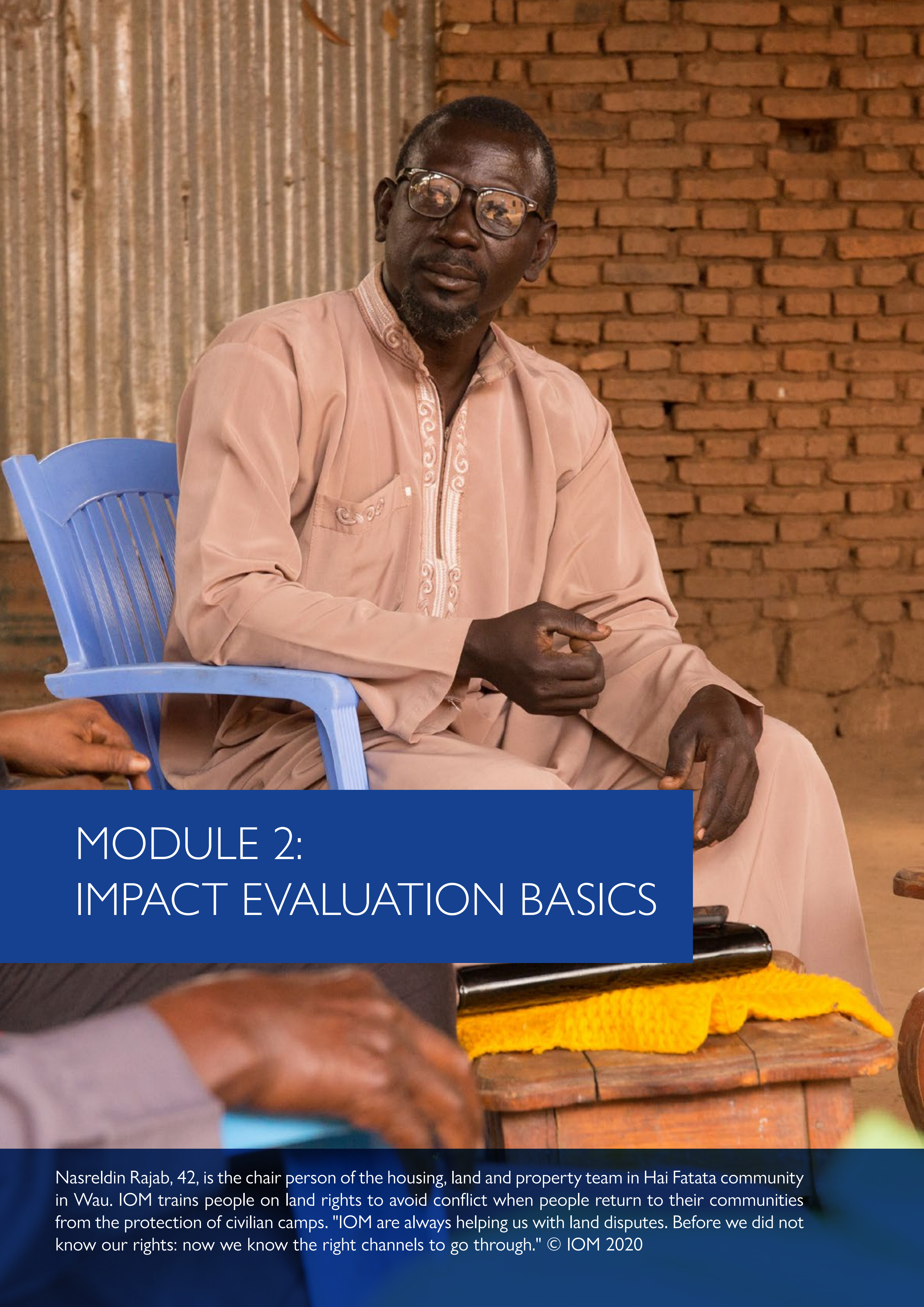
At the end of each module, there is a quiz to check your understanding of that module.

- To complete each module, you must achieve a passing score for the quiz.
- The score needed to pass will be given at the start of the quiz.
- If you do not pass, we recommend you review the module content before retaking the quiz.
- There is no limit to how many times you may retake the quiz.

CONCLUSION

You have completed this module.

You should now know what to expect in the remaining modules and understand how this course can benefit you. When you are ready, proceed to Module 2 to find out what impact evaluation is and how it works.



MODULE 2: IMPACT EVALUATION BASICS

Nasreldin Rajab, 42, is the chair person of the housing, land and property team in Hai Fatata community in Wau. IOM trains people on land rights to avoid conflict when people return to their communities from the protection of civilian camps. "IOM are always helping us with land disputes. Before we did not know our rights: now we know the right channels to go through." © IOM 2020

MODULE 2: IMPACT EVALUATION BASICS

INTRODUCTION

This module aims to provide an overall understanding of what an impact evaluation is, why it is a powerful component of a reintegration programme's monitoring and evaluation strategy and acquaint trainees with key concepts and terminology.

OUTCOMES

At the end of this module, trainees will be able to:

- Explain what an impact evaluation is, how it can contribute to programme design and decision-making and how it differs from other types of evaluation;
- Describe the options available to assess impact, attribution and contribution;
- Outline different stages and constraints in setting up an impact evaluation process;
- Define commonly used jargon and terminology relevant to impact evaluation;
- Give some examples of the complications and challenges specific to impact evaluations in complex development and humanitarian contexts;
- Apply ethical considerations when commissioning and conducting impact evaluations.

CASE STUDY: MEET THE EU-IOM JOINT INITIATIVE

Let's look at an example of a return and reintegration programme: the EU-IOM Joint Initiative for Migrant Protection and Reintegration.

A vocational training programme was launched at Nyala Technological College, South Darfur, by IOM through the EU-IOM Joint Initiative. © IOM 2021/Muse MOHAMMED

> "The EU-IOM Joint Initiative aims at enabling returnees to restart their lives in their countries of origin, grounded in IOM's integrated approach to reintegration. Reintegration assistance under the EU-IOM Joint Initiative supports migrants and their communities, has the potential to complement local development, and mitigates some of the drivers of irregular migration."
>
> "The reintegration support aims to address returnees' economic, social and psychosocial needs and to foster inclusion of communities of return in reintegration planning and support whenever possible. The EU-IOM Joint Initiative does not foresee standard reintegration packages. Instead, reintegration counsellors and returnees jointly define individual reintegration plans, which are tailored to the returnees' needs and vulnerabilities as well as their opportunities and motivations. The support may be provided to individuals, groups or communities."
>
> Web page: *www.migrationjointinitiative.org/reintegration*

This is a large, complex programme, with components in West and Central Africa, North Africa and the Horn of Africa. It involves many kinds of support interventions and requires a large amount of resources.

Find out more on the EU-IOM Joint Initiative web page

Web page: *www.migrationjointinitiative.org/about-eu-iom-joint-initiative*

Watch a short video about the EU-IOM Joint Initiative

Video: *www.youtube.com/watch?v=qVo1bVbPU2M*

Is it working?

So how can we know if a programme like the EU-IOM Joint Initiative is successful in achieving the intended results or impact?

What do we need to know?

We need to assess whether the intervention is generating:

- The desired impact
- For the right people
- Within the intended time frame
- Without any unintended consequences
- And as a result of the activities implemented.

How do we find out?

Assessing this requires:

- The collection of appropriate data
- At the right time
- From the right people, groups or institutions
- Performing relevant analysis to establish whether the intervention was the cause of the changes observed.

What can we do with this information?

The resulting information can inform decisions about:

- Targeting of project participants
- Adapting the current design and funding of ongoing programme interventions
- The design of future similar interventions and programmes.

The solution: impact evaluation

Impact evaluations are carried out to provide this information.

The IMPACT Study

The EU-IOM Joint Initiative programme includes an impact evaluation of its reintegration assistance work in the Horn of Africa, called "The Impact Evaluation of the EU-IOM Joint Initiative Programme for Migrant Protection and Reintegration (Horn of Africa)", otherwise known as IMPACT.

IMPACT will focus on Ethiopia, Somalia and the Sudan, where the number of programme beneficiaries is the highest.

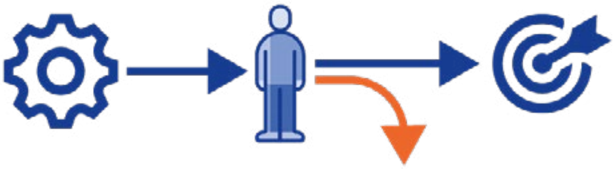
IOM staff verifies documents against the original flight manifests during a verification exercise of the population at the returnee transit site in Juba. © IOM 2012

Impact evaluations are carried out to measure the impacts of a wide range of interventions. This course focuses specifically on impact evaluations for return and reintegration programmes.

WHAT DO WE MEAN BY "IMPACT"?

Put simply, an impact is the effect caused by an action; this means the impacts are things that happen as the result of an intervention. These might be:

Intended or unintended



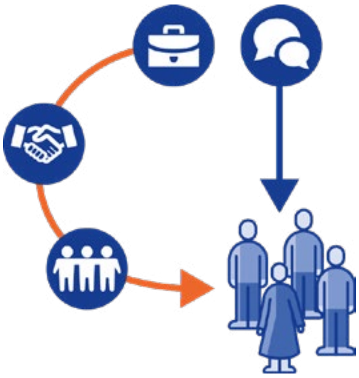
Intended
For example, increased income for returnee households provided with training support to gain employment.

Unintended
For example, returnees receiving training support make social connections and have improved happiness and well-being.

Direct or indirect

Direct
For example, a returnee has increased feelings of integration within the community of return.

Indirect
For example, gaining employment enabled a returnee to interact with local leaders who then elected the returnee to a local planning committee position.



Positive or negative

Positive
For example, a returnee has increased feelings of integration within the community of return.

Negative
For example, a returnee using increased income for non-productive activities, leading to conflict within the household.

When looking at the impacts that a programme activity has had, there will always be variation among programme participants. That is to say, some individuals or households will benefit significantly more or less than others from certain interventions. The methods used to assess the impact of an intervention try to ascertain the overall, or average, effect for the target population.

CAUSAL ATTRIBUTION

Impact evaluations aim to accomplish two key goals:

1. Measure the impact that occurs (including if there was no change, or a combination of positive and negative impact);
2. Establish whether – and to what extent – the intervention being evaluated was the cause of these observed changes. This is referred to as "causal attribution".

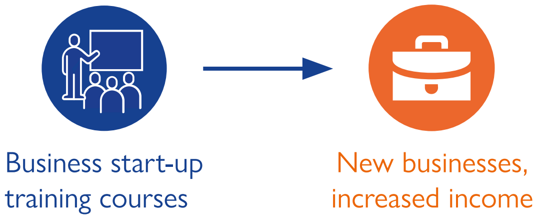
Causal attribution is what sets impact evaluation apart from other types of evaluation.

CAUSAL ATTRIBUTION CHALLENGES

Particularly in a return and reintegration context, there can be a range of factors that make it challenging to isolate the effects of interventions.

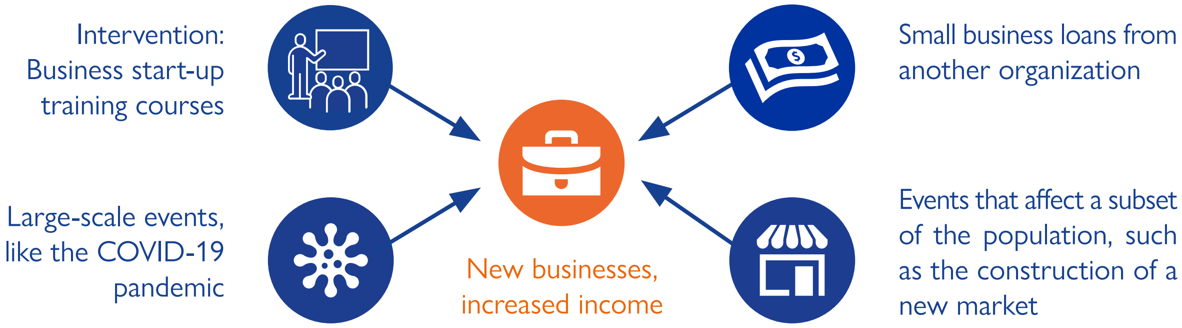
For example, if the programme provides business start-up training courses, they might consider this a successful intervention if new businesses are set up and increased income is observed (Figure 1).

FIGURE 1: THE INTERVENTION AND THE EFFECTS THAT ARE MEASURED TO UNDERSTAND THE IMPACT


Business start-up training courses
New businesses, increased income

However, if another organization working in the location is offering small business loans at the same time, to the same businesses, it can be hard to say which intervention caused the impact observed (Figure 2).

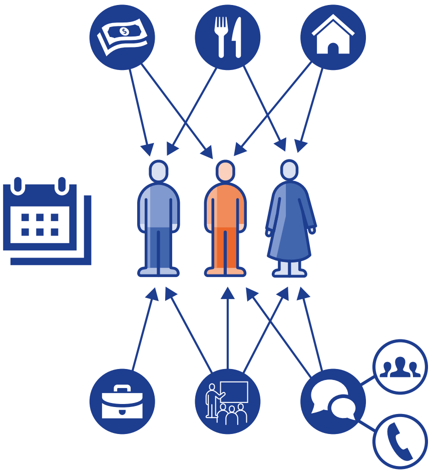
FIGURE 2: OTHER FACTORS CAN HAVE AN EFFECT ON THE IMPACT THAT IS MEASURED


Intervention: Business start-up training courses
Small business loans from another organization
Large-scale events, like the COVID-19 pandemic
New businesses, increased income
Events that affect a subset of the population, such as the construction of a new market

Further difficulty is added when other events affect results.

Failing to get an accurate impression of the impacts caused by the interventions could result in, for example, ineffective strategies continuing, or unknowingly repeating actions that produce negative impacts.

Attribution, contribution, or direct causality?



The term attribution rather than direct causality is often used when designing and analysing impact evaluations for programmes with complex outcomes, such as reintegration.

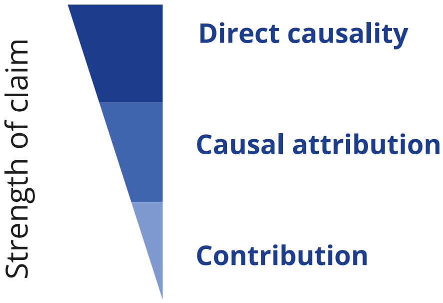
In this kind of programme:

- There are often multiple interventions;
- There may not be a predetermined plan of who receives which interventions when;
- The activities and services themselves may change in their design;
- Beneficiaries may receive different combinations of activities at different points in time.

This makes it very difficult to understand the impacts of individual interventions and activities.

An impact evaluation that examines such a complex situation is unlikely to provide strong evidence to say which impacts were produced by which activities. Instead, the focus would be on trying to evaluate the causal attribution of the programme as a whole, including all the activities.

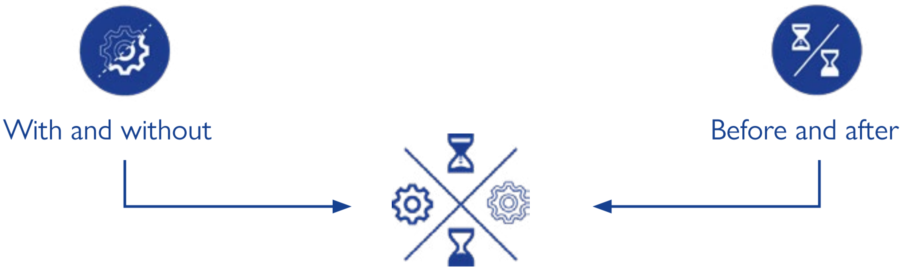
Conversely, in qualitative and mixed qualitative and quantitative impact evaluations, the assessment may be on the "contribution" of the intervention to the impacts observed, rather than "attribution". This alternative terminology acknowledges that there may have been other causes, and it isn't possible to say with certainty that the intervention was the sole reason for the impact observed.



HOW DOES IMPACT EVALUATION WORK?

So, how can an evaluation measure the impact that is attributable to the intervention? Impact evaluations typically accomplish this by making comparisons. In practice, there are two key dimensions which may be used for comparison, shown in Figure 3.

FIGURE 3: COMBINATION OF BEFORE/AFTER AND WITH/WITHOUT COMPARISONS



It is the combination of these two dimensions that allows causal attribution.

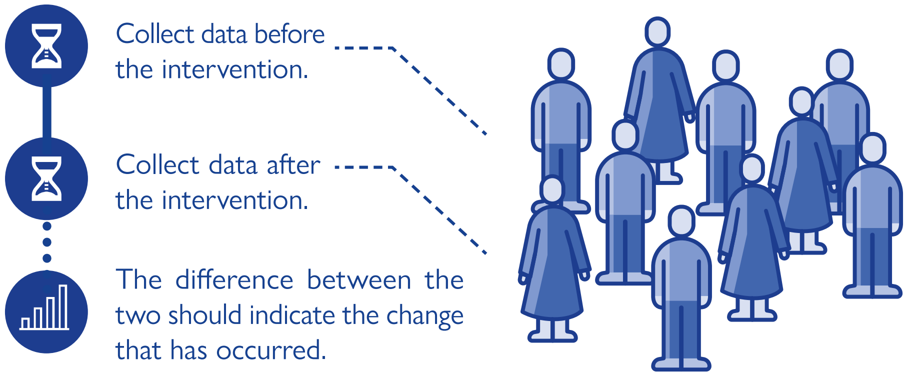
Let’s look at an example.

The assistance that may be offered by a reintegration programme includes psychosocial support (PSS) counselling. Say this counselling is offered to eligible returnees in a programme location, and they have the choice to enrol.

BEFORE AND AFTER

The principle of a before/after comparison is that you can measure an impact by collecting information on your chosen indicators before an intervention and then afterwards and then compare the two, as shown in Figure 4.

FIGURE 4: BEFORE/AFTER COMPARISON



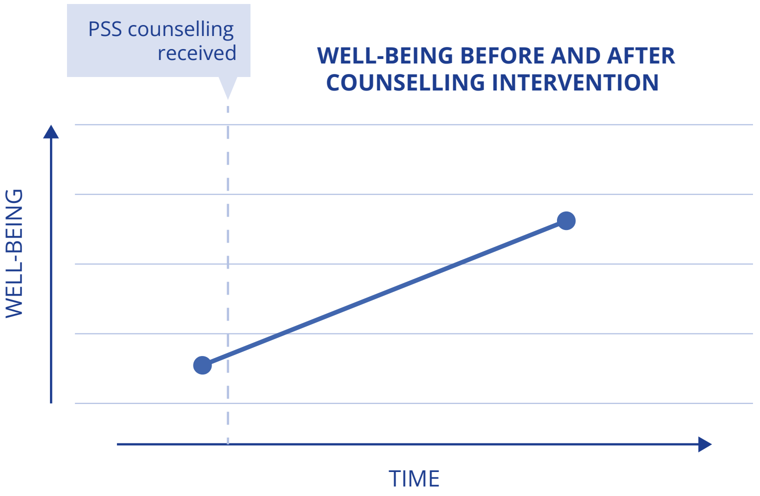
Collect data before the intervention.

Collect data after the intervention.

The difference between the two should indicate the change that has occurred.

Let’s say we take measurements before and after the counselling is provided, and our results are as shown in Figure 5.

FIGURE 5: EXAMPLE OF A BEFORE/AFTER COMPARISON



Question

What does this result tell you? Select one answer.

- □ PSS counselling improved well-being.
- □ There has been an improvement, but we don’t know why.

See the next page for the answer.

Answer

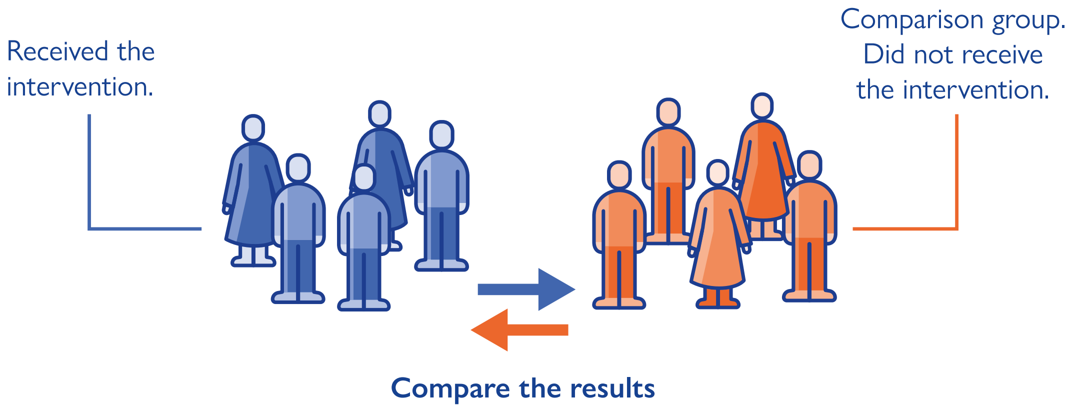
- ☐ PSS counselling improved well-being.
- ☑ **There has been an improvement, but we don't know why.**

The fact that well-being improved after PSS counselling is a good sign, but it's possible there was another cause, so we need more information before we claim the counselling caused the improvement.

WITH AND WITHOUT

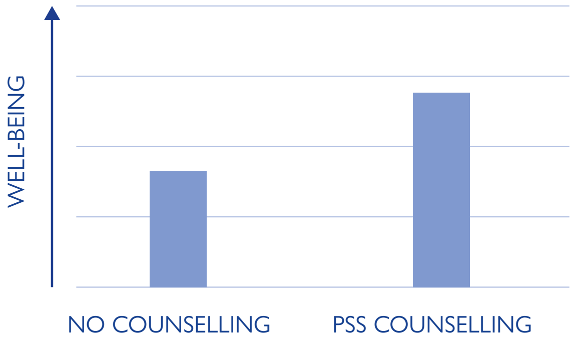
The primary method of trying to establish whether an intervention was the cause of an observed impact is creating comparison groups to make with/without comparisons (Figure 6). We can collect data on subjects who have received an intervention and on subjects who have not (referred to as a comparison group), then compare the results to understand how impacts might have been different without the intervention.

FIGURE 6: WITH/WITHOUT COMPARISON


Received the intervention.
Comparison group. Did not receive the intervention.
Compare the results

This time, we look at those who received counselling compared to people in the same location who didn't enrol in or attend the counselling. Let's say we take measurements after the intervention for the groups that did and did not receive counselling, and the results are as shown in Figure 7.

FIGURE 7: EXAMPLE RESULTS OF A WITH/WITHOUT COMPARISON



Question

What does this result tell you? Chose one answer.

- ☐ PSS counselling improved well-being.
- ☐ Well-being improved for those who received PSS counselling.
- ☐ After the intervention, those who received counselling had better well-being than those who didn't.

See the next page for the answer.

Answer

☐ PSS counselling improved well-being.

☐ Well-being improved for those who received PSS counselling.

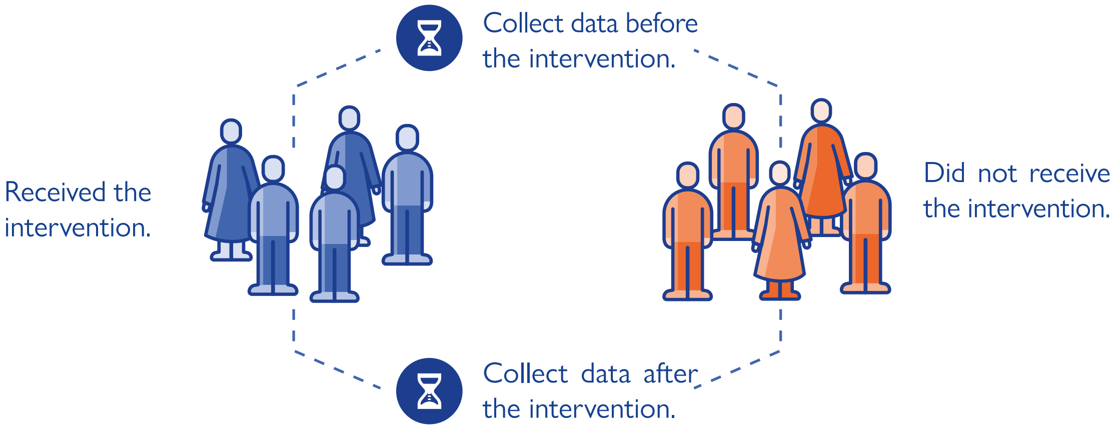
☑ **After the intervention, those who received counselling had better well-being than those who didn't.**

We know that those who received counselling had better well-being afterwards than those who didn't, but we don't know how it has changed over time.

COMBINED BEFORE/AFTER, WITH/WITHOUT COMPARISONS

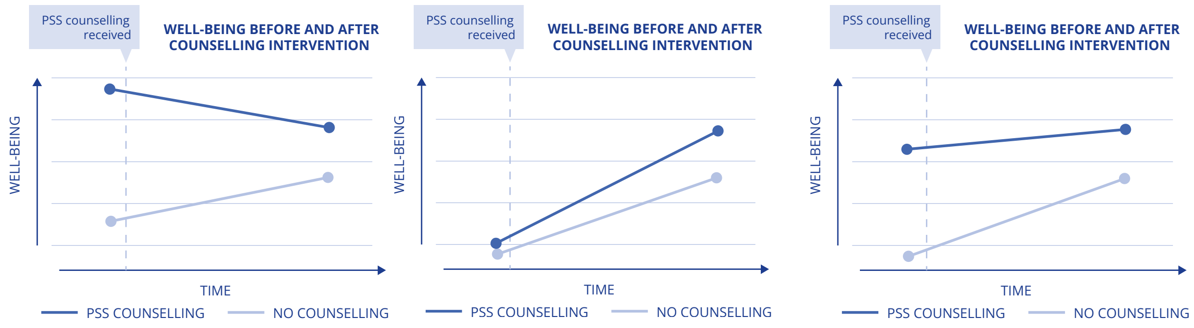
To gain a better understanding of the situation, we need to combine the two dimensions, as shown in Figure 8.

FIGURE 8: COMBINED BEFORE/AFTER AND WITH/WITHOUT COMPARISONS


Collect data before the intervention.
Received the intervention.
Did not receive the intervention.
Collect data after the intervention.

Looking at only Figure 7, any of the outcomes in Figure 9 could be true.

FIGURE 9: EXAMPLES OF POSSIBLE OUTCOMES BASED ON ONLY THE RESULTS FROM A WITH/WITHOUT COMPARISON

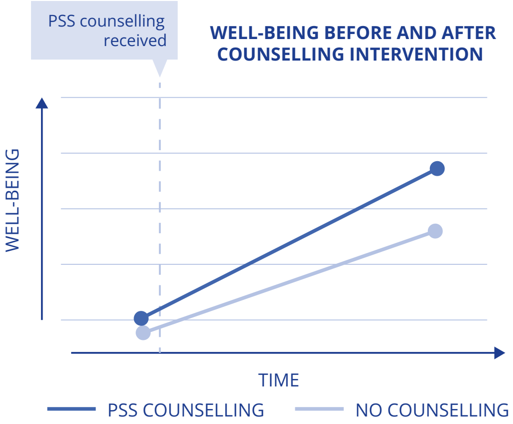


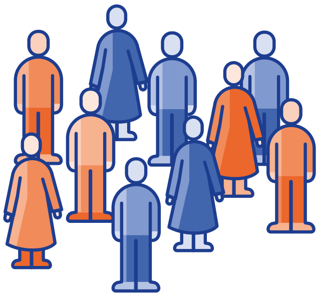
CREATING A COMPARISON GROUP

Let's say we take measurements before and after for both groups and the graph in Figure 10 shows our result.

While it looks like the counselling has had a positive impact on well-being, the situation in reintegration contexts tends to be complicated. Before we can make conclusions about the impact, we need to look again at the comparison group we used.

FIGURE 10: COMBINED BEFORE/AFTER AND WITH/WITHOUT RESULTS FOR EXAMPLE SCENARIO





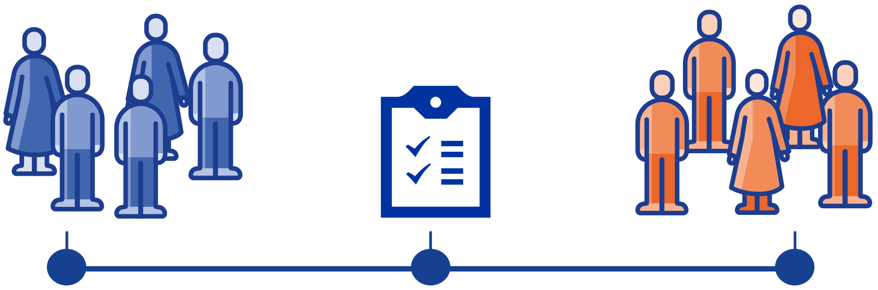
Some factors may have affected the results:

- Returnees in wealthier areas would have been more able to access counselling and may also have generally had better well-being over time;
- Older, less educated, or more traumatized returnees could have been less able or willing to enrol in counselling;
- Returnees with these characteristics would be more likely to have poorer well-being;
- Returnees who received counselling could have also received other support, which could have been the reason for the improvement.

Simply comparing those who attended with those who did not allows these other factors to confuse the results. To be effective in establishing causal attribution, comparison groups need to be as similar as possible.

To create a more helpful comparison group, one option is to deliver the counselling service in cohorts, using a method sometimes called a stepped-wedge approach (Figure 11).

FIGURE 11: STEPPED-WEDGE APPROACH FOR CREATING A COMPARISON GROUP



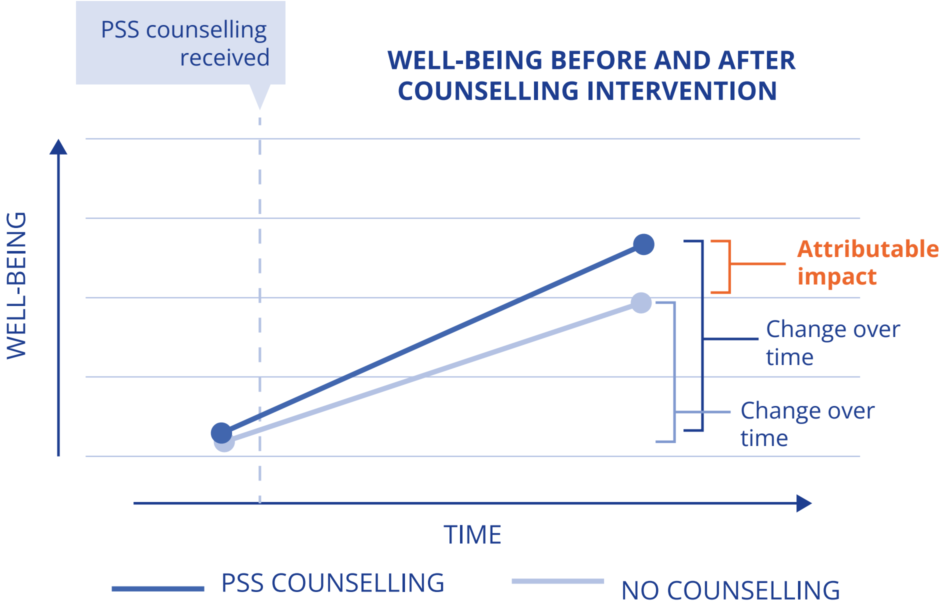
Take everyone who enrols and randomly assign half to receive the counselling.

Take measurements from both groups before and after the first group receives the psychosocial counselling.

The second group will receive the counselling after the measurements have been taken.

Figure 12 below displays the results that are observed.

FIGURE 12: DIFFERENCE IN DIFFERENCE RESULTS FOR EXAMPLE SCENARIO



We compare the change in well-being over time for both groups. The difference in the size of change over time between the two groups is understood to be the attributable impact of the psychosocial support counselling. This technique of combining the before/after and with/without dimensions is called difference in difference.

THE TWO DIMENSIONS OF COMPARISON

Before and after

The "before" measurement is often referred to as the "baseline" and the "after" as an "endline". Some studies also use "midline" or "periodic" observations to capture more points of information between a baseline and an endline.

Challenges

- Having a baseline measurement requires planning of this evaluation strategy very early in the process.
- When do you collect data for the "after" point?
  - Certain impacts may be time sensitive; for example, after food vouchers have been provided, it may take some time before the nutritional benefits are observable in beneficiaries.
  - Depending on what you are measuring, it may be necessary to take multiple "after" observations to look at the sustainability of a change.

With and without

Counterfactual

The principle behind comparing a group who received an intervention with a group who didn't is that we are trying to estimate the counterfactual. The counterfactual is the hypothetical situation that would have occurred without the intervention. By comparing the counterfactual to the result with the intervention, we can isolate the effect of the intervention.

However, it is of course impossible to go back in time and see what would have happened in the alternate reality where there was no intervention! Different strategies are therefore used to try to get as good an estimate as we can.

Comparison groups

Comparison groups are the primary way that studies try to estimate a counterfactual. As seen previously, comparison groups are a group that does not receive the intervention.

Studies should aim to create a comparison group that is as similar as possible to the group receiving the treatment. This should help avoid other factors affecting our results.

Control groups

Control groups are the most rigorous type of comparison group. When creating a control group, "units" (such as individuals, households, etc.) are selected at random to receive the intervention or not. As much as possible, the study aims to control other factors that may affect the measured impacts by ensuring these factors are as similar as possible in both the intervention and control groups.

Challenges

- In development contexts, it can be challenging to identify a comparison or control group. Selection of subjects (or their locations) to receive an intervention(s) will often follow explicit criteria (e.g. vulnerability, household composition – presence of children, people with disabilities). These may be difficult to identify or apply in the same way to a comparison or control group.
- It may be problematic to use a control group for logistical and/or ethical reasons (more discussion on this on page 87).

The challenge of how to make comparisons is an important aspect of the design of an impact evaluation for reintegration programmes; it will be covered in more detail in Module 3 and Module 4.

HOW DOES IMPACT EVALUATION COMPARE TO OTHER TYPES OF EVALUATION?

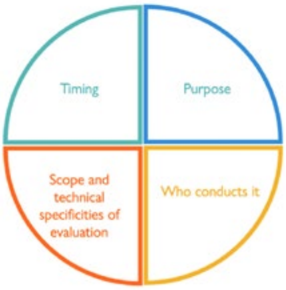
There is a range of evaluation approaches that can provide information about the results of an intervention. What sets impact evaluation apart from other kinds of evaluation?

In its guidance, IOM sets out four elements by which types of evaluation can be characterized (Figure 13). Evaluations can be a combination of these categories.

You may wish to refer to the *IOM Monitoring and Evaluation Guidelines* if these evaluation types are unfamiliar.

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=225*

FIGURE 13: TYPES OF EVALUATION



Impact evaluation is distinct from other types of evaluation in the following ways:

Establishing attribution

Impact evaluations aim to determine both the effects (i.e. the changes and size of these) and whether these can be attributed to the intervention.

Methodology

Impact evaluations require specific methodologies, such as the use of comparison groups and baseline-endline data collection, to show causal attribution. They are quite demanding in terms of technical requirements, budget and data-collection needs.

Because of this, impact evaluations should ideally be planned and integrated into programme implementation.

Rigour

Impact evaluation is also the most rigorous type of evaluation in terms of being able to conclude with confidence whether an intervention is working and how well it is working.

Further reading

Types of evaluation

United States Centers for Disease Control and Prevention.

PDF: https://vetoviolence.cdc.gov/apps/evaluaction/assets/pdf/Types-of-Evaluation.pdf

INTRAC Guide to types of evaluation

INTRAC is a not-for-profit organization that builds the skills and knowledge of civil society organizations to be more effective in addressing poverty and inequality.

PDF: *www.intrac.org/wpcms/wp-content/uploads/2017/01/Types-of-Evaluation.pdf*

Guidance on choosing appropriate designs and methods for impact evaluation

From the Australian Department of Industry, Innovation and Science 2015.

PDF: *www.industry.gov.au/sites/default/files/May 2018/document/pdf/choosing_appropriate_designs_and_methods_for_impact_evaluation_2015.pdf?acsf_files_redirect*

IOM M&E Guidelines

See pages 211–220: Types of evaluation.

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=225*

KNOWLEDGE CHECK

Question 1

Which of the following statements about impact evaluation are true? Select all the answers that apply.

- □ Impact evaluations are useful for informing the planning of future interventions and programming.
- □ Impact evaluations are a cheap way of judging the effectiveness of programme implementation.
- □ Impact evaluations aim to establish the cause of impacts.
- □ Impact evaluations cannot provide information on unexpected outcomes of an implementation.

Question 2

Pick the best definition for the term "impact" in the context of evaluation. Select one answer.

- □ Comparisons that are made to establish causal attribution.
- □ The effects, both intended and unintended, that occur as a result of an intervention.
- □ Actions, such as interventions, that generate effects. These effects can be positive or negative.
- □ A measure of the effectiveness of programme implementation.

Question 3

Match the definitions to the terms.

Terms	Definitions
Contribution	Observations(s) taken before an intervention, to be compared with future observations.
Control group	When we know an intervention was related to a change, but not that it was the sole cause.
Causal attribution	Measure the impact that occurs and establish to what extent the intervention was the cause.
Comparison group	A group that is similar to the intervention group, which does not receive the intervention.
Baseline	A group that is randomly selected to not receive the intervention.

Question 4

Assuming they are possible to successfully implement, which of these strategies would be likely to produce the strongest evidence for causal attribution? Select one answer.

- □ Collecting data after the intervention for a group that received the intervention and a group that did not.
- □ Collecting data at baseline and endline for a group that received the intervention and a control group who did not.
- □ Collecting data at baseline, midline and endline for the group that received the intervention and a non-randomly selected comparison group.

KNOWLEDGE CHECK ANSWERS

Question 1 answer

Which of the following statements about impact evaluation are true? Select all the answers that apply.

☑ **Impact evaluations are useful for informing the planning of future interventions and programming.**
Impact evaluations can provide important information for decision-making and planning of future programme activities.

□ **Impact evaluations are a cheap way of judging the effectiveness of programme implementation.**
Impact evaluations are not cheap and can have a variety of aims beyond judging "effectiveness".

☑ **Impact evaluations aim to establish the cause of impacts.**
An important aspect of impact evaluations is their aim to establish causal attribution.

□ **Impact evaluations cannot provide information on unexpected outcomes of an implementation.**
Impact evaluations aim to understand all impacts resulting from an implementation, both intended and unintended.

Question 2 answer

Pick the best definition for the term "impact" in the context of evaluation. Select one answer.

□ **Comparisons that are made to establish causal attribution.**
Comparisons are a component of the methods used in impact evaluation.

☑ **The effects, both intended and unintended, that occur as a result of an intervention**
This is the correct answer.

□ **Actions, such as interventions, that generate effects. These effects can be positive or negative.**
An impact is the effect, positive or negative, intended or not, that is caused by an action.

□ **A measure of the effectiveness of programme implementation.**
Impacts might be measured as part of understanding the effectiveness of a programme, but they are not the same thing.

Question 3 answer

Match the definitions to the terms.

Definitions		Terms
Contribution	→	When we know an intervention was related to a change, but not that it was the sole cause.
Control group	→	A group that is randomly selected to not receive the intervention.
Causal attribution	→	Measure the impact that occurs and establish to what extent the intervention was the cause.
Comparison group	→	A group that is similar to the intervention group, which does not receive the intervention.
Baseline	→	Observations(s) taken before an intervention, to be compared with future observations.

The following modules make frequent use of these terms. If you are unsure about their definitions, we advise reviewing the course content before continuing.

Question 4 answer

Assuming they are possible to successfully implement, which of these strategies would be likely to produce the strongest evidence for causal attribution? Select one answer.

- □ Collecting data after the intervention for a group that received the intervention and a group that did not.
- ☑ Collecting data at baseline and endline for a group that received the intervention and a control group who did not.
- □ Collecting data at baseline, midline and endline for the group that received the intervention and a non-randomly selected comparison group.

An option that also includes both a with/without and a time-based comparison is ideal.

A control group, which is randomly selected, will produce better evidence than a non-randomly selected comparison group, even with fewer rounds of observations.

THE CASE FOR IMPACT EVALUATION

WHY CONDUCT AN IMPACT EVALUATION?

Impact evaluations are complex and can be expensive, challenging and technically demanding to implement. So why would we want to conduct one? IMPACT's stated goals demonstrate their motivations:

> "The Impact Evaluation of the EU-IOM Joint Initiative Programme for Migrant Protection and Reintegration (Horn of Africa), hereby IMPACT, aims to provide a robust assessment of the impact of IOM's reintegration assistance, providing an accountability mechanism to beneficiaries of the programme, the donor and wider sector, and an evidence base to inform future reintegration programming, while maximizing cost-effectiveness."
>
> *Methodological Report IMPACT*
>
> 🔗 Web page: *https://returnandreintegration.iom.int/en/resources/report/methodological-report-impact-impact-evaluation-eu-iom-joint-initiative-migrant*

Humanitarian and development interventions are not only costly in terms of resources and time, but the well-being of large numbers of people can be hugely affected by their success or otherwise. Therefore, we want to know that the opportunities and resources that are available are being used in a way that will achieve desired outcomes and additionally are not producing unexpected negative impacts.

Impact evaluations can help to...

Provide effective assistance

It is important to evaluate whether the interventions in a programme are having positive impacts, especially as resources and time are often limited. This can support learning and improvement of programme activities.

Inform decision-making

The information provided by an impact evaluation helps with planning future similar interventions or implementing an intervention in more locations. Also, depending on the timing, it may inform decisions about continuing or changing an ongoing intervention.

Justify costs

Having evidence that particular interventions produce the desired effects can justify spending or resource requirements.

Impact evaluations are important because...

There is increased use and focus by international organizations

Another argument for conducting impact evaluations is that there is an increased demand for them. For example, the United Kingdom's Foreign, Commonwealth & Development Office (FCDO), previously Department for International Development, have long invested in sponsoring and recruiting a cadre of statisticians. Along with the World Bank, FCDO funds quantitative impact assessments, where appropriate.

(Note that this shift is not universal among such organizations. There are differing viewpoints on the effectiveness and suitability of impact evaluations, which will be mentioned in the next section.)

This increased demand for impact evaluations:

1. Can be read as an endorsement of the value and importance of impact evaluations in providing accountability for expenditure and providing lessons learned for how programmes can maximize positive outcomes.
2. Means those implementing return and reintegration programmes may at some stage be obliged by donor requirements to carry out an impact evaluation as part of a programme.

They were part of the work that won the Nobel Memorial Prize in Economic Sciences 2019

In 2019, the Nobel Memorial Prize in Economic sciences was awarded to three winners (Abhijit Banerjee, Esther Duflo and Michael Kremer) for their work in applying an experimental approach to assessing development projects.

The award can be seen as an endorsement of the use of impact evaluation methods in humanitarian and development contexts. In addition, the winners' findings suggested that without such evaluations, it is possible to unknowingly continue interventions that are not having the expected effect.

The move towards more prevalent use of impact evaluation has led to more informed decision-making in humanitarian interventions worldwide.

The group used randomized control trials (a method used in impact evaluation; see page 86), already a common method in clinical trials, to evaluate the effects of interventions.

This work has had a large effect in the field, where previously interventions were based on logically reasonable, but largely untested, assumptions about what would benefit vulnerable and disadvantaged communities. Interventions were being rolled out without a full understanding of whether they had been the cause of improvements observed, or whether they were necessarily an effective use of limited resources.

Find out more:

Nobel Memorial Prize Lecture: Abhijit Banerjee

Video: *www.youtube.com/watch?v=XvyMO7CmFlk*

Nobel Memorial Prize Lecture: Esther Duflo

Video: *www.youtube.com/watch?v=KFRnY-5K5OU*

Nobel Memorial Prize Lecture: Michael Kramer

Video: *www.youtube.com/watch?v=hOTHeNZU_JQ*

INTERVIEW TRANSCRIPT: WHY CONDUCT AN IMPACT EVALUATION?

An interview with Jasper Tjaden, Professor for Applied Social Research and Public Policy at the Economic and Social Science Department of the University of Potsdam, and formerly of IOM and the World Bank, talking about why impact evaluations are important for return and reintegration programmes.

WHY CONDUCT AN IMPACT EVALUATION?

Interview with Professor Jasper Tjaden, Professor for Applied Social Research and Public Policy, University of Potsdam

My name is Jasper Tjeden. I am professor for Applied Social Research Policy at the University of Potsdam in Germany. I conduct research on various topics related to migration and migrant integration questions. Before joining the university, I worked for the International Organization for Migration – IOM – for about four years, for their Global Migration Data Centre in Berlin. There, I led a programme around conducting impact evaluations in different areas and different countries.

Before joining IOM, I had the pleasure to work with the World Bank for a year; I was a research assistant, research analyst, in their development impact unit that is primarily concerned with analysing the impact of various development programmes around the world.

There are various reasons why one should conduct an impact evaluation. One, obviously, is accountability. The funding organization provides funding to implement the programme so whoever is implementing it needs to be accountable to ensure that they're actually achieving the kind of impact they're going for. But more than accountability, it's also about learning. So, the implementing organization should conduct, or should consider conducting, an effective elevation, because it's a great way for everyone involved to learn whether how they're doing things is effective and how they could do better in the future. So, this is more about continuous learning – learning how to perform that in the future, how to improve the project, regardless of what the donor thinks or where the money is coming from, but rather how to improve operational efficacy.

Without these more robust ways to measure the impacts, you never really know whether you're achieving what you're planning to achieve. And, without it, we don't know whether the whole approach that projects are taking is worth it, and is the right one and whether we should adjust the programme to go elsewhere. A lot of projects and programmes are decided on intuition or some basic background and look at the literature maybe, or the experience of certain people involved, but rarely do we have actual hard evidence on the effects of programmes. But the effects of programmes, and that hard evidence can really move funding organizations, also, in the direction where they're willing to scale up funding for a certain programme because there are some hard facts on the table about the effectiveness of those programmes.

The area of return and reintegration poses some specific challenges to impact evaluation, but also it is an area where there's a huge demand for these types of impact evaluations. Why? Mainly because it's a rather – not a new field, but a field where there has been a lot of recent activity in funding and investment. A lot of money has gone to return and reintegration programmes, a lot of different strategies have been applied, and we simply don't have a very good understanding of what works, and it is in those situations – where there's a lot of activity but very little evidence – that impact evaluations have their greatest benefit. So, I think especially in the area of return and reintegration, there's great potential for impact evaluations.

Learning from impact evaluations

A lot of United Nations organizations have, still, this implementing mentality of saying we get money, we implement it, and then we show the donor that we did an amazing job. This is not the mentality of impact evaluation. Impact evaluation, you want to learn from mistakes; you want to improve in the future. Certain donors like that attitude, and they fund you even if it's possible that you fail, because you at least have a sincere interest to be better next time. However, there's still a lot of implementing agencies that are not in that mindset. They're really afraid of disappointing donors and then receiving less funding in the future.

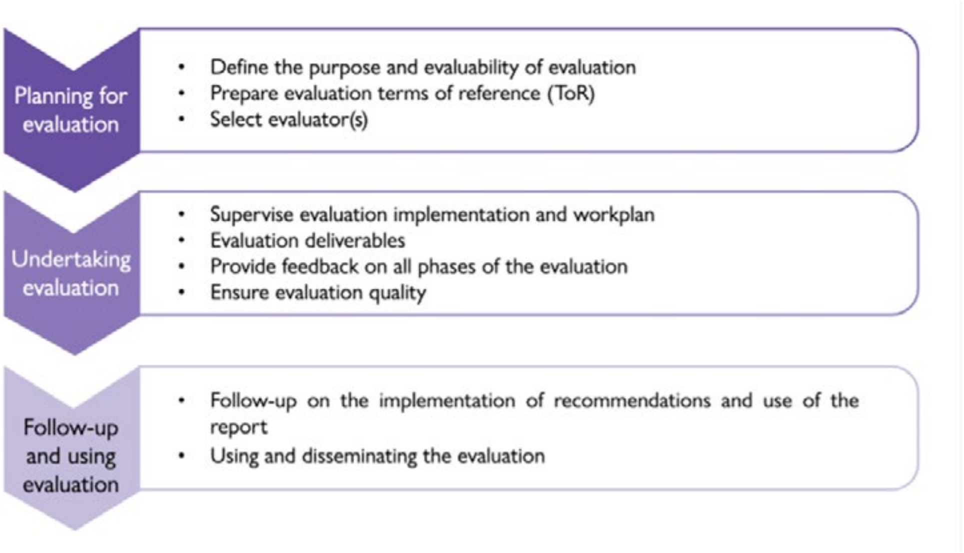
And I think it's important for impact evaluations to work, it's important to move away from that and approach this mentality of continuous learning and improvement. Then everyone will benefit, because in the end it's taxpayer money. And, you know, if you can find and study that certain things work when others don't, you make it public, others can learn from that and maybe avoid your mistakes. Or they can capitalize on the great projects that you have done. I'm aware it's a competitive environment between implementers, but in the end this is a public investment, a public good, that you know that everyone is contributing to and that should be shared.

PLANNING AN IMPACT EVALUATION

STAGES OF AN IMPACT EVALUATION

Before considering setting up an impact evaluation, it would be helpful to be familiar with the more general process for an evaluation.

In the following sections, information specific to impact evaluations is given. However, if the stages here are unfamiliar, it would be sensible to refer to the *IOM M&E Guidelines* to gain more general information about the process of conducting an evaluation.



Planning for evaluation

- Define the purpose and evaluability of evaluation
- Prepare evaluation terms of reference (ToR)
- Select evaluator(s)

Undertaking evaluation

- Supervise evaluation implementation and workplan
- Evaluation deliverables
- Provide feedback on all phases of the evaluation
- Ensure evaluation quality

Follow-up and using evaluation

- Follow-up on the implementation of recommendations and use of the report
- Using and disseminating the evaluation

Stages of Evaluation – IOM M&E Guidelines p.207

Due to the scope and technical requirements of an impact evaluation, the design of the evaluation should ideally be considered before the design of an intervention is finalized.

This is because the impact evaluation design may influence the type and timing of baseline information collection, the implementation schedule of the intervention and the identification or construction of a comparison or control group.

IS AN IMPACT EVALUATION APPROPRIATE?

Impact evaluations are not trivial to carry out; substantial resources are required to conduct them effectively. Therefore, it is vital to have a clear understanding of the purpose of an impact evaluation and how the findings will be used before deciding to implement one. You should conduct an evaluability assessment, ideally during the programme design stage.

Web page: *www.betterevaluation.org/methods-approaches/themes/evaluability-assessment*

Utilization-Focused Evaluation

Michael Quinn Patton developed the approach of Utilization-Focused Evaluation, which is based on the assertion that an evaluation should be judged by its usefulness, and thus the planning and implementation of an evaluation should aim to maximize the potential and likelihood of its findings being used.

Find out more on Utilization-Focused Evaluation:

Web page: *www.betterevaluation.org/methods-approaches/approaches/utilisation-focused-evaluation*

THINGS TO CONSIDER BEFORE DECIDING TO CARRY OUT AN IMPACT EVALUATION

These are five key aspects to consider before deciding to carry out an impact evaluation as part of a programme or intervention.

WHO IS GOING TO USE THE FINDINGS OF AN IMPACT EVALUATION?

Ideally, these intended users would be engaged in the planning process to help ensure that the findings meet their information needs.

IS IT THE RIGHT TIME TO CARRY OUT AN IMPACT EVALUATION?

There is little point in conducting one if the results will arrive too late to be useful in decision-making.

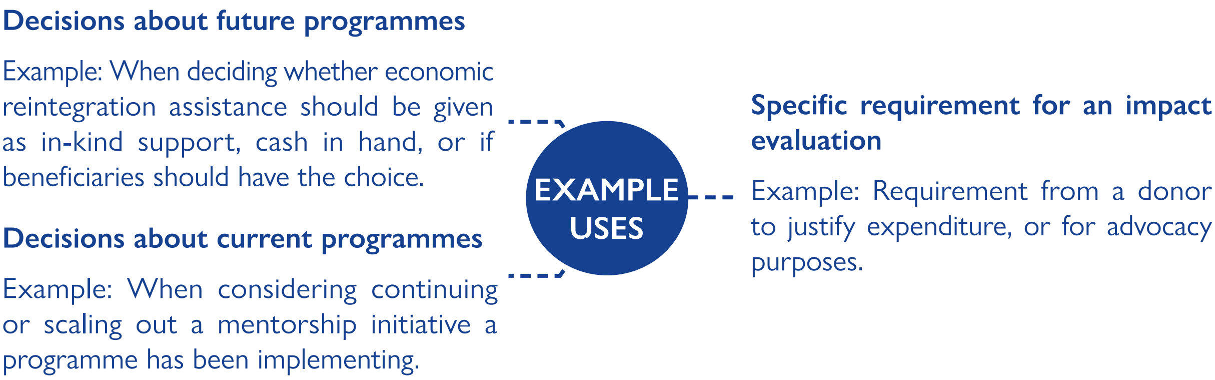
However, consider that the conclusions may be relevant outside of the programme being evaluated, such as informing decisions concerning future programmes. Therefore, even if the programme being evaluated has ended, this does not necessarily mean it is too late to learn from it.

Timing is also crucial for the measurement of some indicator**s**, as impacts may take some time to emerge, or cease to be observable after a certain period.

IS THERE A SPECIFIC PURPOSE FOR THE INFORMATION YOU PLAN TO OBTAIN?

There must be a clear understanding of the uses for an impact evaluation, in order to:

- Justify carrying one out in the first place;
- Inform decisions when planning an evaluation to ensure the information produced is suitable for its intended use.



Decisions about future programmes

Example: When deciding whether economic reintegration assistance should be given as in-kind support, cash in hand, or if beneficiaries should have the choice.

Decisions about current programmes

Example: When considering continuing or scaling out a mentorship initiative a programme has been implementing.

EXAMPLE USES

Specific requirement for an impact evaluation

Example: Requirement from a donor to justify expenditure, or for advocacy purposes.

ARE SUFFICIENT RESOURCES AVAILABLE?

This should be established before committing to implementing an impact evaluation. There is little to be gained from carrying out an impact evaluation poorly due to lack of time, budget or other resources.

IS THERE AN INTENTION TO USE AND APPLY THE FINDINGS?

Is there a genuine appetite and commitment at the relevant levels to use and apply findings from an evaluation in future programmes and interventions?

- Is the strategic focus currently aimed at understanding the impact of interventions?
- Is there an intention to carry out further programmes and interventions to which the findings could be applied?

An impact evaluation is only relevant if the aims and intentions of relevant organizations, partners and stakeholders align with the intended use case.

HOW MUCH DOES AN IMPACT EVALUATION COST?

Impact evaluations are more demanding than other kinds of evaluation, and the budget needs to reflect that.

It is not sensible to suggest a single figure for the cost of an impact evaluation. Examples estimated by just a small number of practitioners working in reintegration or similar contexts range from USD 60,000 to multimillion-dollar evaluations. Impact evaluations funded by 3ie cost an average of about USD 334,000.

There are numerous aspects of an impact evaluation that will affect the cost:

- Sample size
- Data-collection method
- Travel
- Security
- Languages
- Location(s)
- Scale
- Pretesting
- Staff training
- Equipment
- Stratification (read about this on page 66)
- Level of precision of estimated effects required
- And many more.

Certain strategies can reduce costs. For example, conducting phone surveys saves a significant amount of money on travel and related costs. Compromises can be made to the data collected or the information the evaluation can produce. However, cutting costs too far can lead to poor practices and low-quality outputs.

Striking the right balance between quality of information outputs and affordability is a significant challenge in planning an impact evaluation.

BEFORE YOU START: PROGRAMME THEORY AND EVALUATION QUESTIONS

Before commencing an impact evaluation, the following should be established:

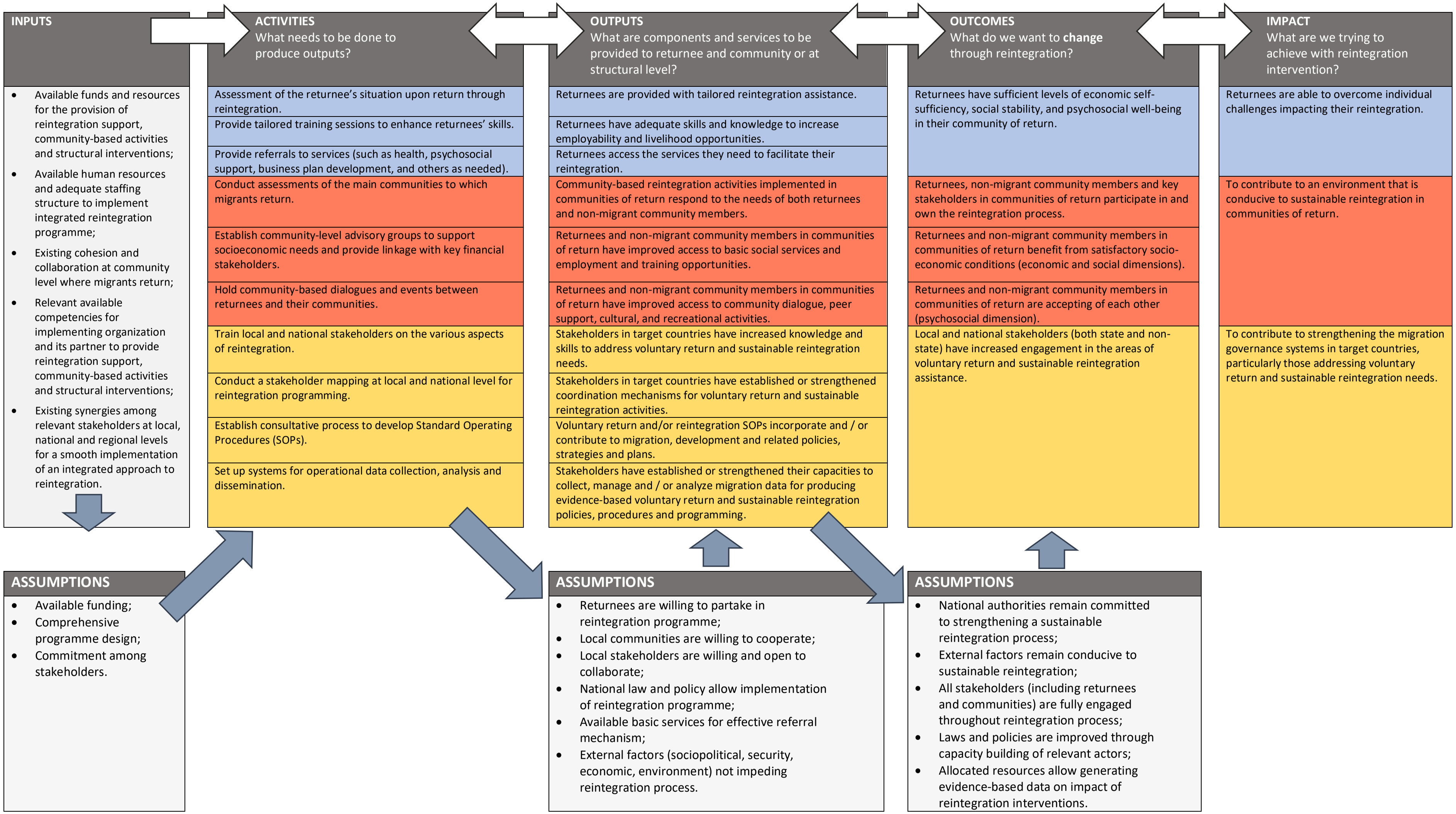
- A programme theory
- Evaluation criteria and evaluation questions

A programme theory

A programme theory, such as a theory of change or impact pathway, that maps out the expected cause and effect framework for an intervention.

On the next page, you can see an example of a theory of change.

Theory of change example



INPUTS

- Available funds and resources for the provision of reintegration support, community-based activities and structural interventions;
- Available human resources and adequate staffing structure to implement integrated reintegration programme;
- Existing cohesion and collaboration at community level where migrants return;
- Relevant available competencies for implementing organization and its partner to provide reintegration support, community-based activities and structural interventions;
- Existing synergies among relevant stakeholders at local, national and regional levels for a smooth implementation of an integrated approach to reintegration.

ASSUMPTIONS

- Available funding;
- Comprehensive programme design;
- Commitment among stakeholders.

ACTIVITIES
What needs to be done to produce outputs?

- Assessment of the returnee's situation upon return through reintegration.
- Provide tailored training sessions to enhance returnees' skills.
- Provide referrals to services (such as health, psychosocial support, business plan development, and others as needed).
- Conduct assessments of the main communities to which migrants return.
- Establish community-level advisory groups to support socioeconomic needs and provide linkage with key financial stakeholders.
- Hold community-based dialogues and events between returnees and their communities.
- Train local and national stakeholders on the various aspects of reintegration.
- Conduct a stakeholder mapping at local and national level for reintegration programming.
- Establish consultative process to develop Standard Operating Procedures (SOPs).
- Set up systems for operational data collection, analysis and dissemination.

OUTPUTS
What are components and services to be provided to returnee and community or at structural level?

- Returnees are provided with tailored reintegration assistance.
- Returnees have adequate skills and knowledge to increase employability and livelihood opportunities.
- Returnees access the services they need to facilitate their reintegration.
- Community-based reintegration activities implemented in communities of return respond to the needs of both returnees and non-migrant community members.
- Returnees and non-migrant community members in communities of return have improved access to basic social services and employment and training opportunities.
- Returnees and non-migrant community members in communities of return have improved access to community dialogue, peer support, cultural, and recreational activities.
- Stakeholders in target countries have increased knowledge and skills to address voluntary return and sustainable reintegration needs.
- Stakeholders in target countries have established or strengthened coordination mechanisms for voluntary return and sustainable reintegration activities.
- Voluntary return and/or reintegration SOPs incorporate and / or contribute to migration, development and related policies, strategies and plans.
- Stakeholders have established or strengthened their capacities to collect, manage and / or analyze migration data for producing evidence-based voluntary return and sustainable reintegration policies, procedures and programming.

ASSUMPTIONS

- Returnees are willing to partake in reintegration programme;
- Local communities are willing to cooperate;
- Local stakeholders are willing and open to collaborate;
- National law and policy allow implementation of reintegration programme;
- Available basic services for effective referral mechanism;
- External factors (sociopolitical, security, economic, environment) not impeding reintegration process.

OUTCOMES
What do we want to **change** through reintegration?

- Returnees have sufficient levels of economic self-sufficiency, social stability, and psychosocial well-being in their community of return.
- Returnees, non-migrant community members and key stakeholders in communities of return participate in and own the reintegration process.
- Returnees and non-migrant community members in communities of return benefit from satisfactory socio-economic conditions (economic and social dimensions).
- Returnees and non-migrant community members in communities of return are accepting of each other (psychosocial dimension).
- Local and national stakeholders (both state and non-state) have increased engagement in the areas of voluntary return and sustainable reintegration assistance.

ASSUMPTIONS

- National authorities remain committed to strengthening a sustainable reintegration process;
- External factors remain conducive to sustainable reintegration;
- All stakeholders (including returnees and communities) are fully engaged throughout reintegration process;
- Laws and policies are improved through capacity building of relevant actors;
- Allocated resources allow generating evidence-based data on impact of reintegration interventions.

IMPACT
What are we trying to achieve with reintegration intervention?

- Returnees are able to overcome individual challenges impacting their reintegration.
- To contribute to an environment that is conducive to sustainable reintegration in communities of return.
- To contribute to strengthening the migration governance systems in target countries, particularly those addressing voluntary return and sustainable reintegration needs.

Evaluation criteria and evaluation questions

Evaluation criteria and evaluation questions, based on the programme theory, define what the programme needs to evaluate.

Below, you can see an example of evaluation questions from the IMPACT study.

EVALUATION QUESTIONS EXAMPLE

These are the evaluation questions used in the Impact Evaluation of the EU-IOM Joint Initiative for Migrant Protection and Reintegration in the Horn of Africa region (IMPACT).

You can find out more in the methodological report.

Objective 1

What is the impact of the EU-IOM Joint Initiative (Horn of Africa (HoA)) on sustainable reintegration of supported migrant returnees?

- Have changes in programme implementation, such as the transition to mobile money, effected outcomes of reintegration assistance and, if so, how?
- How has delay in providing assistance to returnees affected/impacted on their reintegration?
- How have the EU-IOM Joint Initiative (HoA) adapted the assistance provided to meet changes in context and what has the impact of these changes been on the reintegration of returnees?

Objective 2

How can sustainable reintegration metrics be improved?

- Does the current assisted voluntary return and reintegration (AVRR) data chain collect sufficient information to assess "sustainable reintegration"?
- Does the Reintegration Sustainability Index appropriately capture local context and provide the empirical basis for appropriate programme intervention decisions, including opportunities for analysis of drivers of reintegration and drivers of remigration, and determine which of those can be affected by AVRR programme implementation?

Objective 3

How can we effectively evaluate impact of reintegration programmes in the future and what are the methodological requirements to do so?

- As definitions of reintegration often reference the non-migrant residents as a comparison, how can this cohort be meaningfully included in the data chain and contribute to an understanding of sustainable reintegration?
- Is there evidence to support the W model theory and what are the implications for evaluative methodologies assessing the effects of reintegration assistance?

The process of planning an intervention should include the development of a programme theory that maps out the changes (the outcomes and impacts) that the intervention is expected to produce and how these will come about (i.e. the activities and outputs).

Specifying the desired outcomes and impacts allows for the definition of how these will be evaluated – which indicators will be used and how they will be measured.

The programme theory can take different forms, varying from a logic or cause model through to a theory of change.

Types of programme theory

Theory of change

A theory of change (ToC) maps out how, why and for whom change is expected to occur in the intended context, describing the overall theory of how changes happen for an intervention to achieve its intended goal.

- Illustrates the chain of results from activities > outputs > outcomes > objectives. This is called the "pathway of change" or the "causal pathway"; a ToC may include several of these, both those directly related to the intervention and those that are not.
- Captures the dynamic and complex network of pathways to change.
- Explains the connection between an intervention and the effect it causes, by examining the logic and assumptions at each stage in the causal pathways. This includes external influences as well as those related to the intervention.

Logical framework (also called a results matrix)

Outlines the linear hierarchy of inputs, outputs, outcomes and impacts, including specifying the indicators, sources of information and means of verification that can be used to measure progress towards the intended results.

- Focused on the causal sequence for the intervention, rather than the wider picture.
- Assumptions made are the mostly external factors that need to be in place for one stage to lead to the next.

Find out more

Reintegration Handbook - Practical guidance on the design, implementation and monitoring of reintegration assistance

Section 5.2.1: p.175 Theory of Change

Section 5.2.2: p.178 Results framework

Section 5.4: p.186 Managing an Evaluation

PDF: *https://publications.iom.int/system/files/pdf/iom_reintegration_handbook.pdf#page=183*

IOM Monitoring and Evaluation Guidelines

Section 3.2: p.44 Programme Theory

Section 3.3: p.45 Theory of Change

Section 3.4: p.56 Results Matrix

Section 5.2: p.220 Evaluation criteria

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=58*

KNOWLEDGE CHECK

Read this case study and think about whether or not an impact evaluation would be an appropriate choice in this scenario:

A project was set up, funded by Country A's government to provide support to voluntary returnees to assist them in returning and reintegrating in their countries of origin.

The project provided transportation and a single reintegration support payment six months after their return for each returnee. The programme lasted two years with returnees joining the programme at different times.

The government initiative funding the project wants some evidence that their money was spent wisely and that the programme was effective compared to the cost of the migrants remaining in Country A. They have provided a budget for programme evaluation in the funding.

IOM provides non-food item kits, medical assistance, water and sanitation facilities and onward transportation assistance to vulnerable and stranded returnees in Renk. © IOM 2012/Samantha DONKIN

Question 1

Review the five key aspects to consider when deciding whether to conduct an impact evaluation:

Do you know who the intended users are?

□ Yes

□ No

Is there a **cl**ear use case for an impact evaluation?

□ Yes

□ No

Would the evaluation be complete in time to be useful?

□ Yes

□ No

Would an evaluation be operationally relevant for stakeholders?

□ Yes

□ No

Are there sufficient resources to carry out the evaluation?

□ Yes

□ No

Question 2

Taking the above into consideration, answer the following question:

Is an impact evaluation appropriate in this scenario?

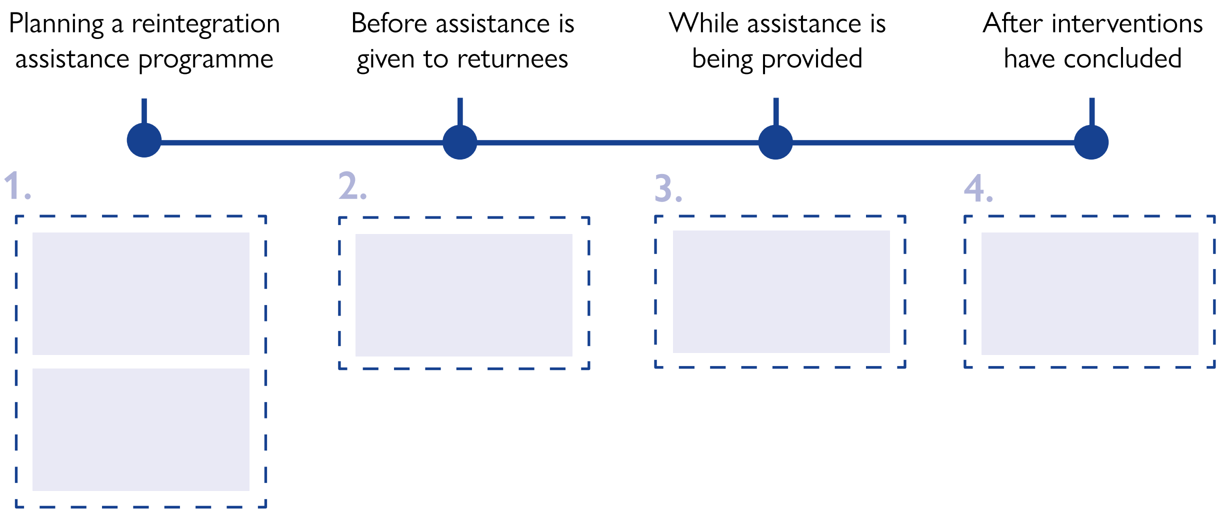
□ Yes

□ No

Question 3

Match each of the actions to their correct position on the "Stages of an impact evaluation" timeline.

Stages of an impact evaluation



Planning a reintegration assistance programme

Before assistance is given to returnees

While assistance is being provided

After interventions have concluded

Actions

A. Baseline data collection

B. Collection of midline data

C. Decide whether to carry out an impact evaluation

D. Developing a programme theory

E. Collection of endline data

Question 1 answer

Review the five key aspects to consider when deciding whether to conduct an impact evaluation:

Do you know who the intended users are?

☑ Yes

□ No

The government initiative has requested an evaluation of the programme.

Is there a clear use case for an impact evaluation?

□ Yes

☑ No

There does not appear to be an interest in repeating or collecting lessons learned from the programme. The intended uses for information indicated by the stakeholders are about justifying use of funds.

Would the evaluation be complete in time to be useful?

□ Yes

☑ No

As it is a fairly short-term programme, there likely wouldn't be time for an impact evaluation to be useful for improving implementation during the programme activity. There is not an intention to repeat it, so an impact evaluation after the conclusion would be too late.

Would an evaluation be operationally relevant for stakeholders?

□ Yes

☑ No

There does not appear to be an interest in repeating or collecting lessons learned from the programme, nor of undertaking similar programmes in the future.

Are there sufficient resources to carry out the evaluation?

☑ Yes

□ No

We don't have information about resources other than money, but there is funding for an evaluation in the budget.

Question 2 answer

Is an impact evaluation appropriate in this scenario?

Answer: No

While of course this is a simplified example and there may be other factors, we would argue that an impact evaluation is not appropriate in this situation.

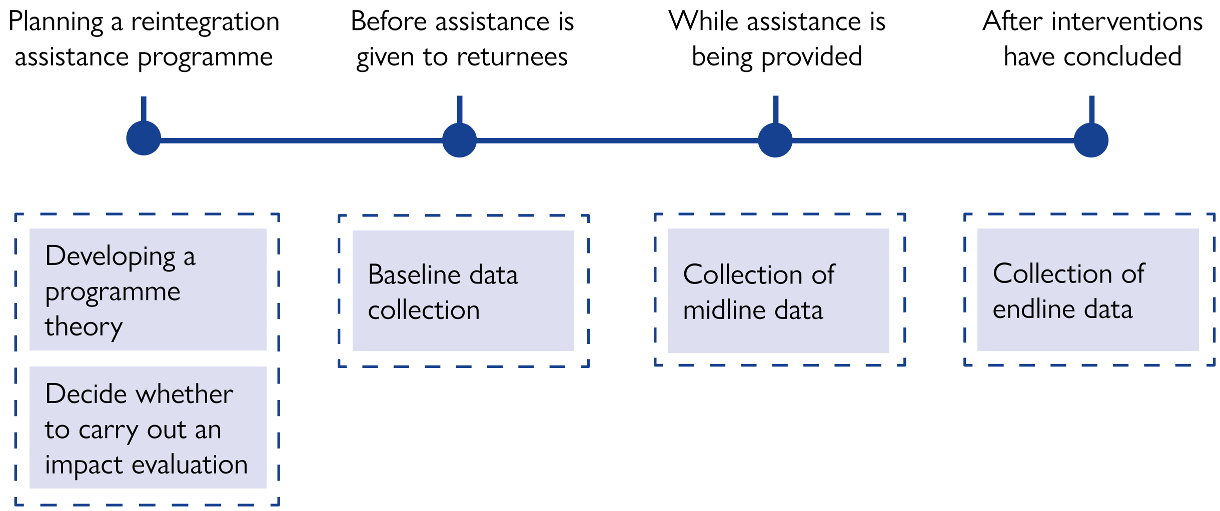
- The project was aimed at short-term support, and any long-term impacts that might be observed post-programme would be unlikely to be attributable to this support.
- The project itself was intended to be a one-off and was fairly short-lived. There does not seem to be a need or appetite to capture lessons learned.
- While the project is required to produce an evaluation, they also seem interested in efficient use of their money, so it would be hard to justify a costly impact evaluation when a smaller-scale evaluation would meet their information requirements.

Question 3 answer

Match each of the actions to their correct position on the “Stages of an impact evaluation” timeline.

Answers: 1D, 1C, 2A, 3B, 4E

Stages of an impact evaluation


Planning a reintegration assistance programme
Before assistance is given to returnees
While assistance is being provided
After interventions have concluded
Developing a programme theory
Decide whether to carry out an impact evaluation
Baseline data collection
Collection of midline data
Collection of endline data

Developing a programme theory
A programme theory is a crucial aspect of the planning stages of a programme.

Decide whether to carry out an impact evaluation
This should be incorporated into the planning of the programme, to allow for things like control groups and baseline data collection to be implemented.

Baseline data collection
Baseline data needs to capture the situation before interventions take place.

Collection of midline data
Midline refers to data being collected midway through the intervention.

Collection of endline data
Endline data are collected at the end of the activity being evaluated.

THE RETURN AND REINTEGRATION CONTEXT

HOW A RETURN AND REINTEGRATION CONTEXT AFFECTS THE IMPACT EVALUATION PROCESS

For return and reintegration programmes, there are a number of considerations and contextual factors that influence the way an impact evaluation can be designed and conducted.

Question

Which of the following factors do you think might affect the design of an impact evaluation for return and reintegration programme activities?

- □ Difficulty of identifying returnees that are not participating in the programme.
- □ Newly returned returnees don’t always go back to their community of origin and may change residences shortly after returning.
- □ Distressing events experienced by returnees prior to their return.
- □ Multiple/tailored interventions received by different participants.

Answer

- ☑ Difficulty of identifying returnees that are not participating in the programme.
- ☑ Newly returned returnees don’t always go back to their community of origin and may change residences shortly after returning.
- ☑ Distressing events experienced by returnees prior to their return.
- ☑ Multiple/tailored interventions received by different participants.

All of these are considerations that must be taken into account when planning and conducting an impact evaluation for return and reintegration interventions.

CONTEXTUAL CHALLENGES

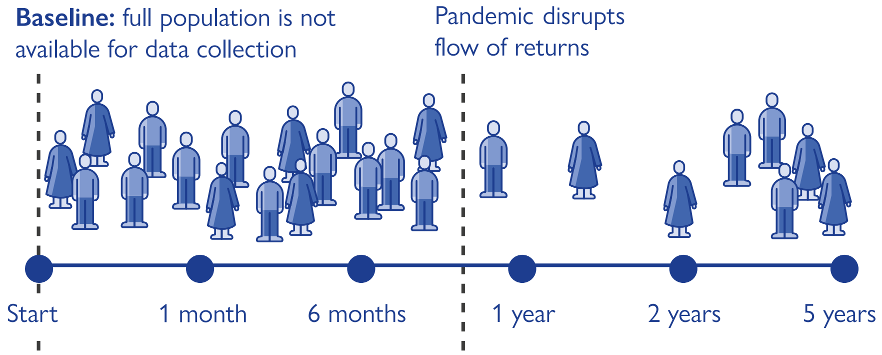
The IMPACT study we have been looking at is an example of an impact evaluation that is taking place in a return and reintegration context. Let's look at some key challenges that needed to be considered for this study.

ROLLING RECRUITMENT

Programme participants are not all recruited at once into the programme, as returns may happen over a long period of time. This means you don't have a full population to select from for the impact evaluation baseline data collection (Figure 14).

The flow of returns may also be disrupted by external factors such as a pandemic or political instability, with consequences for the determination of sample sizes and the planning of data-collection activities.

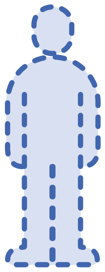
FIGURE 14: ROLLING RECRUITMENT INTO A PROGRAMME


Baseline: full population is not available for data collection
Pandemic disrupts flow of returns
Start
1 month
6 months
1 year
2 years
5 years

MOBILE POPULATION

Returnees can:

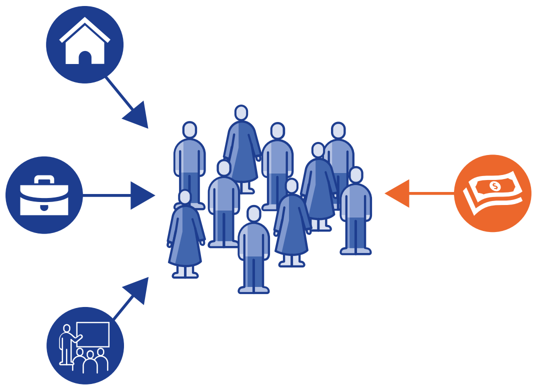
- Change phone number frequently
- Change addresses frequently
- Have limited connectivity
- Remigrate



This makes it difficult to follow returnees and get in touch to collect data.

PRESENCE OF OTHER PROGRAMMES

While IOM was offering interventions like in-kind support to establish a micro-business, housing support and business training, other related assistance was being provided by different organizations: for example, the United Nations High Commissioner for Refugees (UNHCR) was providing cash support to some returnees, on top of the assistance provided to IOM, making it harder to isolate the effect of IOM's activities alone.

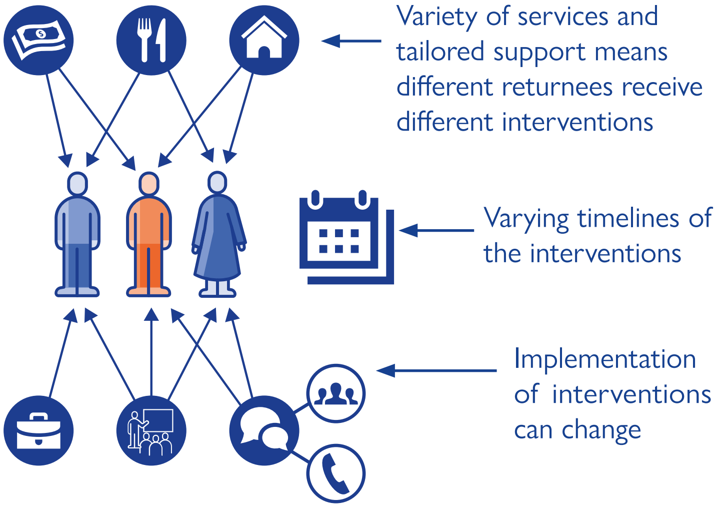


RANGE OF ASSISTANCE

The tailored interventions and breadth of services offered by the programme make it challenging to understand the effects of different interventions. These interventions have different timelines and are expected to contribute to successful reintegration in different ways. This complicates the process of collecting and analysing data.

To add further complication, the methods for providing interventions can change over time; for example, some interactions might begin to be conducted via phone due to the COVID-19 pandemic. (Figure 15)

FIGURE 15: RANGE OF ASSISTANCE



Variety of services and tailored support means different returnees receive different interventions

Varying timelines of the interventions

Implementation of interventions can change

CONDUCTING INTERVIEWS

The IMPACT study faced some difficulties with conducting interviews:

- COVID-19 posed risks for face-to-face interviews.
- Establishing trust via phone interviews is tricky; this makes it hard to get quality data, especially for qualitative information.
- Conflict and insecurity may also limit returnees' ability and willingness to speak openly.

These are just a few examples; there are many possible ones.

There is no single design "recipe" for any impact evaluation, and this is especially true in return and reintegration contexts, where the situation is complex and sensitive. As you will see in the following modules, adaptations will need to be made throughout an impact evaluation design and implementation process to allow for the specifics of the setting.

CRITICISMS OF IMPACT EVALUATION

OVERVIEW OF CRITICISMS OF IMPACT EVALUATION

As has been shown previously in this module, impact evaluations have many potential benefits and are becoming increasingly accepted and encouraged. However, there are some criticisms of the use of impact evaluations and it is important to be aware of these.

Structural

There is an argument that impact evaluations emphasize individual agency and behaviour over structural problems.

For example, large structural factors, such as global trade or the COVID-19 pandemic, could well affect reintegration and migration routes and motives, but these are outside the scope of what can be measured using impact evaluation methods.

Generalizability

Most impact evaluations are done on small scale and are limited in terms of locality. Once the programme scale increases, some of the effects of the programme might not be observed. It is also uncertain to what extent the results observed are applicable to other contexts.

For example, although the Joint Initiative programme covers West and Central Africa, North Africa and Horn of Africa, the IMPACT study is focused on Ethiopia, Somalia and the Sudan. Can the findings be generalized to West African countries? Moreover, even within the three countries, there are limits to how much results can be generalized; for example, depending on the sample, the results may not apply to some of the "older" returnees who may be underrepresented or not represented at all in the sample.

Partiality

Often, impact evaluation is limited to individual interventions within a broader programme, because these are easier to measure. What is evaluated might get excessive focus and importance placed upon it, even if it is a smaller component of a large programme.

For example, an impact evaluation might be conducted specifically on the economic component of a reintegration programme.

ETHICAL CONSIDERATIONS FOR IMPACT EVALUATIONS

Ethical issues need to be considered carefully as, in impact evaluation designs more than elsewhere, they have practical consequences.

- Is it justifiable to exclude potential beneficiaries from the assistance to populate a control or comparison group?
- Should you provide incentives for participation of returnees in the evaluation?
- The evaluation results have real potential to affect decision-making.

As with any humanitarian and development work, awareness of and adherence to ethical standards is paramount. It is recommended to review the ethical guidance from your organization for a full understanding.

Find out more

Reintegration Handbook - Practical guidance on the design, implementation and monitoring of reintegration assistance

Page 174 - Section 5.1.1. Ethical Considerations for M&E

PDF: *https://publications.iom.int/system/files/pdf/iom_reintegration_handbook.pdf#page=182*

IOM Monitoring and Evaluation Guidelines

Page 24 - Section 2.1. Professional norms and standards in monitoring and evaluation

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=38*

United Nations Evaluation Group Ethical Guidelines

The guidelines can be downloaded from the website

Web page: *www.unevaluation.org/document/detail/2866*

Let's think about some ethical considerations that affect a study.

For each of the considerations below, consider the ethical implications and make a note your answers to the questions. Suggested answers and feedback are given on page 44.

CONSIDERATION 1

An IOM study on migrant debt reveals that 56 per cent of returnees incurred debt to finance their migration project. © IOM 2021

For many return and reintegration programmes, the vast majority of beneficiaries are young men. What ethical concerns do you think may arise from this?

Take some time to think and note down your thoughts, then compare them to our suggestions on page 44.

CONSIDERATION 2

The most rigorous possible design for any impact evaluation of a reintegration programme would use a control group of randomly allocated returnees who receive no intervention. Which of the following statements do you think apply to a study that is considering using this approach? Select all the answers you think are correct.

- □ It would be important to show transparency in the random allocation of returnees to intervention or control.
- □ All participants must voluntarily agree to participate and be aware that they might be in the control group.
- □ This approach could allow harm and suffering by withholding support from those who need it.
- □ It is always best to use a control group, due to the benefits of the information gained from the process.

CONSIDERATION 3

Since it is often ethically unacceptable to exclude some returnees from receiving assistance for the purpose of populating a comparison group, what alternative options can be considered that might reduce ethical challenges?

Take some time to think and note down your thoughts, then compare them to our suggestions on page 45.

A monitoring and evaluation mission to Kismayo, Somalia. © IOM 2016/ Mary-Sanyu OSIRE

CONSIDERATION 4

Hargeisa Group Hospital, Somalia.
© IOM 2013/Mary-Sanyu OSIRE

In the IMPACT study, it was decided to use non-migrants residing in the same communities as the returnees as a comparison group. This was partly due to the ethical concerns of using a comparison group made up of returnees in the programme. What ethical considerations might be relevant to this option?

Take some time to think and note down your thoughts, then compare them to our suggestions on page 45.

Suggested answers and solutions to the questions are given below. Note that our suggestions are not the only valid answers, and you may have thought of something quite different.

CONSIDERATION 1

Here are our suggestions:

- There are many ethical aspects related to the fact that female returnees may be under-represented in the sample.
- Culturally, it may not be appropriate to interview women from male-headed households. However, without interviewing them, how can you understand the impact of the programme on women?
- Interview questions may not be tailored to female returnees and this may limit the study's ability to identify gender-sensitive effects of the intervention.

CONSIDERATION 2

Answers	Feedback
☑ **It would be important to show transparency in the random allocation of returnees to intervention or control.**	There should not be a possibility or suspicion of unfairness in any process that would affect the assistance returnees receive.
☑ **All participants must voluntarily agree to participate and be aware that they might be in the control group.**	Informed consent is vital, and it would be unjustifiable to create a situation where someone agrees to participate in assisted return on the understanding they will receive support, only for this support to be withheld.
☑ **This approach could allow harm and suffering by withholding support from those who need it.**	Returnees are often dependent on programme assistance and withholding it from those who need it could put them at further risk.
☐ It is always best to use a control group, due to the benefits of the information gained from the processs.	If creating a control group entails actively causing harm by withholding crucial support, this is not an option that can be considered for that situation, regardless of the potential benefits of the study.

CONSIDERATION 3

Here are our suggestions:

Grouping returnees according to certain characteristics

Consider grouping returnees according to certain characteristics – different levels of participation in the programme, for example – and using those distinctions to create comparison groups. This approach would come with issues of analysis and self-selection bias.

Evaluate a single intervention from a wider programme

Conduct the study not at programme level, but for a single intervention or in a subset of locations, whereby the "comparison" group are receiving other programme interventions rather than no intervention. For example, a project chose people from a wider programme to form the comparison group for participants receiving a mentoring intervention.

Stepped-wedge approach

As we saw in an earlier section of this module, it's possible to create a "stepped-wedge" design that staggers interventions such that comparison group participants join the programme later.

Consider ethics early

Such decisions can have a significant effect on the scope of the evaluation, so the ethical implications of the evaluation should be considered early, not as an afterthought.

CONSIDERATION 4

Here are our suggestions:

Reluctance to be interviewed

Host communities are not receiving direct interventions and so might be (rightly) reluctant to be repeatedly interviewed.

Benefits should balance out risk

Benefits should balance out the risk taken on by participating. Members of the host community may be sacrificing their time and taking on some risk (e.g. sharing of their data) by participating in the study, for potentially no benefits. For any comparison group which does not receive direct interventions there is an issue of "fairness".

Non-migrants can benefit from programme interventions

One option would be to provide benefits to non-migrant comparison groups later, or to provide community-level interventions, although they do not receive the individual-level interventions.

Compensating respondents

Another possibility would be to compensate respondents for the time spent in interviews. This can be a helpful solution, but there are some ethical issues and other concerns, including:

- The common practice for compensation in the programme locations (i.e. what other programmes have been doing);
- The appropriate level of compensation given the time requirements expected from respondents Whether compensation is only paid to those who have no likelihood of benefiting from the programme;
- The potential influence of the compensation, both on the respondents' willingness to participate (i.e. is their consent given voluntarily?) and possible bias in their responses.

We provide further references on this topic page 230.

QUIZ

This quiz will check your understanding of the topics covered in this module. There are six questions. You must get a score of at least five out of six to pass.

1. Which of the following is the best definition of an impact evaluation? Select one answer.

   □ An evaluation conducted at the end of a programme to report on the effectiveness and quality of implementation.

   □ An evaluation that measures the effects of an intervention and which can attribute those effects to the intervention.

   □ A study required by donors for justification of expenditure.

   □ A model that maps out the expected cause-effect framework for programme activities.

2. Which of the following are possible reasons why an impact evaluation should be carried out? Select all the answers that apply.

   □ It is a requirement from the donor.

   □ To meet with IOM guidance, every programme should include an impact evaluation.

   □ Information about the effectiveness of an intervention is required to help decide whether to continue with it.

   □ To provide evidence for advocacy and fundraising purposes.

3. Which of these is the key difference between impact evaluation and other common kinds of evaluation? Select one answer.

   □ It is summative – it is carried out at the end of a programme activity.

   □ It is formative – it is carried out during a programme activity.

   □ It aims to establish that the intervention being evaluated was the cause of the observed impact(s).

   □ It does not need to be planned until the intervention has begun.

4. Which of the following tasks are necessary to complete before commencing an impact evaluation? Select all the answers that apply.

   □ Develop a programme theory.

   □ Identify or establish a control group.

   □ Have a clear understanding of what the information produced by the evaluation will be used for.

   □ Obtain sufficient funding, confirm staff availability and time to conduct the evaluation.

   □ Receive a mandate from the donor that you must carry out an impact evaluation.

5. Which of the following is a method for assessing causal attribution? Select one answer.

   □ Developing a programme theory.

   □ Comparing recipients of an intervention with a comparison group both before and after.

   □ Conducting a survey asking intervention beneficiaries how useful an intervention has been.

6. Which of the following are ethical considerations that should be taken into account when designing and carrying out an impact evaluation? Select all the answers that apply.

- □ The inclusion of a group (comparison or control) who have no chance of benefiting from the programme.
- □ Ensuring that evaluation participants voluntarily agree to contribute.
- □ Making sure that only those who have been involved in the programme implementation are involved in the evaluation.
- □ Ensuring that sensitive and personally identifiable data are kept confidential.
- □ Where possible, collecting data without the knowledge of the subjects so that the results are not influenced.

QUIZ ANSWERS

There are six questions. You must get a score of at least five out of six to pass.

1. Which of the following is the best definition of an impact evaluation? Select one answer.

   □ An evaluation conducted at the end of a programme to report on the effectiveness and quality of implementation.

   ☑ An evaluation that measures the effects of an intervention and which can attribute those effects to the intervention.

   □ A study required by donors for justification of expenditure.

   □ A model that maps out the expected cause-effect framework for programme activities.

2. Which of the following are possible reasons why an impact evaluation should be carried out? Select all the answers that apply.

   ☑ It is a requirement from the donor.

   □ To meet with IOM guidance, every programme should include an impact evaluation.

   ☑ Information about the effectiveness of an intervention is required to help decide whether to continue with it.

   ☑ To provide evidence for advocacy and fundraising purposes.

3. Which of these is the key difference between impact evaluation and other common kinds of evaluation? Select one answer.

   □ It is summative – it is carried out at the end of a programme activity.

   □ It is formative – it is carried out during a programme activity.

   ☑ It aims to establish that the intervention being evaluated was the cause of the observed impact(s).

   □ It does not need to be planned until the intervention has begun.

4. Which of the following tasks are necessary to complete before commencing an impact evaluation? Select all the answers that apply.

   ☑ Develop a programme theory.

   □ Identify or establish a control group.

   ☑ Have a clear understanding of what the information produced by the evaluation will be used for.

   ☑ Obtain sufficient funding, confirm staff availability and time to conduct the evaluation.

   □ Receive a mandate from the donor that you must carry out an impact evaluation.

5. Which of the following is a method for assessing causal attribution? Select one answer.

   □ Developing a programme theory.

   ☑ Comparing recipients of an intervention with a comparison group both before and after.

   □ Conducting a survey asking intervention beneficiaries how useful an intervention has been.

6. Which of the following are ethical considerations that should be taken into account when designing and carrying out an impact evaluation? Select all the answers that apply.

☑ The inclusion of a group (comparison or control) who have no chance of benefiting from the programme.

☑ Ensuring that evaluation participants voluntarily agree to contribute.

□ Making sure that only those who have been involved in the programme implementation are involved in the evaluation.

☑ Ensuring that sensitive and personally identifiable data are kept confidential.

□ Where possible, collecting data without the knowledge of the subjects so that the results are not influenced.

SUMMARY

In this module, we have seen that:

1. An impact evaluation is a particularly intensive kind of evaluation, which measures impacts that can be attributed to an activity.

2. Impact evaluation outputs can provide extremely useful information about the effects of interventions, helping inform decision-making, provide evidence for stakeholders and improve reintegration assistance in the future.

3. Attributable impact is measured using comparisons over time and using comparison groups.

4. Impact evaluations require careful consideration and they should ideally be planned as the programme activity is being designed.

5. The reintegration context has a large effect on how impact evaluations can be conducted and introduces challenges in terms of ethics, the practicalities of conducting a study and interpreting the information produced.

MODULE 3 PART 1: WHAT, WHO AND HOW?

Abdulhalim had been working as a welder for a number of years when his landlord kicked him out for failing to cover his rent. That's when he decided to go to Libya. But he returned to the Sudan in August 2019 after six months as part of the assistance provided by the EU- Joint Initiative in 2019. After receiving economic reintegration assistance, he bought equipment and materials for his business, which is based in his hometown, El Geneina, West Darfur. © IOM 2021/Muse MOHAMMED

MODULE 3 PART 1: WHAT, WHO AND HOW?

INTRODUCTION

The first part of this module explains the key decisions that make up the process of designing an impact evaluation.

OUTCOMES

At the end of this module, trainees will be able to:

- List design options for an impact evaluation of return and reintegration programming using quantitative methods.
- Explain advantages and disadvantages of commonly used impact evaluation designs in the return and reintegration context.
- Explain the role of randomized control trials and quasi-experimental designs in evaluating return and reintegration programme impacts.
- Describe commonly used quantitative methods for impact evaluations.

CASE STUDY

In this module, we will continue to follow the EU-IOM Joint Initiative for Migrant Protection and Reintegration through the process of designing the quantitative aspects of the IMPACT study to evaluate the impact of the programme.

See page 9 for the description of the EU-IOM Joint Initiative and the associated IMPACT study.

Once the EU-IOM Joint Initiative programme decided that an impact evaluation was necessary, the next step was to design the evaluation that they would carry out.

DESIGNING AN IMPACT EVALUATION: OVERVIEW

When you measure the impact of a programme, you need empirical evidence as the basis for the evaluation. Often, the bulk of the empirical evidence will come from quantitative impact evaluation; the methods for this are the focus of this module.

This will sometimes be proceeded or followed up by qualitative research to help formulate the quantitative questions, or to help explain patterns in the quantitative results. More about qualitative methods and how they can be combined with quantitative approaches can be found in Module 5.

When discussing evaluation studies, the generic term "treatment" is often used to refer to an activity, such as the intervention, programme or policy (or combination of these), that is being evaluated.

Designing an impact evaluation study is essentially the process of making a series of decisions.

Broadly speaking, it means answering the questions:

What will be measured?

What are you evaluating?

- What is the treatment and how is it expected to work?
- What change do you expect to see?
- What evaluation questions are you answering?
- What indicators can be used to measure the change you expect to see?

Who will you take measurements about?

On whom are you measuring the impact and at which levels?

- Who and where are the population you are evaluating?
- Is it looking at individuals, households, or communities? Is it looking at the structural level – such as the local government?
- Is there a need to "stratify" observed respondents – i.e. separate subgroups of interest, such as livelihood zones, location, respondents' gender and/or age?

How will you measure the impact?

What methods will you use to conduct and compare measurements?

- How are you getting the information about the effect of the treatment? (Primary data, secondary data such as administrative data.)
- How many observations are required to get the data to inform the impact evaluation? e.g. baseline, midline, endline.
- How can you establish a meaningful counterfactual?
- How, and to what extent, can causal attribution be established?

In the following sections of this module, we will work through each of the questions, "what?", "who?" and "how?" for the IMPACT case study and examine the quantitative options for the design of an impact evaluation.

There is no single "correct" impact evaluation design. In return and reintegration work, as with many other fields, details of a project, users' information needs and contextual factors vary, and the evaluation design must reflect these.

Balancing priorities

The information given in this module is largely focused on the implications of quantitative design choices for the relevance of the collected data and subsequent findings. However, it is important to consider what is financially and practically feasible to implement, as well as what will produce useful data.

The quantitative enquiry can often form the largest cost of an impact evaluation. Very often, there are budgets and cost limitations that prevent the "best" impact evaluation design from being implemented and thus design compromises must be made.

PREPARING TO DESIGN AN IMPACT EVALUATION

THEORETICAL BASIS

To answer any of the questions of "what will be measured?", "who will you take measurements about?" and "how will you measure the impact?" that make up your study design, it is necessary to first have a clear theoretical basis for the programme or intervention you intend to evaluate.

Question

Which of the following would ideally be established before designing an impact evaluation?

(Hint: this was covered in Module 2)

- □ A baseline survey.
- □ The expected chain of cause and effect that shows how the programme activities are expected to generate the intended impacts.
- □ Evaluation criteria and evaluation questions.
- □ The unintended consequences of an intervention.

The answers are on the next page

Answer

□ **A baseline survey.**
While this should certainly be carried out before the treatment is implemented, the baseline survey is often, but not always, a component of an impact evaluation.

☑ **The expected chain of cause and effect that shows how the programme activities are expected to generate the intended impacts.**
This is the programme theory. It is crucial for an impact evaluation and should also be established when planning a programme.

☑ **Evaluation criteria and evaluation questions.**
These should be established as part of planning a programme and are a vital initial component of planning an impact evaluation.

□ **The unintended consequences of an intervention.**
These are more likely to be discovered during the impact evaluation. However, it is still useful to consider these when planning an impact evaluation and there may be options to control or measure these during the impact evaluation.

The process of planning a programme should include the development of a programme theory, such as a theory of change, which maps out the causal pathways of how the intervention is expected to produce the desired impacts. Ideally, the evaluation questions and criteria would also be established either prior to the planning of an impact evaluation or early in the process. These are important to have in place before conducting any evaluation.

Please refer to the Reintegration Handbook and IOM Monitoring and Evaluation Guidelines for more information on these concepts:

Reintegration Handbook - Practical guidance on the design, implementation and monitoring of reintegration assistance

Section 5.2.1: p.175 Theory of change

Section 5.2.2: p.178 Results framework

Section 5.4: p.186 Managing an evaluation

PDF*: https://publications.iom.int/system/files/pdf/iom_reintegration_handbook.pdf#page=183*

IOM Monitoring and Evaluation Guidelines

Section 3.2: p.44 Programme theory

Section 3.3: p.45 Theory of change

Section 3.4: p.56 Results matrix

Section 5.2: p.220 Evaluation criteria

PDF*: https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=58*

Once the programme theory is in place, it should be clear what is expected to change as a result of the programme or intervention and what needs to be measured to understand this change.

There is already a clear idea of what changes are expected; and in some cases, such as a logical framework, the extent of that change may also be set as a milestone target.

THEORETICAL BASIS FOR AN IMPACT EVALUATION

Before planning an intervention or carrying out an impact evaluation, it is necessary to have a clear theoretical basis for the programme or intervention. Let's clarify what this means in terms of where impact evaluation fits in the broader process, illustrated in Figure 16.

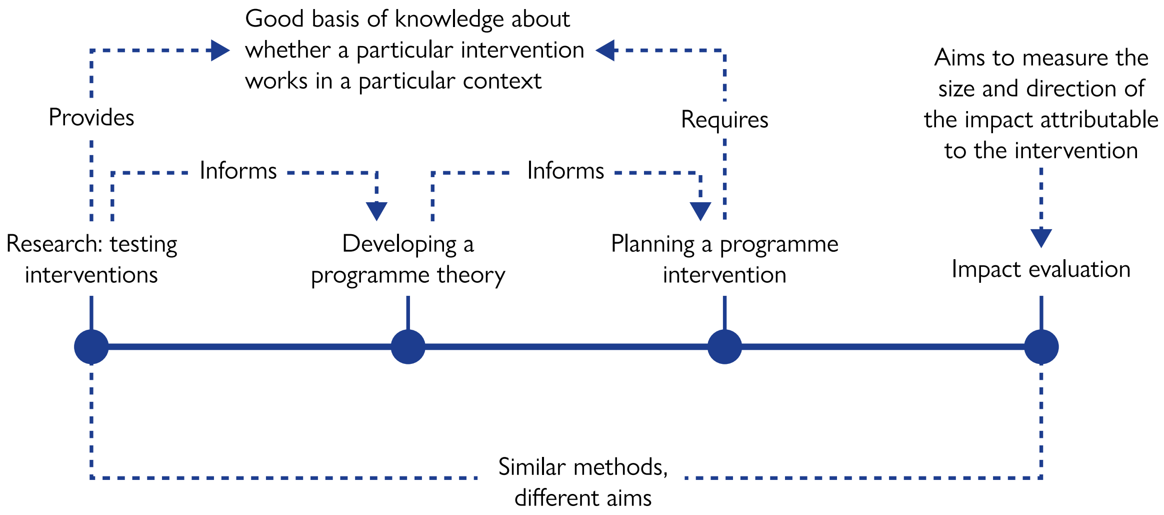
The main task of an impact evaluation is to measure the size and direction of the impact that can be attributed to the intervention being evaluated.

When developing a programme intervention, it is necessary to have a good basis of existing knowledge to justify application at scale. If there is no theoretical basis or existing knowledge about whether an intervention will work, then it shouldn't be implemented at scale. Instead, we need a system of testing and learning about whether a particular intervention works in a particular context. This might involve small-scale empirical trials or experiments to test a series of feasible, promising options.

While this process might include some of the same methods that are used for impact evaluation, such as randomized control trials or quasi-experiments, it is important to make the distinction between the aims of such work, which would be termed research, and the aims of carrying out an impact evaluation. The empirical evidence gathered during the research informs the development of the programme theory used to plan full-scale interventions.

By the time a full-scale intervention is in place and an impact evaluation is being carried out, the question should no longer be what type of effect is expected from the intervention but rather about the extent or size of the expected effect and about optimization – looking at whether particular implementation strategies or approaches work better than others.

FIGURE 16: THEORETICAL BASIS FOR AN IMPACT EVALUATION


Good basis of knowledge about whether a particular intervention works in a particular context
Provides
Requires
Aims to measure the size and direction of the impact attributable to the intervention
Informs
Informs
Research: testing interventions
Developing a programme theory
Planning a programme intervention
Impact evaluation
Similar methods, different aims

To give a simplified example, an impact evaluation is not used to establish whether we should feed people who are starving, but rather to determine whether it is most effective to provide food parcels, money, or vouchers.

WHAT WILL BE MEASURED?

The first key question when designing an impact evaluation is what will be measured. This is concerned with defining precisely what activities are to be evaluated, what mechanisms of cause and effect are expected and may be tested and what indicators should be used to assess the impact.

With the programme theory and evaluation in place, much of this is already established.

- What is the treatment and how it is expected to work?
- What change do you expect to see?
- What evaluation questions are you answering?
- What indicators can be used to measure the change you expect to see?

Please refer to the Reintegration Handbook and Monitoring and Evaluation Guidelines for more information on programme theories and developing evaluation questions:

Reintegration Handbook - Practical guidance on the design, implementation and monitoring of reintegration assistance

Section 5.2.1: p.175 Theory of change

Section 5.2.2: p.178 Results framework

Section 5.4: p.186 Managing an evaluation

PDF: *https://publications.iom.int/system/files/pdf/iom_reintegration_handbook.pdf#page=183*

IOM Monitoring and Evaluation Guidelines

Section 3.2: p.44 Programme theory

Section 3.3: p.45 Theory of change

Section 3.4: p.56 Results matrix

Section 5.2: p.220 Evaluation criteria

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=58*

What is the treatment?

The term “treatment” refers to the activity that is being evaluated. It is crucial to define precisely what the treatment is for an evaluation; what, specifically, is being evaluated? As has been mentioned, the programme theory sets out the intended programme activities. However, an impact evaluation may aim to evaluate only a subset of the programme – perhaps a small number of interventions – or in only specific locations.

In the IMPACT study example, even though the Joint-Initiative has components in West and Central Africa, North Africa and the Horn of Africa, the impact evaluation is focused on the reintegration assistance work in Ethiopia, Somalia and the Sudan only.

SELECTING INDICATORS

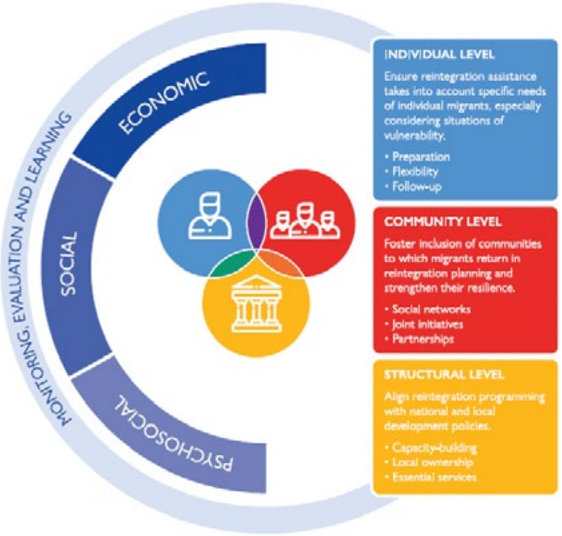
Determining the indicators that will be measured in an impact evaluation for return and reintegration can be challenging. The indicators you choose need to measure the change you expect to see as a result of the treatment, and they should be suitable to answer your evaluation questions.

Various factors affect this decision, including:

- Project goals
- Agency priorities
- Data availability
- Uses for information (see page 29)
- Users of information (see page 29)

In many cases, the aim is to look at the treatment's (the activity that is being evaluated) impact on "reintegration". However, reintegration is not a simple concept to define or measure. There is not one harmonized definition, and it is a combination of multiple factors (see Figure 17).

FIGURE 17: IOM INTEGRATED APPROACH TO REINTEGRATION


MONITORING, EVALUATION AND LEARNING
ECONOMIC
SOCIAL
PSYCHOSOCIAL
INDIVIDUAL LEVEL
Ensure reintegration assistance takes into account specific needs of individual migrants, especially considering situations of vulnerability.
• Preparation
• Flexibility
• Follow-up
COMMUNITY LEVEL
Foster inclusion of communities to which migrants return in reintegration planning and strengthen their resilience.
• Social networks
• Joint initiatives
• Partnerships
STRUCTURAL LEVEL
Align reintegration programming with national and local development policies.
• Capacity-building
• Local ownership
• Essential services

MEASURES OF REINTEGRATION

There is currently not a single, agreed-upon way reintegration is measured and calculated. Possible approaches include:

- Using a "composite index", which combines different factors into a single "score"
- Using multiple separate values in a "dashboard" approach.

Establishing how these are calculated, and which measurements are used to do so, is part of the challenge of deciding how to measure reintegration. Various groups, including IOM, have used a range of approaches and de facto definitions. For example, IOM has established this definition:

> "Reintegration can be considered sustainable when returnees have reached levels of economic self-sufficiency, social stability within their communities, and psychosocial well-being that allow them to cope with (re)migration drivers. Having achieved sustainable reintegration, returnees are able to make further migration decisions a matter of choice, rather than necessity."

Towards an Integrated Approach to Reintegration in the context of return (IOM, 2017) p.3

PDF: *www.iom.int/sites/g/files/tmzbdl486/files/our_work/DMM/AVRR/Towards-an-Integrated-Approach-to-Reintegration.pdf#page=5*

Based on this definition, IOM and its partners developed the Reintegration Sustainability Survey and the related composite index. The index combines 31 indicators of reintegration across economic, social and psychosocial dimensions to produce a single reintegration score.

Reintegration Sustainability Survey and the related composite index:

PDF: *https://returnandreintegration.iom.int/sites/g/files/tmzbdl341/files/documents/knowledge_bite_1_-_introduction_0.pdf*

When planning an impact evaluation for return and reintegration programmes, it may be possible to use one of the existing ways of measuring reintegration, perhaps with some adjustments.

Find out more about defining and measuring reintegration on page 111.

The definition and measurements of impacts for an evaluation need to be agreed with the intended users of the evaluation's outputs to ensure the information produced is relevant and useful.

For example, if an organization is interested in the financial and health-related concerns of returnees, evaluating their programme's impact based on returnees' perceptions of reintegration may not provide sufficient detail on the financial and health-related impacts.

KNOWLEDGE CHECK

Question 1

Which of the following terms is being described in the definition below?

"The specific activities that are being evaluated."

- □ Intervention
- □ Programme
- □ Treatment
- □ Reintegration assistance

Question 2

Which of the following statements are true? Select all that apply.

- □ An impact evaluation for return and reintegration programmes must measure reintegration.
- □ Reintegration is often measured by combining multiple factors into a single "composite index".
- □ The definition and selection of indicators should be based only on the evaluation questions.
- □ The definition of reintegration is subjective, and there is not one harmonized definition that is used.

Question 1 answer

Which of the following terms is being described in the definition below?

"The specific activities that are being evaluated."

- □ Intervention
- □ Programme
- ☑ Treatment
- □ Reintegration assistance

In an impact evaluation, the "treatment" could include reintegration assistance. It could be a single intervention or multiple interventions. The treatment is whatever is going to be evaluated.

Question 2 answer

Which of the following statements are true? Select all that apply.

- □ **An impact evaluation for return and reintegration programmes must measure reintegration.** While many such impact evaluations will be looking at the impact on reintegration, this is not always the case and can vary depending on the goals of the evaluation.
- ☑ **Reintegration is often measured by combining multiple factors into a single "composite index".** This is indeed a common option for measuring reintegration.
- □ **The definition and selection of indicators should be based only on the evaluation questions.** While the evaluation questions are important to consider when determining the indicators that will be measured in an impact evaluation, other factors – such as project goals, agency priorities, data availability and the intended uses for the evaluation outputs – are also relevant.
- ☑ **The definition of reintegration is subjective, and there is not one harmonized definition that is used.** This is true. While agencies like IOM may have established definitions, it is a complex concept that can mean different things to different people.

CASE STUDY EXAMPLE

Here is an extract from the evaluation questions for the IMPACT study:

Objective 1

What is the impact of the EU-IOM Joint Initiative (Horn of Africa (HoA)) on sustainable reintegration of supported migrant returnees?

- Have changes in programme implementation, such as the transition to mobile money, effected outcomes of reintegration assistance and, if so, how?
- How has delay in providing assistance to returnees affected/impacted on their reintegration?
- How have the EU-IOM Joint Initiative (HoA) adapted the assistance provided to meet changes in context and what has the impact of these changes been on the reintegration of returnees?

Objective 2

How can sustainable reintegration metrics be improved?

- Does the current assisted voluntary return and reintegration (AVRR) data chain collect sufficient information to assess "sustainable reintegration"?
- Does the Reintegration Sustainability Index appropriately capture local context and provide the empirical basis for appropriate programme intervention decisions, including opportunities for analysis of drivers of reintegration and drivers of remigration, and determine which of those can be affected by AVRR programme implementation?

Objective 3

How can we effectively evaluate impact of reintegration programmes in the future and what are the methodological requirements to do so?

- As definitions of reintegration often reference the non-migrant residents as a comparison, how can this cohort be meaningfully included in the data chain and contribute to an understanding of sustainable reintegration?

With this in place, we have the starting point to answer many of the what/who/how questions we introduced earlier in the module.

The evaluation questions show that there are multiple dimensions to this evaluation. As well as evaluating the impact of the programme, there is a desire to use the IMPACT study as a test case for the Reintegration Sustainability Index (RSI) measure of reintegration and to learn about effective impact evaluation.

Looking at objective 1 and the related subquestions, can we make any decisions about what indicators we should measure to answer this question?

Question

What indicator are we expecting to see impacted? Choose one answer.

□ Sustainable reintegration

□ Returnee well-being

□ Delays to receiving reintegration assistance

□ Transition from in-kind assistance to mobile money

Answer

What indicator are we expecting to see impacted? Choose one answer.

- ☑ Sustainable reintegration
- □ Returnee well-being
- □ Delays to receiving reintegration assistance
- □ Transition from in-kind assistance to mobile money

The questions under objective 1 ask specifically about the effect on reintegration. Returnee well-being is an aspect of successful reintegration, but not the same thing. Delays to receiving reintegration assistance and transition from in-kind assistance to mobile money are factors whose effect on sustainable reintegration the study aims to discover.

WHO WILL YOU TAKE MEASUREMENTS ABOUT?

Having decided what to measure, the next question is: from whom (or what) are you taking measurements?

We need to know:

Hierarchy

Are the units part of a hierarchical structure that is relevant for data collection and analysis?

Units of observation

Who or what are you planning to collect data from? This might be individuals, families, households, organized social structures, administrative structures, etc.

Units of analysis

At what level of the hierarchy are you looking for differences and making conclusions?

Stratification

Do you need to collect data from different groups of units?

Identifying the above is a very important part of planning data collection, especially when sampling is to be used.

This module assumes an understanding of basic sampling concepts. To review this topic, refer to the IOM M&E Guidelines, page 121.

PDF: *https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=135*

This course only covers aspects of sampling that are specific to impact evaluations. However, resources about sampling, as well as other relevant topics, are provided in Module 7.

DESIGN HIERARCHY

Let’s look at an example.

A project that supports reintegration of returnees has established job centre offices where returnees can find information about jobs in their location and link up with organizations that are interested in supporting them with training, placements and/or jobs.

The programme works in three provinces of a country.

Within each province, six districts were selected to receive the intervention.

In each of those districts, two job centres were set up to provide services targeted to returnees.

These job centres are located quite far apart from each other, maximizing coverage of the district and meaning that returnees in a district only really have access to one of the two job centres.

An impact evaluation is being carried out. It aims to compare the impact of two different approaches provided by job centres to support returnees:

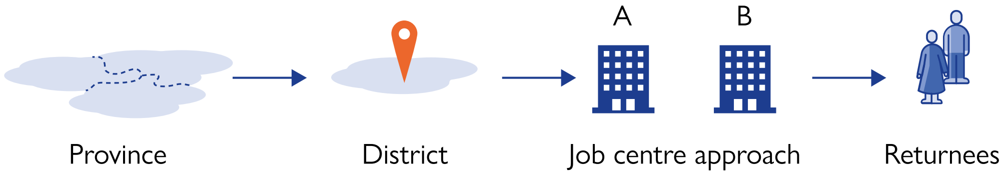
Approach A

Job centres offering services to groups of returnees, including group training on business skills and group sessions on how to interview for a job.

Approach B

Job centres offering services to individual returnees, focusing on one-to-one guidance and connecting returnees with organizations and employment opportunities.

In each district, one job centre provides services using approach A, and the other uses approach B. There is a hierarchy here, which is important to consider when both designing and analysing an impact evaluation:



UNITS OF OBSERVATION AND ANALYSIS

The example in the previous section showed a hierarchy of units and how interventions may happen at levels of the hierarchy other than the individual level. In such cases it is likely that the unit of observation and the unit of analysis will be different.

Unit of observation

- The unit or item you actually observe or measure;
- Determined by the data-collection method.

Unit of analysis

- The unit at the level at which the intervention takes place and the analysis conducted;
- Determined by the programme design (including hierarchy) and evaluation questions.

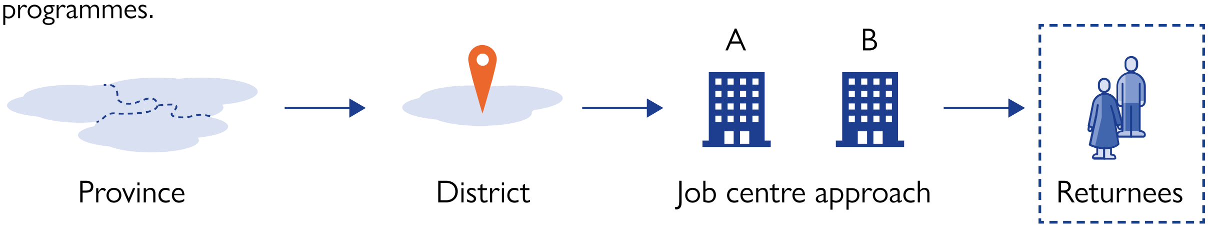
The most common units of analysis encountered in impact evaluations are individuals, households/families, communities, villages and geographical areas.

IDENTIFICATION OF UNITS OF OBSERVATION AND UNITS OF ANALYSIS

At what level of this hierarchy should the results be measured?

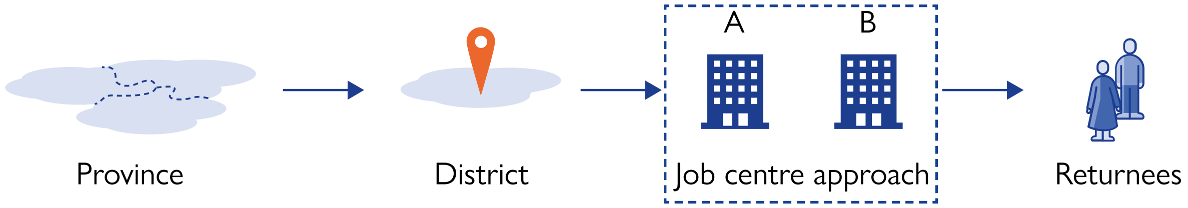
Units of observation

Reintegration, and the components used to define it, apply to individual people. Therefore, the units of observation will be the returnees. This is common for impact evaluations of return and reintegration programmes.



Units of analysis

However, the intervention in this example does not happen at the individual level. The impact evaluation is interested in seeing the differences among, and making conclusions about, job centres – not individual returnees. Therefore, the units of analysis are the job centres.



The unit of analysis needs to match the unit at which the intervention takes place. Returnees attending the same job centre are provided with the same services, so for the purposes of the evaluation, there are no important differences between the experiences of individual returnees attending the same job centre. This means that the intervention happens at job centre level.

When we consider the job centre as the unit of analysis for the impact evaluation, we will likely need to summarize the effects observed in the individual returnees served by a job centre; for example, we might look for an improvement in the total number of people employed, or the average income for returnees using that job centre.

ACTVITY

Let’s look at a few examples of potential studies and think about what the units of observation and unit of analysis would be. Add the missing units to the gaps.

Research question	Unit of observation	Data collection	Unit of analysis
What is the change in reintegration sustainability index since returnees returned to country of origin?	Returnee	Survey of returnees just after return to country of origin and 12–18 months later	
How does returnee-perceived trust in institutions vary by location?		Survey of returnees	
What proportion of returnee households have access to education?		Survey of returnee capturing data on their household children’s enrolment in school	
How does returnee reintegration vary by implementing partner?		Survey of returnees just after return to country of origin and 12–18 months later	

Units:

- **Implementing partner**
- **Household**
- **Returnee**
- **Location**
- **Index**

ACTVITY ANSWERS

Let's look at a few examples of potential studies and think about what the units of observation and unit of analysis would be.

Research question	Unit of observation	Data collection	Unit of analysis
What is the change in reintegration sustainability index since returnees returned to country of origin?	**Returnee**	Survey of returnees just after return to country of origin and 12–18 months later	**Returnee**

Feedback

Reintegration is measured at the level of individual returnees.

How does returnee-perceived trust in institutions vary by location?	**Returnee**	Survey of returnees	**Location**

Feedback

We have to ask returnees to find out about their trust in institutions. The survey is to be conducted on returnees. The research question specifies that we want to know about the differences between locations.

What proportion of returnee households have access to education?	**Returnee**	Survey of returnee capturing data on their household children's enrolment in school	**Household**

Feedback

- The survey is to be conducted with individual returnees, asking about their household.
- The question specifies that the interest is in the proportion of households that have access.

How does returnee reintegration vary by implementing partner?	**Returnee**	Survey of returnees just after return to country of origin and 12–18 months later	**Implementing partner**

Feedback

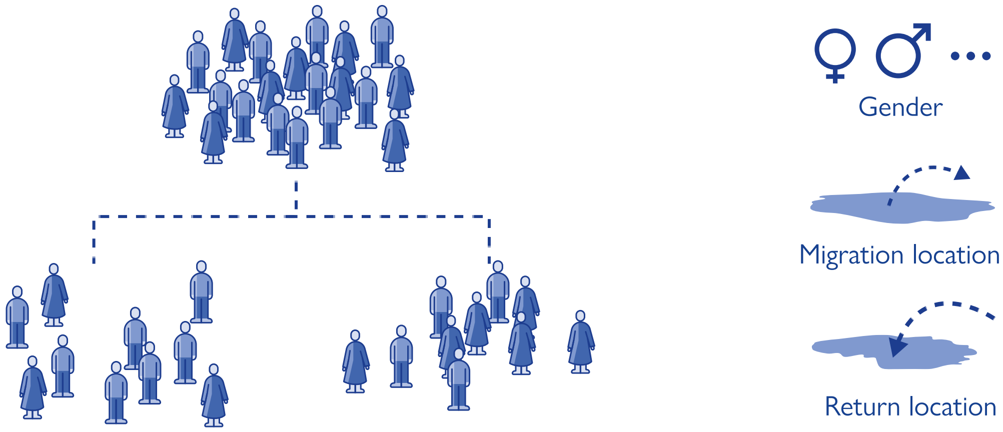
- The survey is to be conducted with returnees.
- The research question specifies that we want to know about the variation between different implementing partners. This will likely mean comparing the average reintegration index for the returnees for each implementing partner.

STRATIFICATION

Stratification is when the units are divided up into groups – or "strata" – for the sake of sampling and/or reporting. There are two major purposes for stratification: reporting by subpopulation and the representation of specific subgroups

Report by subpopulation

Thinking about the information needs of the evaluation, is it only necessary to make statements for the target population as a whole? Or is it also necessary to have separate results for subpopulations of this group?



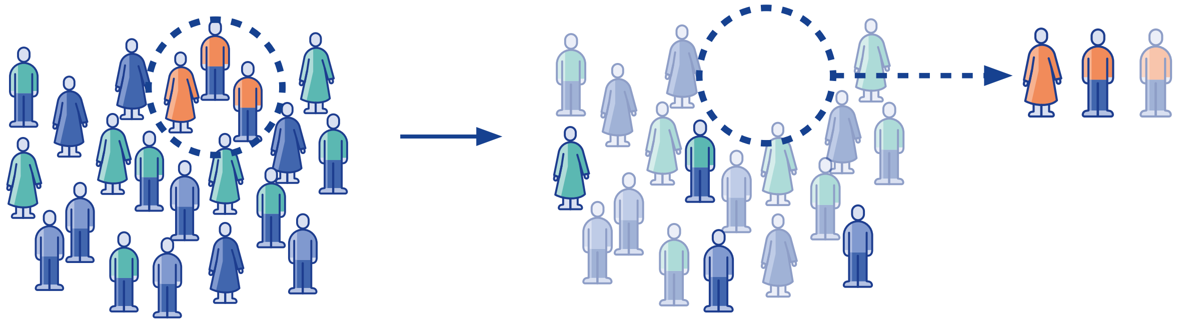
For example, do you need to be able to report separate results for different genders, migration location or return location? This last example could be relevant for programmes that want to use evaluation results to guide adaptive programming.

This increases the overall sample size requirements for a study; whatever sample size would have been needed from the whole population without stratification will be needed for each subpopulation, so that key impact evaluation indicators can be estimated with similar precision in each one. This will increase the cost and resources required for data collection.

Representation of specific subgroups

In some cases, it is possible that there are certain groups that will be missed if they are not specifically sampled. For example, there could be a particularly small subgroup, or one that is difficult to access. It is possible to create a strata for the group to ensure they are not missed from the sampled population (Figure 18).

FIGURE 18: STRATIFICATION TO ALLOW REPRESENTATION OF A SUBGROUP



A small subgroup of the population

A strata is created for the subgroup and sampled

However, this would then likely mean overrepresentation of that group, but this can then be corrected for later in the process, using a method called "weighting".

Find out more: Video: *www.youtube.com/watch?v=RXWuQkWcjPk*

There is also an approach we can take earlier in the process called sampling "probability proportional to size", where the resulting sample size in each subgroup will be different but approximately represent the same proportion of the total population within the subgroup.

Find out more: Video: *www.youtube.com/watch?v=I34wLgRgjQM*

After defining stratification, a sampling strategy is developed within each of the strata.

> Sampling strategies are outside the scope of this course, but you can find references to learn more about sampling on page 222.

KNOWLEDGE CHECK

Imagine you are planning an impact evaluation for a programme intervention that provides cash grants for returnees. The evaluation wants to measure the impact that the grants have on reintegration compared to the programme's standard approach of providing in-kind support.

There is reason to believe, based on previous trials, that the grants may have a bigger impact on reintegration for returnees on average, but benefit women less than the in-kind support.

Question 1

What will the unit of observation be for the study?

- □ Country
- □ Village
- □ Intervention
- □ Women
- □ Individual returnees

Question 2

In the scenario described above, do you need to use stratification in your study design?

- □ Yes
- □ No

KNOWLEDGE CHECK ANSWERS

Question 1 answer

What will the unit of observation be for the study?

- □ Country
- □ Village
- □ Intervention
- □ Women
- ☑ Individual returnees

The support is given to individual returnees, and we are interested in the differences at the level of individuals who were given different treatments.

Question 2 answer

In the scenario described above, do you need to use stratification in your study design?

- ☑ Yes
- □ No

Stratification is needed in this design, as the evaluation aims to report on the impact for men and women as well as the overall impact.

HOW WILL YOU MEASURE THE IMPACT?

DIMENSIONS FOR COMPARISON

The third key decision to make in designing an impact evaluation is how you will measure the impact.

Impact evaluations are all based on comparison. Any statement that something has "improved" or that one option "is better" only makes sense in relation to a point of comparison. After deciding what you will measure and from whom, the third major decision in designing a study is to establish what you are going to compare against.

As we have seen in Module 2, the two major dimensions for comparison are about time (before/after comparisons) and the presence of the treatment being evaluated (with/without comparisons). Generally, impact evaluations would include a combination of the two.

TIME-BASED COMPARISONS: RECAP

Let's recap what we learned in Module 2.

Question 1

Which of the following statements best describes how the change of an indicator over time can be measured?

- □ Collecting data on the indicator of interest and comparing it to the average for the population.
- □ Collecting data on the indicator of interest at two or more points in time, then calculating the difference between the two measurements.
- □ Using a survey that measures the indicator of interest at the time of data collection and also includes a question about what it was in the past.

Question 2

To measure the effect of a treatment, impact evaluations commonly use baseline and endline studies.

Match the term to the definition.

Terms	Definitions
Baseline	Measurement taken after the treatment has commenced but before the end of the study.
Midline	Measurement of the indicators of interest before the treatment is provided.
Endline	Measurement of the indicators of interest after the end of the treatment.

TIME-BASED COMPARISONS: RECAP ANSWERS

Question 1 answer

Which of the following statements best describes how the change of an indicator over time can be measured?

☐ **Collecting data on the indicator of interest and comparing it to the average for the population.**
This does not include a time dimension – it merely tells you about the indicator you measure compared to the average.

☑ **Collecting data on the indicator of interest at two or more points in time, then calculating the difference between the two measurements.**
Feedback: This is what we refer to as a before/after comparison.

☐ **Using a survey that measures the indicator of interest at the time of data collection and also includes a question about what it was in the past.**
This method is used sometimes but is not always ideal as people's recollection is often inaccurate.

Question 2 answer

To measure the effect of a treatment, impact evaluations commonly use baseline and endline studies.

Match the term to the definition.

Terms		Definitions
Baseline	→	Measurement of the indicators of interest before the treatment is provided.
Midline	→	Measurement taken after the treatment has commenced but before the end of the study.
Endline	→	Measurement of the indicators of interest after the end of the treatment.

When planning for data collection, the timing should be considered carefully:

Baseline

- It is crucial that this takes place before the treatment begins.
- A baseline should not be taken too long before treatment, as changes might occur between the baseline and the beginning of treatment that could be misattributed to the treatment.
- When does the treatment you are evaluating begin? Consider this carefully. For example, if a programme provides training assistance to returnees once they have returned from abroad, then the treatment begins when the returnee arrives at the location where they plan to settle.
- What change are you measuring? A treatment may begin before the returnee has left the migration location, but if the evaluation aims to look at the change in local reintegration, then the baseline would be conducted post-return, once the returnee is settled in a location but before they receive post-return benefits of the programme.

Midline

- Midlines can be taken at any time between baseline and endline for a variety of purposes. They might aim to capture trends and patterns, changes due to shocks, or short-term effects of aspects of the treatment, to give just a few examples.
- Timing of the midline measurements depends heavily on the purpose. For example, if the aim is to capture short-term effects caused by particular aspects of the treatment, then the midline should be late enough that there has been time for the treatment to have an effect and not so late that the effect can no longer be observed.
- The expected time for impacts to emerge and be observable will depend on the specifics of the treatment. This requires careful thought at the planning stage.

Endline

- The endline must take place after the treatment has finished.
- Much like the midline, deciding how long after the end of the treatment to take the endline will depend on the time it can reasonably be expected for the treatment to create an impact.

Read more about the issue of when to measure to observe impacts on page 133.

TIME-BASED COMPARISONS: LONGITUDINAL AND CROSS-SECTIONAL STUDIES

When planning to make before-after comparisons, there are two key options, depending on the availability of data and practicalities of data collection:

- Longitudinal design: observations are made on the same units for all the data-collection rounds (such as baseline, midline and endline);
- Repeated cross-sectional: a different sample of units is used each time.

The longitudinal design (also known as panel design) is the most commonly used in the humanitarian context.

LONGITUDINAL STUDIES: ADVANTAGES

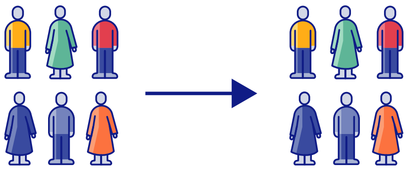
Less variation between respondent samples

As studies use the same sample each time, the characteristics of the respondents are fairly consistent throughout the study. This means that differences between baseline and later rounds of data collection show the change over time with little additional variation caused by the natural differences between people.

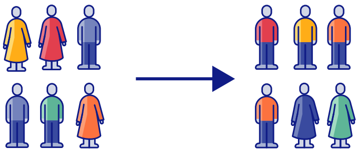
This means that the basic sample size can be smaller than for repeated cross-sectional surveys.

Bear in mind that the characteristics of people (and households, communities, etc.) do naturally change over time, so respondent characteristics may not be completely consistent.

Longitudinal study



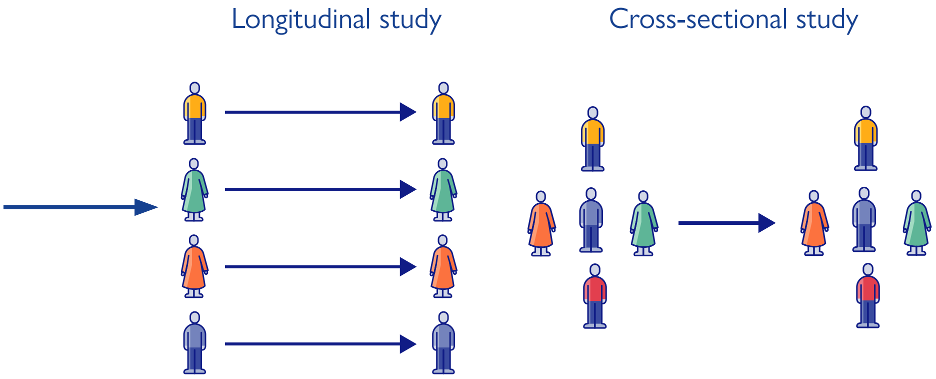
Cross-sectional study



Better understanding of change over time

With a longitudinal study, comparisons can be made of the same respondent at different points in time, so it is possible to measure the change for each respondent rather than the change between the aggregated results for all the respondents, as shown in Figure 19.

FIGURE 19: CHANGE OVER TIME IN LONGITUDINAL AND CROSS-SECTIONAL STUDIES



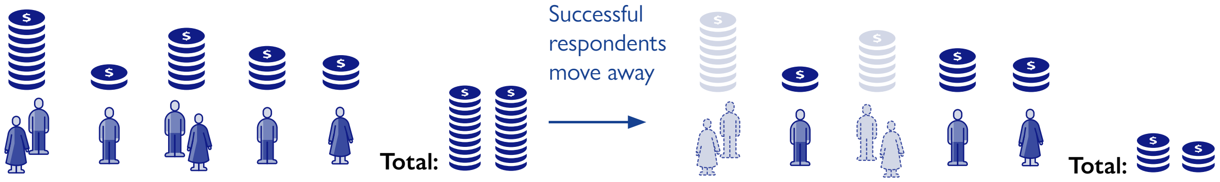
The results can be analysed at the respondent level, which can be helpful for investigating the characteristics that affect changes.

LONGITUDINAL STUDIES: DISADVANTAGES

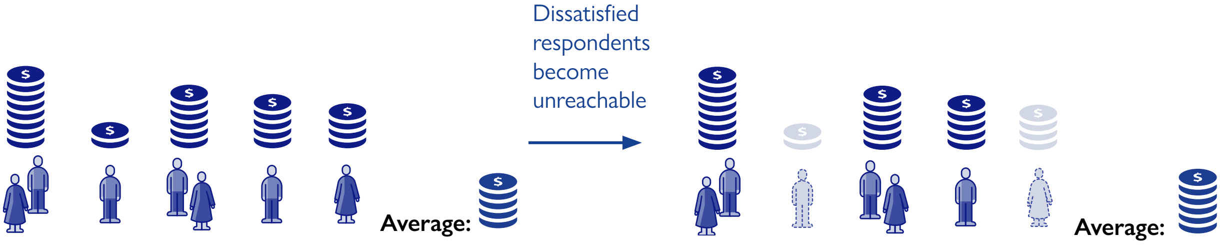
Attrition

Loss of respondents before all observations have been completed. Respondents that were interviewed for the baseline may then become unreachable or refuse to participate in later interviews.

It is possible for the causes of attrition to be related to the treatment, and you should try to understand what the causes for attrition are and how they may affect (bias) the results. For example, imagine households are doing very well due to programme activities and thus move to a location outside of the survey area after the baseline interview and become unreachable: this could lead to an underestimate of the effect of the treatment, as the results for the remaining respondents would be reduced.



Conversely, some respondents to the baseline interview may become unreachable because they are dissatisfied with the programme (e.g. in a particular area or receiving assistance from a particular implementing partner) and they don't see any benefit in being part of it: this could lead to an overestimation of the effect of the treatment.

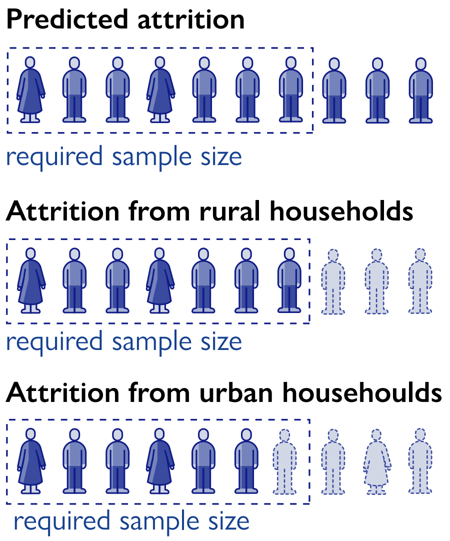


A well-designed longitudinal study anticipates some attrition and increases the planned sample size to allow for this. Decisions about this might be informed by learning from previous studies. For example, a baseline-endline survey previously conducted in Somalia recorded an attrition rate of 30 per cent – meaning that 30 per cent of the respondents to the baseline interview could not be reached or refused to do the endline interview.

Based on this experience, when planning another impact evaluation in Somalia, the project increased the baseline sample size by 30 per cent assuming that this would be the attrition rate.

Later in the process, after data collection took place, the project realised that they had different attrition rates within the sample: 30 per cent for rural households and 40 per cent for urban ones.

The +30 per cent adjustment was enough to compensate for the attrition in the rural households, but it was insufficient for the other group. This resulted in a loss of precision in the impact estimate for urban households.



Work needed to locate the respondents

Longitudinal studies rely on being able to return to the same respondents for each round of data collection. This can be challenging, particularly when the respondents are returnees, as they are more likely than other groups to change address and phone number. Significant time and effort may need to be expended to maintain contact and locate respondents who have moved. (See page 137 for more information.)

Personally identifiable information

To conduct longitudinal data collection, it is necessary to collect and store personally identifiable information on the respondents, such as their names, addresses and phone numbers. Collecting these data creates risk for the respondents and requires allocation of resources to keep them secure and confidential.

REPEATED CROSS-SECTIONAL STUDIES: ADVANTAGES

Ideally suited for assessing trends at the population or other aggregate levels

If there is no need to look at the changes over time at the observational unit level (e.g. returnee) then the repeated cross-sectional can provide this information without the complexities of conducting a longitudinal study.

Can save time and effort compared to longitudinal studies

When using a different sample each time, there is no need to maintain contact or work to locate the exact same respondents from previous rounds of data collection. Samples can simply be taken from the available population.

No need to account for attrition

The risk of attrition is not a concern for repeated cross-sectional studies and so the initial sample size does not need to be increased to allow for it. This can make cross-sectional studies the cheaper and easier option in many cases.

Less need for sensitive data

As there is no need to facilitate future contact with respondents, this reduces the likelihood that personally identifiable data will need to be collected. This means that the respondents may not need to take on the risk associated with sharing their potentially sensitive information and avoids the increased responsibility and data security requirements for the project.

- Note that if the survey itself involves collecting personally identifiable information, then the associated risks and requirements still apply regardless of whether a longitudinal or cross-sectional approach is used.

REPEATED CROSS-SECTIONAL STUDIES: DISADVANTAGES

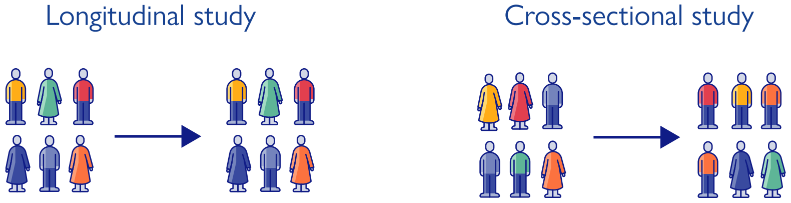
Change can only be understood at the aggregate level

Cross-sectional studies give more limited options for making comparisons compared to longitudinal studies; it is not possible to determine the actual change for specific households, only to compare aggregated results at different times.

Variation among observations

As cross-sectional studies use a different sample each time, there will naturally be some differences in the characteristics of the respondents in each sample. This adds variation into the results which can make it more difficult to isolate the real change over time.

This also leads to larger basic sample size requirements than for longitudinal studies.



Need to use same selection criteria each time

Although cross-sectional studies are, in general, easier to carry out, drawing repeated samples does require work to ensure that the selection criteria for the respondents remains the same over the different rounds of data collection.

In some disciplines, the term “longitudinal” is sometimes used more broadly to refer to any study that uses repeated measurements over time.

Throughout this course, we are using term specifically to refer to studies that take repeated measurements from the same units in each round of data collection.

TIME-BASED COMPARISONS: KNOWLEDGE CHECK

Question 1

What are the two options for time-based comparisons?

□ Longitudinal
□ Control group
□ Repeated cross-sectional
□ Counterfactual

Question 2

Match the definitions to the term:

Terms	Definitions
Longitudinal study	A study in which a new sample of units is used for each round of data collection.
Repeated cross-sectional study	A study in which the same units are used in each round of data collection.

Question 3

Which of the following statements are true? Select all that apply.

□ Repeated cross-sectional studies have a greater risk of attrition than longitudinal.
□ Longitudinal studies are preferable to repeated cross-sectional if they are practical to undertake.
□ Repeated cross-sectional studies allow analysis of changes at the individual (or respondent) level.

TIME-BASED COMPARISONS: KNOWLEDGE CHECK ANSWERS

Question 1 answer

What are the two options for time-based comparisons?

- ☑ **Longitudinal**
- □ **Control group**
  This is actually an option for a with/without comparison.
- ☑ **Repeated cross-sectional**
- □ **Counterfactual**
  This is not an aspect of time-based comparisons.

Question 2 answer

Match the definitions to the term:

Terms		Definitions
Longitudinal study	→	A study in which the same units are used in each round of data collection.
Repeated cross-sectional study	→	A study in which a new sample of units is used for each round of data collection.

Question 3 answer

Which of the following statements are true? Select all that apply.

- □ **Repeated cross-sectional studies have a greater risk of attrition than longitudinal.**
  Cross-sectional studies do not have a risk of attrition because they do not aim to retain units for multiple rounds of data collection.
- ☑ **Longitudinal studies are preferable to repeated cross-sectional if they are practical to undertake.**
  The downsides of longitudinal studies are generally related to the practical difficulties of carrying them out; the data they provide are in many circumstances more useful than data from repeated cross-sectional studies.
- □ **Repeated cross-sectional studies allow analysis of changes at the individual (or respondent) level.**
  Longitudinal studies allow this kind of analysis. Repeated cross-sections only allow change to be measured in aggregate; the change on average between the two samples can be measured but there is no way to look at the change for individuals.

WITH/WITHOUT COMPARISONS

We covered this kind of comparison in Module 2. Let’s review some key points.

Question 1

What is being described in the definition below?

“What would have happened if the treatment had not taken place. In practice, it is impossible to observe this “alternate reality”, so strategies are used to estimate what would have happened. We can compare an estimate of what would have happened without the treatment with the results for those that received the treatment to determine the likely impact that the treatment had.”

Choose one answer.

□ Control group

□ Counterfactual

□ Causal attribution

□ Comparison group

Question 2

Connect the statements below to the relevant label:

A group that has not received the treatment or has received an alternative

Produces the strongest possible evidence for casual attribution

At risk of bias

Might not have been randomly assigned

Studies should aim to make this as similar to the group receiving treatment

A randomly assigned group of units that do not receive the treatment

Can be challenging to establish due to issues of practicality and ethics

Control group

Comparison group

Both

Question 3

Which of the following is a definition of “causal attribution”?

□ Using a comparison group to estimate a counterfactual.

□ Determining that an action (such as an intervention or policy) was the cause of changes that have been observed.

□ Combining the with/without and before/after dimensions for comparison.

□ When factors other than the treatment have had an effect on the indicators being measured in an evaluation, making it difficult to understand the impact of the treatment.

WITH/WITHOUT COMPARISONS - ANSWERS

Question 1 answer

What is being described in the definition below?

"What would have happened if the treatment had not taken place. In practice, it is impossible to observe this "alternate reality", so strategies are used to estimate what would have happened. We can compare an estimate of what would have happened without the treatment with the results for those that received the treatment to determine the likely impact that the treatment had."

Choose one answer.

- □ Control group
- ☑ Counterfactual
- □ Causal attribution
- □ Comparison group

The principle behind comparing a group who received an intervention with a group who didn't is that we are trying to estimate the counterfactual. The counterfactual is the hypothetical situation that would have occurred without the intervention. By comparing the counterfactual to the result with the intervention, we can isolate the effect of the intervention.

However, it is of course impossible to go back in time and see what would have happened in the alternate reality where there was no intervention! Different strategies are therefore used to try to get as good an estimate as we can.

Question 2 answer

Connect the statements below to the relevant label:

Control group

Produces the strongest possible evidence for casual attribution

A randomly assigned group of units that do not receive the treatment

Comparison group

At risk of bias

Might not have been randomly assigned

Both

A group that has not received the treatment or has received an alternative

Studies should aim to make this as similar to the group receiving treatment as possible

Can be challenging to establish due to issues of practicality and ethics

Comparison group

A group that has not received the treatment or has received an alternative treatment. A comparison group is compared to a group that has received the treatment (sometimes called the "treatment group"), to understand the difference the treatment makes. Ideally, a comparison group would be as similar to the group receiving the treatment as possible.

Depending on the method used to determine who receives the treatment and who does not, there is a risk of the factors that affect whether a unit is assigned to receive the treatment also affecting the results. Various strategies are used to try to mitigate this.

The method for selecting a comparison group may be constrained by practical limitations or ethics.

Control group

A special kind of comparison group that is made by randomly selecting units to not receive the treatment (or to receive an alternative). In terms of being able to attribute an effect to a specific cause, this is a powerful design. However, it is not always practical, feasible, ethical or necessary to have a control group in an impact evaluation.

In development situations, it is difficult to have the level of flexibility needed to randomly assign programme activities to communities or households. The more usual situation is that the target communities and households are defined by programme design without room for randomization. Therefore, it is often necessary to estimate the counterfactual another way, such as using a comparison group that has not been randomly selected.

Question 3 answer

Which of the following is a definition of "causal attribution"?

□ **Using a comparison group to estimate a counterfactual.**
This is a strategy that is used to make a with/without comparison and may help establish causal attribution.

☑ **Determining that an action (such as an intervention or policy) was the cause of changes that have been observed.**
In an impact evaluation, causal attribution is the act of establishing whether – and to what extent – changes that have been observed were caused by the treatment. It is a very important goal of an impact evaluation.

□ **Combining the with/without and before/after dimensions for comparison.**
This is a method, sometimes called "difference in difference", that studies use to try to establish causal attribution. We will look at this method in more detail later in this module.

□ **When factors other than the treatment have had an effect on the indicators being measured in an evaluation, making it difficult to understand the impact of the treatment.**
This is referred to as "contamination" and is a common problem that can make it hard to achieve causal attribution. We will see more about this later in the module.

Note that the "without" in "with/without" does not always mean we are comparing against a group that has received no treatment at all – it may mean comparing a treatment to a different kind of treatment. In development work, it is rare to compare against no treatment. Returnees assisted in return and reintegration programmes may often be in need of immediate assistance and withholding treatment from them would be problematic. Unless you have a clear reason for why this is necessary, it's not usually an option to be considered, both for ethical reasons and for the effectiveness of the study.

The aim of an impact evaluation is often about optimization – to discover if an approach or kind of treatment is more effective (in terms of use of resources, overall impact, etc.) than others. Therefore, it is more informative to compare a treatment to other promising options or to the current standard practice.

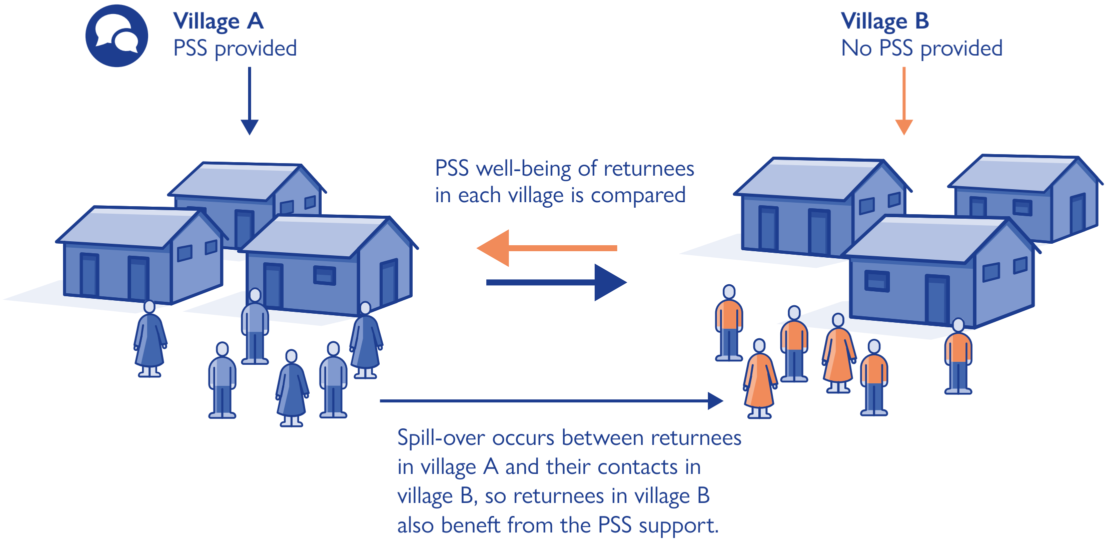
CONTAMINATION WHEN ESTABLISHING ATTRIBUTION

When you are conducting an impact evaluation, you need to ascertain that what you're measuring is the consequence of the treatment and not anything else. A major challenge to be addressed in any impact evaluation design is that of contamination. This could happen in three main ways:

TREATMENT EFFECTS "SPILL-OVER" FROM THE INTENDED BENEFICIARIES.

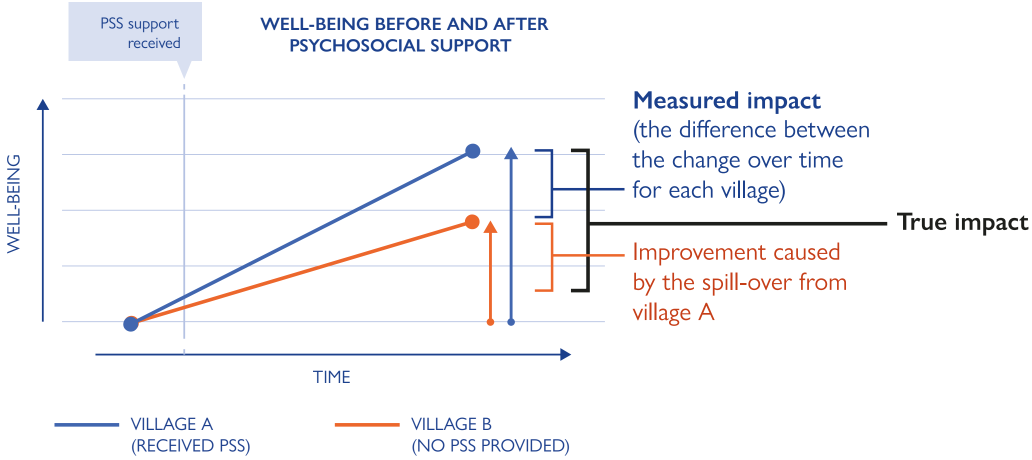
The treatment has effects that "spill-over" from the intended beneficiaries. If groups or individuals from the comparison group are indirectly affected by this, it will result in an inaccurate comparison.

For example, psychosocial support (PSS) session are provided to returnees in Village A (all enrolled in a return and reintegration programme) in an attempt to help them maintain a positive attitude after return. In a nearby village, Village B, there are also returnees enrolled in the same programme, but no PSS sessions have been organised there.


Village A
PSS provided
Village B
No PSS provided
PSS well-being of returnees in each village is compared
Spill-over occurs between returnees in village A and their contacts in village B, so returnees in village B also beneft from the PSS support.

The impact of the initiative in village A is measured by comparing the PSS well-being of returnees between the two villages. However, unbeknownst to the evaluators, several returnees of Village A know returnees in Village B through a local migrant association and shared what they had learned in the PSS sessions with these contacts. We can see the effect of this in Figure 20.

FIGURE 20: THE EFFECT OF SPILL-OVER ON THE MEASURED IMPACT



The PSS sessions resulted in improvements in PSS well-being for returnees in Village A. The sharing of the PSS session contents also resulted in improvement for returnees in Village B. As the PSS initiative is evaluated by comparing returnees in Village A with returnees in Village B, the improvement in Village B leads to the impact of the training in Village A being underestimated. This is a problem because it could lead to the PSS initiative not being repeated or funded further because it is believed to have less of an impact than it really did.

STUDY POPULATION IS AFFECTED BY SOMETHING OTHER THAN THE TREATMENT

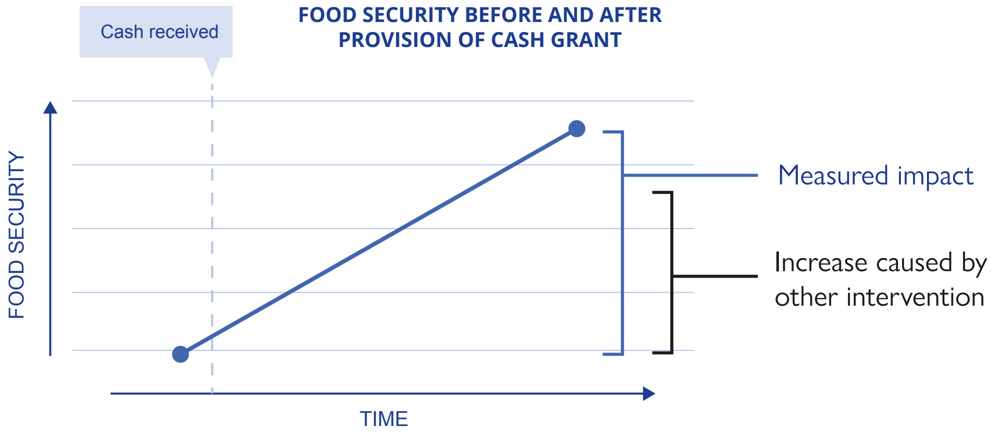
The study population – either the treatment group or the comparison group – is significantly affected by something other than the treatment. This could be another intervention, perhaps carried out by a different agency. If this results in similar impacts to those intended by the treatment, this could confuse the results. It could cause an underestimation as seen in the previous example, or an overestimation if the impact of the other interventions are mistakenly measured as the impact of the treatment.

For example, a return and reintegration programme is providing economic reintegration assistance using cash transfers. This intervention is being evaluated; food security is being measured to determine the impact of the assistance.

At the same time, a separate organization is running a programme in the area that provides "safety net" cash grants, aimed at vulnerable people. This includes some of the returnees receiving the economic reintegration assistance.

The evaluation records a significant increase in food security in the recipients of the economic reintegration assistance, and thus the implementation of this intervention is understood to be effective and considered for wider implementation elsewhere. However, the safety net cash grants were partly responsible for the increase in food security, and this could lead to the impact of the cash transfers being overestimated, as shown in Figure 21.

FIGURE 21: OVERESTIMATION OF IMPACT DUE TO CONTAMINATION



EXTERNAL FACTORS

External factors, like a government policy change or a major shock such as an earthquake, has effects on the results being measured to evaluate impact. Particularly if this has a disproportionate impact on one group or the other, this could result in over- or underestimation in a similar way to the previous examples.

A lot of effort in conducting impact evaluations goes into avoiding this contamination of results. How do we deal with contamination when it occurs?

In quantitative studies, it might be possible to reduce or eliminate the effect of contamination by "controlling for" it. This means that we collect data on potential contamination (such as shocks experienced, other programmes operating in a community) and essentially try to subtract the effect of the contamination from the change measured in the study; what is left is deemed to be the attributable effect of the intervention.

If contamination cannot be measured (as it is often the case in development contexts), the best we can do is to document and include it in the reporting of results, making it clear that it was not possible to confirm that the results obtained could be purely attributed to our programme. In such situations, the reporting of the results would make weaker statements about "contribution" rather than "attribution".

KNOWLEDGE CHECK

Question 1

Which of the following statements is true? Select all that apply.

□ It is important to document the contamination of results.

□ If you cannot prevent contamination, you can still claim attribution so long as it is recorded and acknowledge in the report.

□ A major effort in conducting impact evaluations goes into avoiding contamination of results.

□ Members of the treatment group sharing benefits of the treatment with members of the comparison group is not contamination, but simply an unintended impact of the treatment.

Question 2

Which one of the following statements best describes the comparisons that need to be made to establish attribution? Select one answer.

□ Comparing a group that has received the treatment with a group that has not received any treatment.

□ Comparing a group that receives the treatment with a group that is as similar as possible except for the treatment received.

□ Compare returnees who have received assistance with non-migrants.

KNOWLEDGE CHECK ANSWERS

Question 1 answer

Which of the following statements is true? Select all that apply.

☑ **It is important to document the contamination of results.**
This is important to ensure that the results are properly understood; not documenting contamination could lead to users of the evaluation believing the evidence for attribution to be much stronger than it actually is.

□ **If you cannot prevent contamination, you can still claim attribution so long as it is recorded and acknowledge in the report.**
While you may be able to claim contribution, it is not possible to claim pure attribution if your study is subject to contamination that cannot be accounted for in the analysis.

☑ **A major effort in conducting impact evaluations goes into avoiding contamination of results.**
Many of the requirements for conducting a "rigorous" impact evaluation are intended to avoid contamination.

□ **Members of the treatment group sharing benefits of the treatment with members of the comparison group is not contamination, but simply an unintended impact of the treatment.**
Comparison groups strive to represent how things would be in the absence of the treatment, so unintended impacts of the treatment on this group are contamination.

Question 2 answer

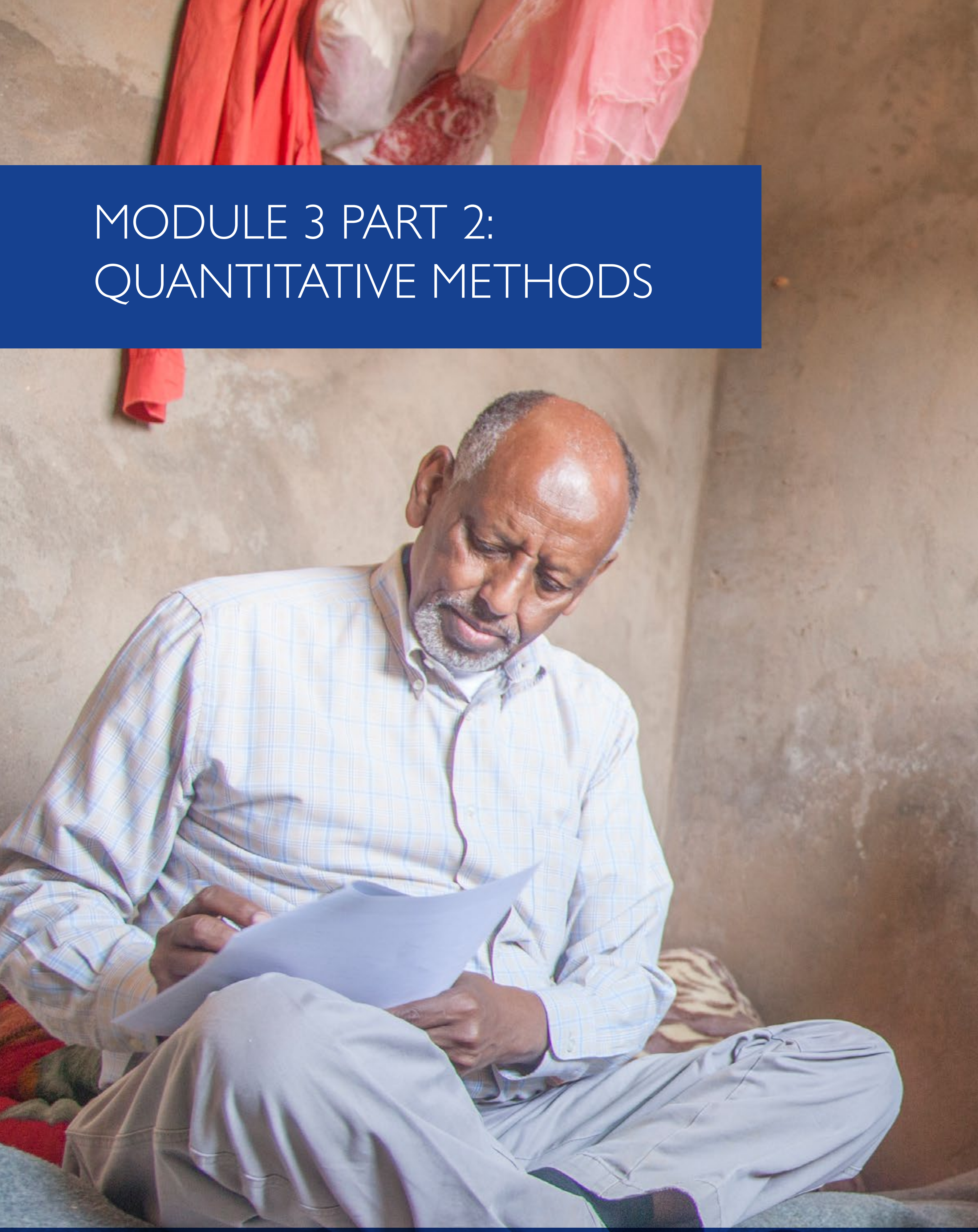
Which one of the following statements best describes the comparisons that need to be made to establish attribution? Select one answer.

□ **Comparing a group that has received the treatment with a group that has not received any treatment.**
While this may be the case, it is not necessary to compare with a group receiving no treatment at all; it may be more relevant to the aims of the study to compare with a different treatment.

☑ **Comparing a group that receives the treatment with a group that is as similar as possible except for the treatment received.**
This is correct. The other statements are options that may be applicable in some cases.

□ **Compare returnees who have received assistance with non-migrants.**
While this is sometimes an option that is used, it is not always the case.

MODULE 3 PART 2: QUANTITATIVE METHODS



On a monitoring and evaluation mission to Kismayo, Somalia in November 2016, the team met with beneficiaries at health facilities managed by IOM and the Ministry of Health, inspected a newly constructed settlement site for internally displaced persons and refugee returnees, and met with the stakeholders including the Government. © IOM 2016/Mary-Sanyu OSIRE

MODULE 3 PART 2: QUANTITATIVE METHODS

INTRODUCTION

This part of the module presents options that can be used to design a study to establish causal attribution and measure the effect of a treatment on indicators of interest.

OUTCOMES

At the end of this module, trainees will be able to:

- List design options for an impact evaluation of return and reintegration programming using quantitative methods.
- Explain advantages and disadvantages of commonly used impact evaluation designs in the return and reintegration context.
- Explain the role of randomized control trials and quasi-experimental designs in evaluating return and reintegration programme impacts.
- Describe commonly used quantitative methods for impact evaluations.

COMMONLY USED METHODS IN IMPACT EVALUATION

This part of the module presents options that can be used to design a study to establish causal attribution and measure the effect of a treatment on indicators of interest. An impact evaluation design might include a combination of more than one of these options.

Designs

- Randomized controlled trials
- Matching
- Propensity score matching
- Stepped-wedge design
- Regression continuity design
- Natural experiments

Analysis techniques

- Difference in difference
- Regression
- Instrumental variables

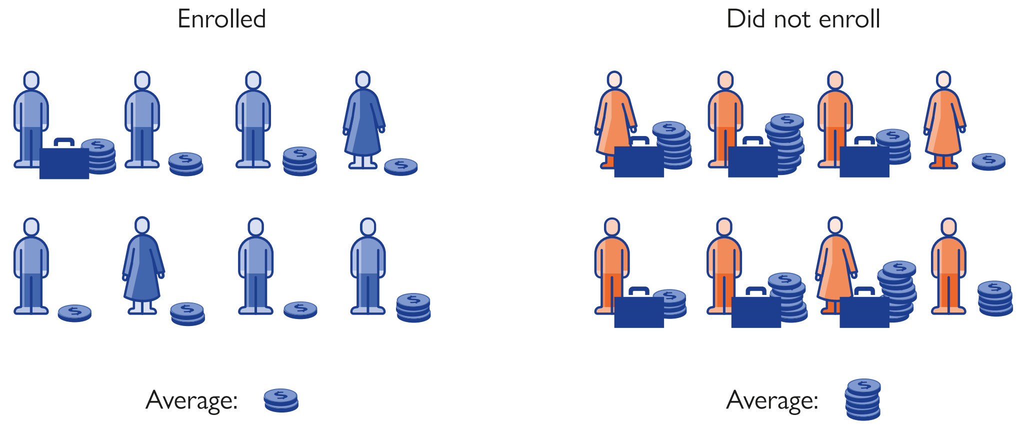
DESIGNS

EXPERIMENTAL DESIGNS: RANDOMIZED CONTROLLED TRIALS (RCTs)

A pure experimental design requires truly random selection of a group to receive the treatment. This randomization avoids the issue of selection bias.

Selection bias happens when there are characteristics that can make units more likely to receive the treatment, which can also affect the results of the treatment. For example, individuals who are unemployed would be more likely to enrol in an employment support or training scheme, as illustrated in Figure 22.

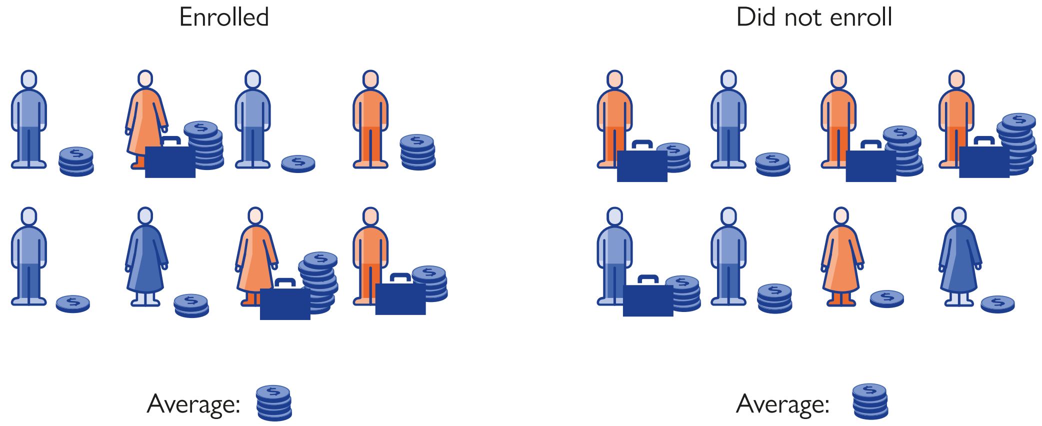
FIGURE 22: NON-RANDOM SELECTION CAN LEAD TO SELECTION BIAS



Evaluating the programme by, say, comparing the income of participants to those who did not join the scheme would likely result in skewed results, because the non-participants would likely have, on average, started out with better employment and income situations.

When using random assignment to create control and treatment groups (Figure 23), this randomness ensures that there is no link between an individual's likelihood of being assigned to receive the treatment and the outcome of the treatment, as none of the characteristics (observed or unobserved) that would make a difference to the effect had any influence on whether the individual was selected.

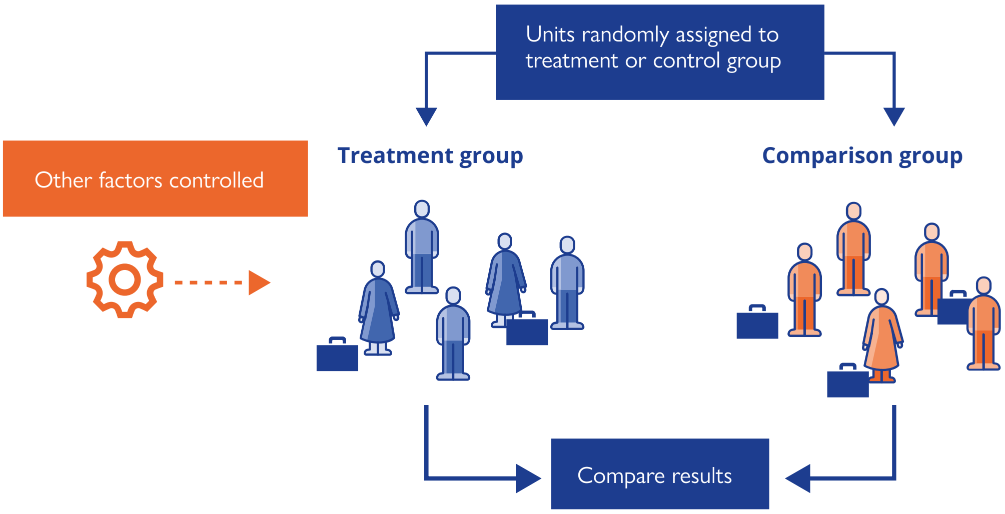
FIGURE 23: RANDOM SELECTION AVOIDS SELECTION BIAS



Therefore, when we calculate the effect of the treatment, we can be confident that there were no other influences on that effect.

This approach is called a randomized controlled trial (RCT), and is illustrated in Figure 24.

FIGURE 24: RANDOMIZED CONTROLLED TRIAL



The word "controlled" in "randomized controlled trial" does not refer to a control group, but rather to controlling factors other than the intervention that may affect the indicators being measured, to prevent contamination.

An RCT does not require the comparison group to not receive any intervention, but can instead use an alternative intervention to make a comparison against. The key thing is that units are assigned the intervention or alternative randomly.

Note that it may sometimes be appropriate for randomization to take place at higher levels – such as by community, rather than at the level of individual people.

Advantages

- RCTs, when well designed and executed, are generally considered to be a very robust way of establishing attribution because the randomization of who receives the treatment or not (or who receives one type of treatment or another) prevents selection bias.
    - This means that the influence of external factors should be the same for both the treatment and control group. This makes the analysis simpler, as there is no need to try to isolate any influence from external factors. (Note, however, that contamination from spill-over is still possible.)

Disadvantages

- There are ethical concerns – even if an alternative intervention is provided, the returnees in the control group will be prevented from receiving the treatment, which may be better than the alternative being compared against.
- A traditional RCT requires advance knowledge of where and to whom the treatment will and will not be implemented. Programmes in the return and reintegration context can rarely define this and need to be adaptable to changes, such as a new migration location or additional intervention being added to the programme.
    - There are alternative adapted RCT designs that can be applied in this context, such as a stepped-wedge design, but these are not simple to implement.
- Conducting an RCT requires that the planning of the evaluation is initiated very early on in the programme intervention to design the randomization process and accommodate it before programme activities commence. This is often not the case.

ⓘ Some people argue that the results of an impact evaluation are only credible if an RCT design was used, whereas others argue that they are one of several options and are not always an appropriate method to address all types of questions.

While they can be very useful in providing strong evidence for attribution, they may not be appropriate or feasible, depending on various contextual factors and the aims of the evaluation. You can find out more about this on page 142.

QUASI-EXPERIMENTAL DESIGNS

It is not always possible to randomize allocation of treatments. In such a situation the next best option is to use what is called a "quasi-experimental" approach.

Why is it called "quasi-experimental"?

The prefix "quasi-" means to resemble something, or to be nearly but not entirely the same as something. Quasi-experimental designs resemble experiments in that they have many of the demands of a full experimental approach, such as management of contamination, but are different because they lack the random selection of the control group.

Random selection is the best way to prevent selection bias, but may not always be a feasible or sensible option. Quasi-experimental designs use other approaches to assign treatments to attempt to create a comparison group that is similar to the group receiving the intervention.

Advantages

- Often more practical to implement in development interventions – in fact, sometimes it is the only feasible option.
- May be more ethical to conduct, e.g. targeting programming on the poorest or most needy.

Disadvantages

- Lack of randomization increases risk of selection bias.
- Evidence produced by quasi-experiments do not offer the same level of robustness of a well-executed experimental study. For this reason, it may be seen to have less credibility.

MATCHING

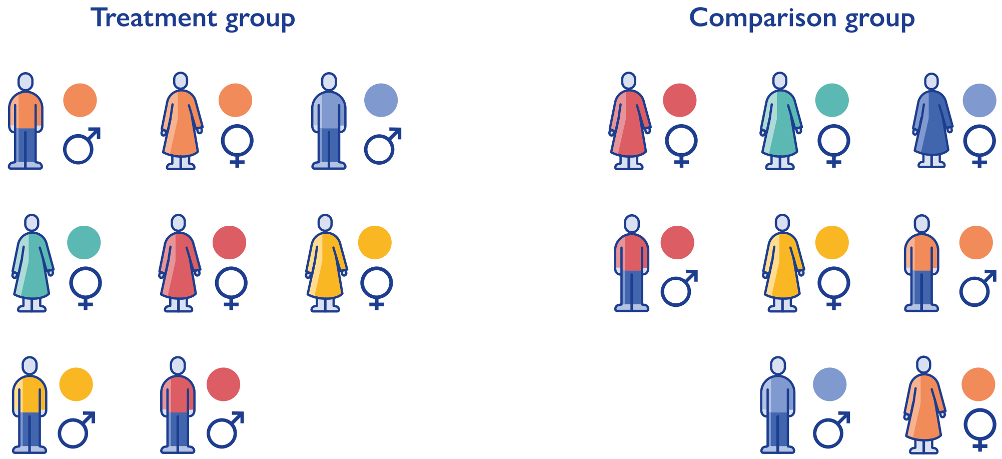
This is a strategy in which a comparison group is created by finding units that are similar to each unit in the treatment group, according to the characteristics that are understood to be relevant to the effect of the intervention. The purpose of matching is to try to avoid selection bias when it is not possible or sensible to use random selection to create a comparison group.

Matching can be incorporated into study designs – particularly quasi-experimental designs – to create a comparison group.

Return to the "Experimental methods" section on page 86 to review selection bias.

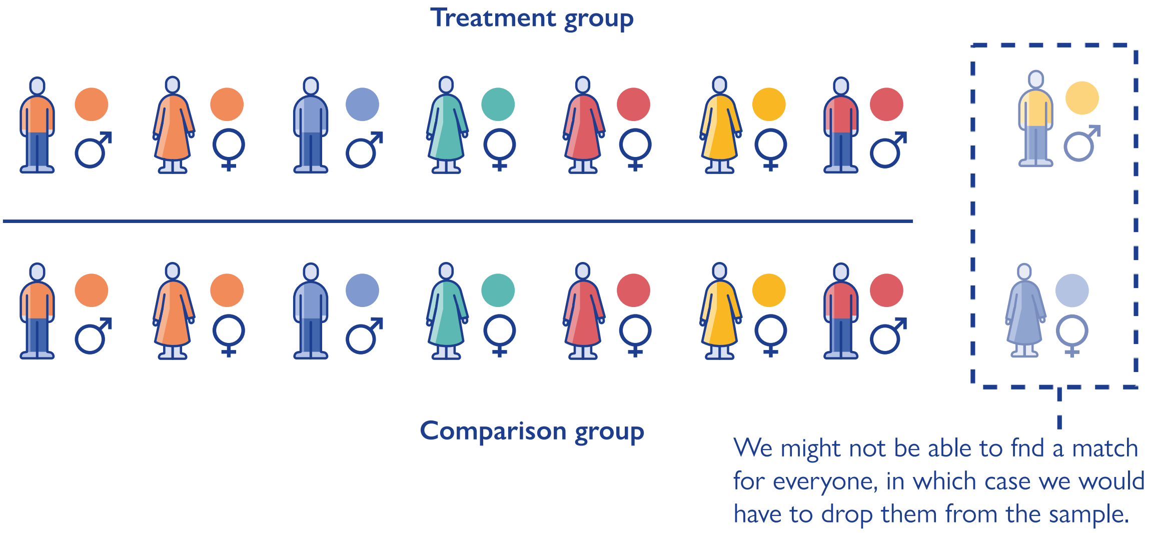
To use matching to create a comparison group, we start by taking samples to create a group to receive the intervention and a potential comparison group. We identify the relevant characteristics of the units in the group receiving the intervention (Figure 25).

FIGURE 25: MATCHING: IDENTIFYING RELEVANT CHARACTERISTICS



Then we look at our comparison group sample and try to find matches for each unit in our intervention group (Figure 26). In this way, we construct a comparison group that is very similar to the group receiving the intervention, at least according to the characteristics that we know are relevant and which we can identify.

FIGURE 26: MATCHING: CONSTRUCTING A COMPARISON GROUP


Treatment group
Comparison group
We might not be able to fnd a match for everyone, in which case we would have to drop them from the sample.

Advantages

- This strategy can be an effective way of reducing selection bias.
- It can be a useful option in scenarios where there is limited control over who receives what interventions – something that is fairly common in humanitarian and development scenarios.

Disadvantages

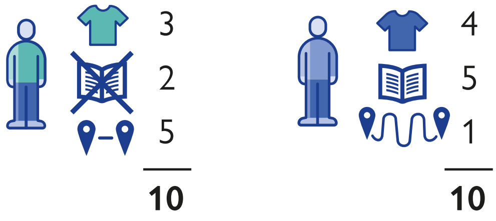
- Matching is demanding in terms of resources in that it requires data for each unit on the characteristics used for the matching process.
- The process of seeking out and identifying matches for the comparison group can be challenging; methods such as propensity score matching are used to make it more practicable.
- Where matches cannot be found, there may be some data that have been collected that cannot be used for the evaluation, because there is nothing to compare them against.
- It can be an expensive option, as an inflated sample size is needed to allow for the loss of units that cannot be matched.

PROPENSITY SCORE MATCHING

Propensity score matching (PSM) is a type of matching approach.

Matching attempts to avoid selection bias by creating a comparison group that has roughly the same characteristics as the treatment group. It does this by constructing the comparison group from units that "match" the units in the treatment group.

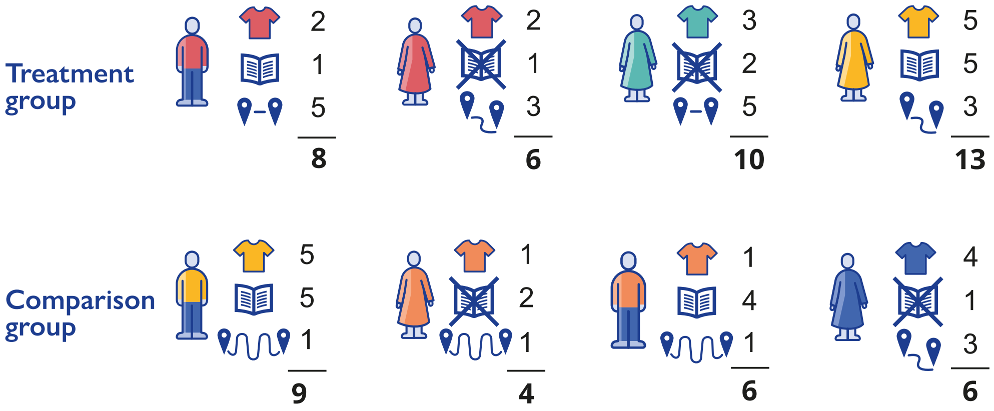
Propensity score matching focuses on characteristics that make a unit more likely to receive the treatment and combines them to create a "propensity score" for each unit, that summarizes how likely they are to receive the treatment. Units with similar scores are matched, even if their specific characteristics are different.



Remember that selection bias is caused when there are characteristics that affect a unit's probability of receiving the treatment and which also affect the outcome of the treatment. By creating a comparison group out of units that match the units in the treatment group in terms of their likelihood to receive the treatment, this method therefore aims to avoid this bias.

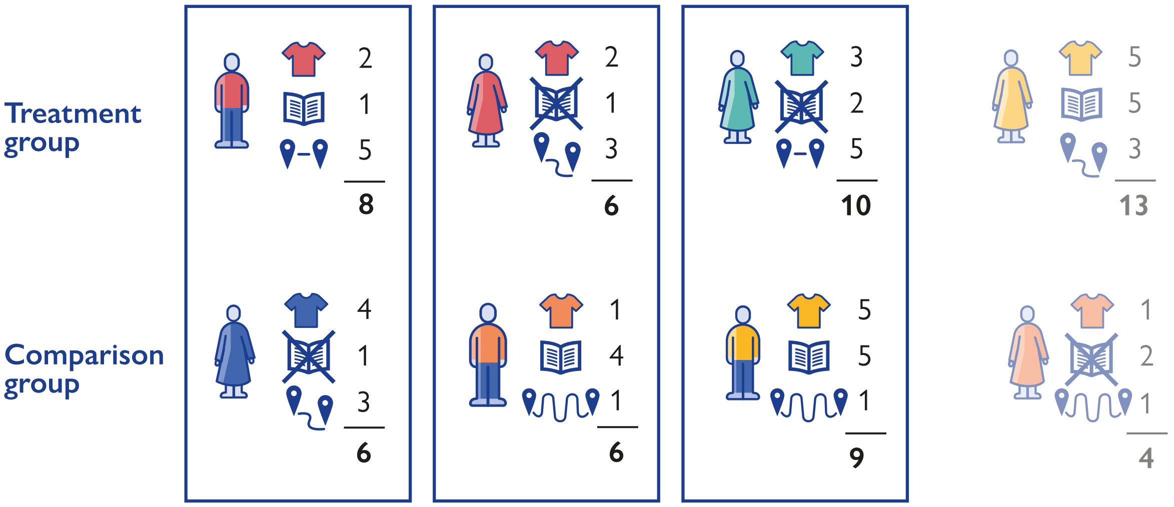
For example, perhaps a programme is evaluating the impact of employment support workshops for returnees. Each potential workshop attendee is given a "propensity score" based on their distance from the workshop venue, their literacy level and other relevant factors (Figure 27).

FIGURE 27: PROPENSITY SCORE MATCHING: CREATING PROPENSITY SCORES



For each unit that received the treatment – in this example, each workshop attendee – a unit who didn't receive the treatment with a similar propensity score is sought (Figure 28). Note that these matched units might have quite different characteristics; a returnee with good literacy who lives far from the venue could have the same propensity score as a returnee with limited literacy who is located close by, and these could be matched.

FIGURE 28: PROPENSITY SCORE MATCHING: MATCHING BASED ON PROPENSITY SCORE



This creates a comparison group that is matched to the group that received the treatment, according to their propensity scores. The logic behind this is that the two groups are roughly the same in terms of their likelihood to receive the treatment, even if they have different characteristics. This is deemed to make the groups similar enough to avoid selection bias.

Advantages

- PSM may be a good option to get some of the benefits of matching with less difficulty. As the matching is based on just the propensity score, it can be easier to find matches.

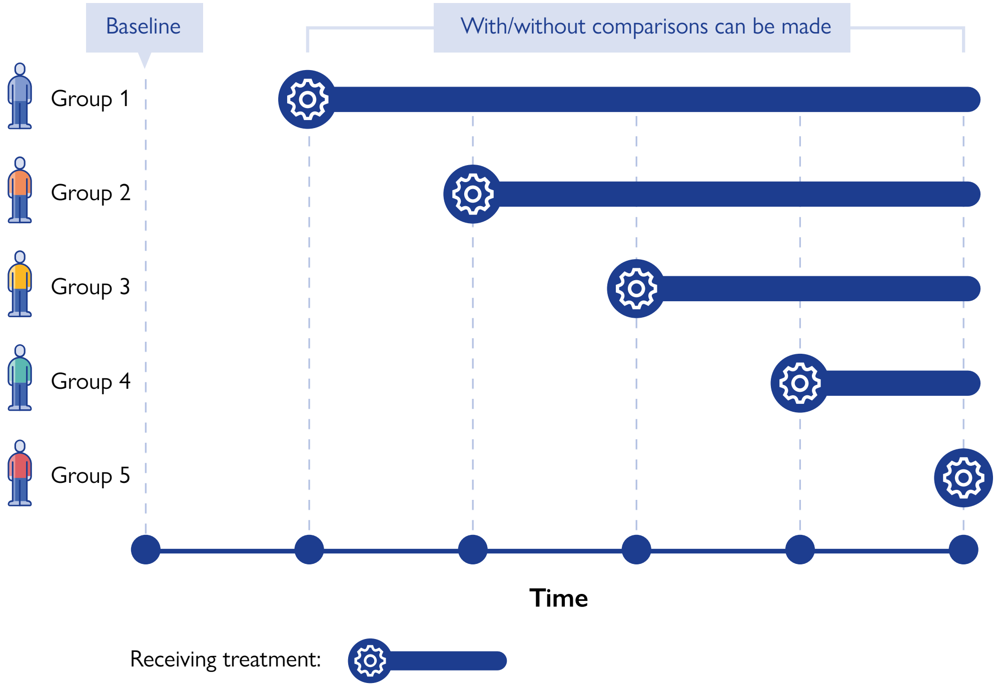
Disadvantages

- Deciding on the characteristics to use as the basis for calculating the propensity score needs careful consideration. If a characteristic is missed that significantly affects participation and also affects the outcome of the treatments, the matching will be incomplete, and the results of the comparison is at risk of bias.

STEPPED-WEDGE DESIGN

One option for creating a comparison group is to use a "stepped-wedge" design (Figure 29). This is where, rather than having a group that does not receive the treatment, the treatment is rolled out to groups of participants in a staggered manner over time:

FIGURE 29: A STEPPED-WEDGE DESIGN



By collecting baseline, endline and carefully timed midlines, a stepped-wedge design can be used to provide both with/without and before/after comparisons.

If the participants are assigned to groups randomly, this design can create a control group for a randomized controlled trial. It can also be used as part of a quasi-experimental design where randomization is not possible.

If the assignment of units to groups is not randomized – for example, it might be important to ensure those with the greatest need receive the treatment soonest – then the risk of selection bias should be considered and included in interpretations of the results.

Advantages

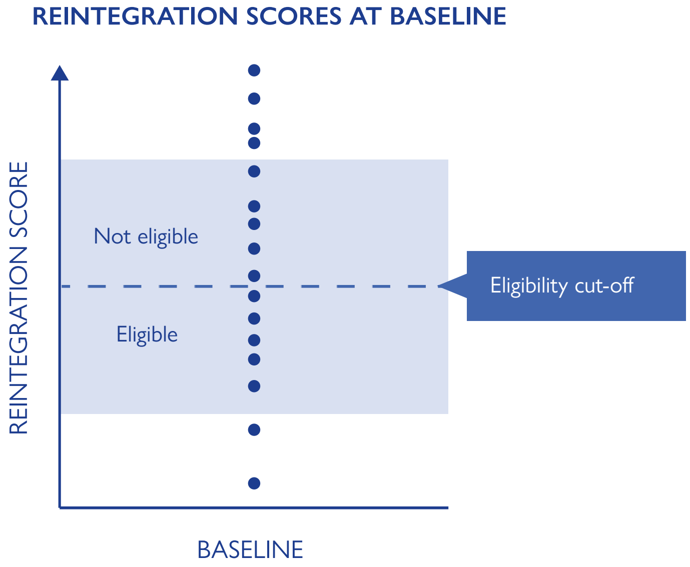
- The stepped-wedge can be a good choice where there are practical constraints – for example, if it is not possible to provide a treatment to all beneficiaries at once due to budgetary or staffing reasons.
- This is a possible solution to the ethical challenges of creating some kinds of comparison groups, as the treatment will eventually be given to all study participants.

Disadvantages

- It is necessary to conduct midline data collection for a stepped-wedge approach; potentially, several rounds are needed (in the example above you would need to collect data at each of the five steps). This can make it an expensive or difficult option.
- A study using a stepped-wedge approach takes longer to carry out than many other designs as the data collection must also be continued for as long as it takes for the last beneficiary group to show the effect of the treatment. This might prevent it from being a feasible option in some cases. It would be best suited for studies where there is a short amount of time between treatment and outcomes.

REGRESSION DISCONTINUITY DESIGN

Instead of random assignment, regression discontinuity designs use some system of ranking or eligibility based on a continuous variable, such as income, (called a "running variable") and assign a "cut-off" point, beyond which units are not eligible to receive the intervention.

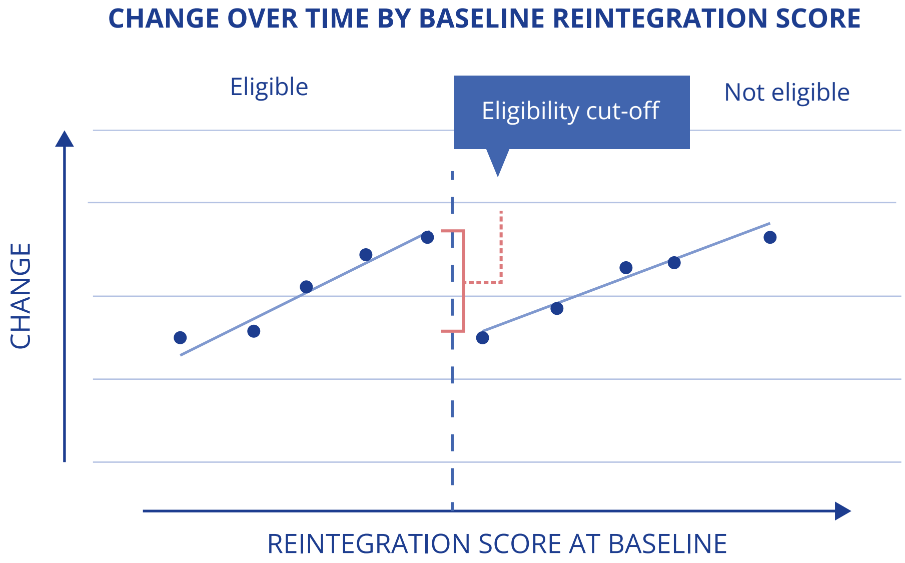


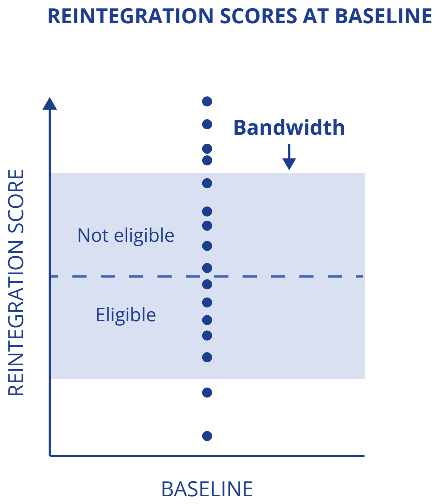
Let's say we have a programme that has conducted a baseline survey to measure a reintegration index score for returnees. The programme offers remedial support to returnees considered most in need. It determines a cut-off point for the reintegration score.

Returnees below the cut-off score are eligible for the remedial support.

Returnees who were close to the cut-off at either side are very similar and can be considered similar enough to be used as the treatment group and comparison group for a with/without comparison.

A baseline-endline comparison is used to measure the change for both groups. The difference in change over time can be calculated to determine the impact of the remedial support.





Bandwidth

There is a careful decision-making process involved in deciding how close to the cut-off point units need to be in order to be included in the comparison – this range is sometimes called the "bandwidth".

A compromise needs to be found between including a sufficient sample size and reducing the effectiveness of the approach in reducing selection bias.

Data are collected from a much larger sample than is likely to be included in the evaluation, to allow for the loss of units that are too far from the cut-off point.

Advantages

- Any non-observable characteristics between the two groups that could affect the outcome are dealt with under the premise that if returnees have very similar scores at the outset then the non-observable characteristics affecting reintegration performance have been accounted for.

Disadvantages

- Single-criteria eligibility cut-offs are not available for many programmes for which impact evaluations are conducted.
- If this prebaseline data are not readily available from secondary sources, then the evaluation has to account for the extra expense of collecting these preliminary data.
- The number of units within the range above and below the threshold may not be enough for a statistically representative sample.
- The observed treated just below the cut-off threshold may not be truly representative of the larger targeted population that are further below the threshold.

NATURAL EXPERIMENTS AND EMERGENT DOMAINS

A natural experiment is a scenario where a treatment and comparison group occur "naturally" – by chance and circumstance, rather than having been designed from the outset of the impact evaluation.

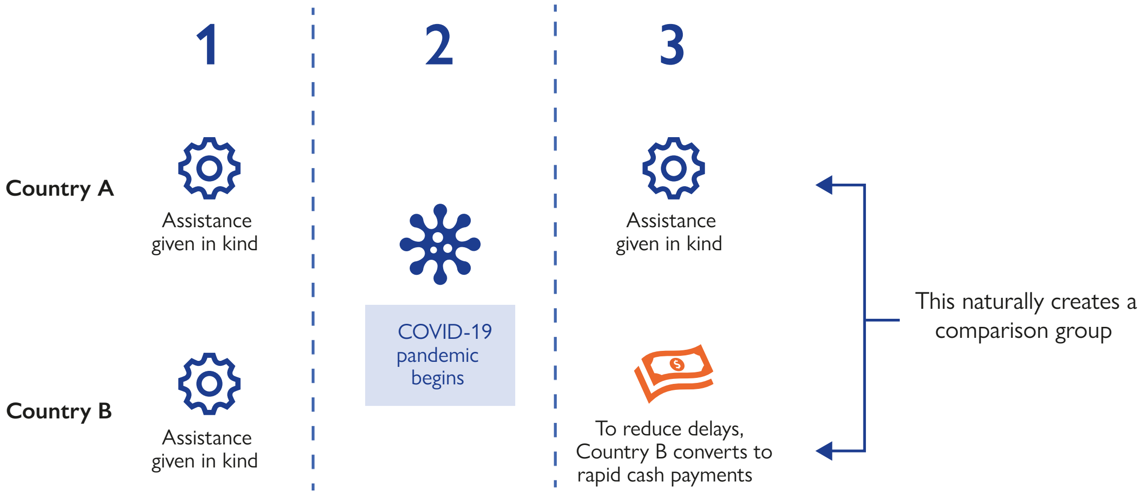
Sometimes, these scenarios can be referred to as "emergent domains", indicating that they are not anticipated at the outset, but emerged during the impact evaluation period.

IOM COVID-19 natural experiment in the Horn of Africa

The COVID-19 pandemic provided the EU-IOM Joint Initiative programme of IOM an opportunity to create a natural experiment study. This natural experiment will assess how returnees and their families were affected by the COVID-19-linked shock (such as lockdown, effect on labour market and mortality shock) which varied across Ethiopia, Somalia and the Sudan in its severity, duration, timing and how each country responded. These differences, which occurred "naturally", create groups of returnees that can be used like comparison groups.

Another component of the natural experiment was the changes to IOM's assistance to returnees, which again varied across countries. The assistance in these countries that was previously given in kind was either partially or completely converted to rapid cash payments. The differences between these adjustments in each country allows for comparisons to measure the impact of the differing approaches. This example is illustrated below in Figure 30.

FIGURE 30: AN EXAMPLE OF A NATURAL EXPERIMENT: CHANGES TO IOM'S ASSISTANCE TO RETURNEES



Natural experiments come with risk of bias and should be considered carefully. However, they are practical in that they can be cheaper and may allow for comparisons that would be otherwise impractical or unethical to conduct.

KNOWLEDGE CHECK

Question 1

Match the description with the quantitative method.

Quantitative method	Description
Randomized controlled trials	A technique for constructing a comparison group that is similar to the treatment group.
Quasi-experiments	The most robust experimental approach, using fully randomized assignment of treatment.
Matching	Creates comparison groups by delaying provision of the treatment, not witholding it.
Stepped-wedge	Methods that use various strategies to create a comparison group without randomization.

Question 2

Match the quantitative method with the relevant prerequisite.

Quantitative method	Prerequisite
Regression discontinuity design	Requires circumstances that create treatment and comparison groups without being designed.
Propensity score matching	Treatment is assigned based on an eligibility criteria with a cut-off point.
Natural experiment	Requires prior knowledge of characteristics that affect likelihood of receiving treatment.

KNOWLEDGE CHECK ANSWERS

Question 1 answer

Match the description with the quantitative method.

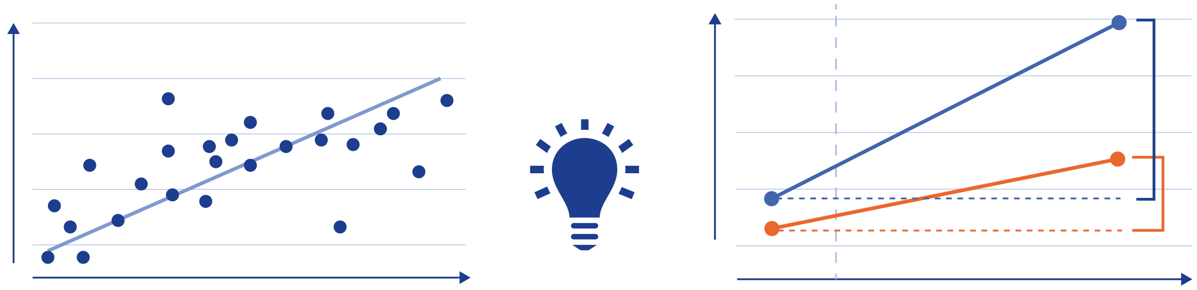
Quantitative method	Description
Randomized controlled trials	The most robust experimental approach, using fully randomized assignment of treatment.
Quasi-experiments	Methods that use various strategies to create a comparison group without randomization.
Matching	A technique for constructing a comparison group that is similar to the treatment group.
Stepped-wedge	Creates comparison groups by delaying provision of the treatment, not witholding it.

Question 2 answer

Match the quantitative method with the relevant prerequisite.

Quantitative method	Prerequisite
Regression discontinuity design	Treatment is assigned based on an eligibility criteria with a cut-off point.
Propensity score matching	Requires prior knowledge of characteristics that affect likelihood of receiving treatment.
Natural experiment	Requires circumstances that create treatment and comparison groups without being designed.

QUANTITATIVE ANALYSIS



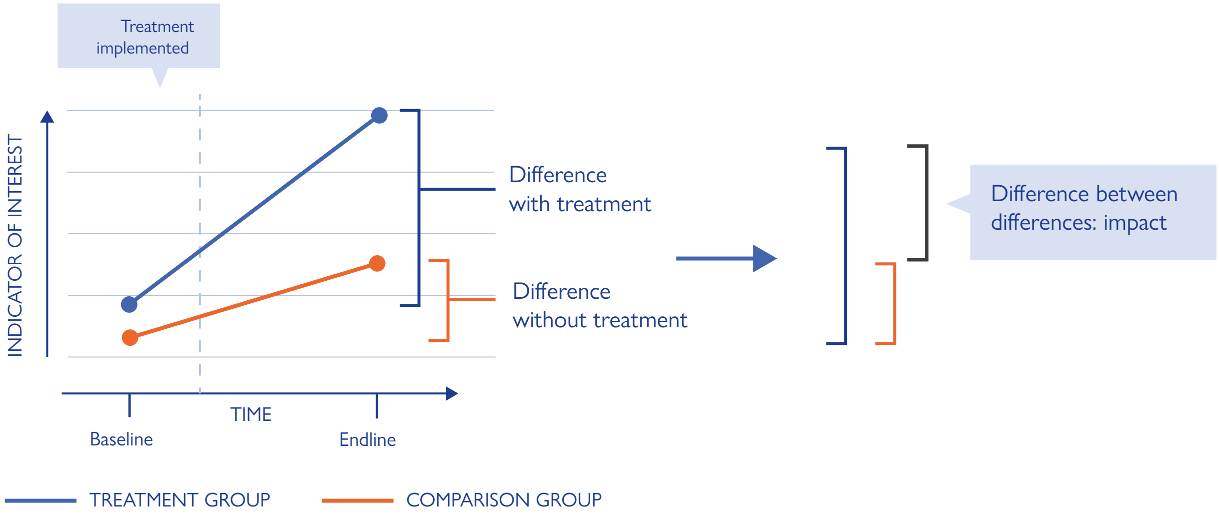
This section contains options for the methods of interpreting data from quantitative impact evaluations. Although data analysis should be conducted by a qualified expert, it is good to consider the analysis at the design stage, to ensure that the appropriate data are collected.

DIFFERENCE IN DIFFERENCE

This is essentially the approach of combining the before/after and with/without dimensions of comparison. This methodology is very commonly used and is particularly helpful for designs that do not or cannot include random allocation between the treated group and the comparison group.

The averages of indicators of interest are taken from the treatment group and from the comparison group before and after the treatment has taken place. Comparing the change over time for the treatment compared with the comparison group provides an estimate for the effect of the treatment. So, as we see in Figure 31, difference in difference means measuring the difference between "before" and "after" for each group and then looking at the difference between those differences.

FIGURE 31: DIFFERENCE IN DIFFERENCE



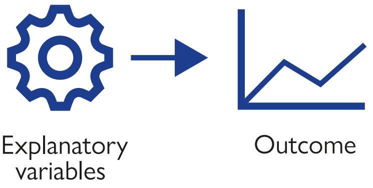
The difference in difference approach is central to quantitative analysis of designed impact evaluations, whether this involves experimental or quasi-experimental methods.

Note for advanced trainees:

Regression is an important method by which data are often analysed in quantitative impact evaluations, and it can be very useful to have an understanding of this when planning for the analysis of data in an impact evaluation. However, as a statistical technique, it is a topic that requires familiarity with concepts that are beyond the scope of this course.

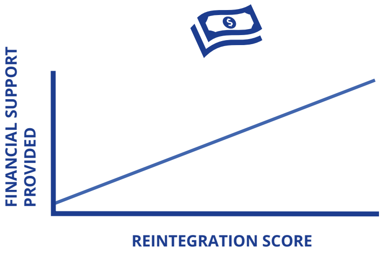
The following section is therefore included to be of use to a subset of trainees and assumes an existing knowledge of basic concepts of statistical modelling.

Regression analysis is central to the analysis of impact evaluation. In principle, a regression model helps us to assess, quantify and generate a mathematical representation of the relationship between a dependent variable (outcome) and a series of explanatory variables.

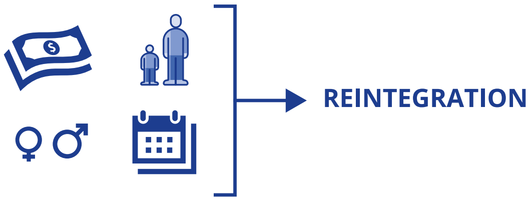


For example, regression can be used to explore the extent to which:

- The reintegration index score (outcome) is affected by the amount of financial support (explanatory variable) provided by a programme. This is a case of one dependent variable and its relationship with one explanatory variable.

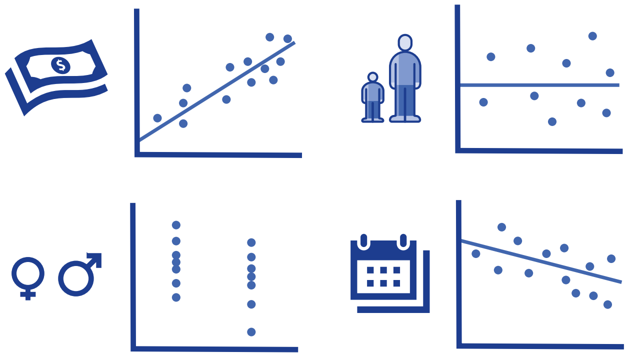


- The reintegration index score (outcome) is affected by different types of support available to the returnee and personal characteristics of the returnee, such as gender, age, length of time since returning to the country. This is a case of one dependent variable and its relationship with a set of explanatory variables.



In impact evaluation we use regression analyses for three common purposes:

Build a model that explains the dependent variable.



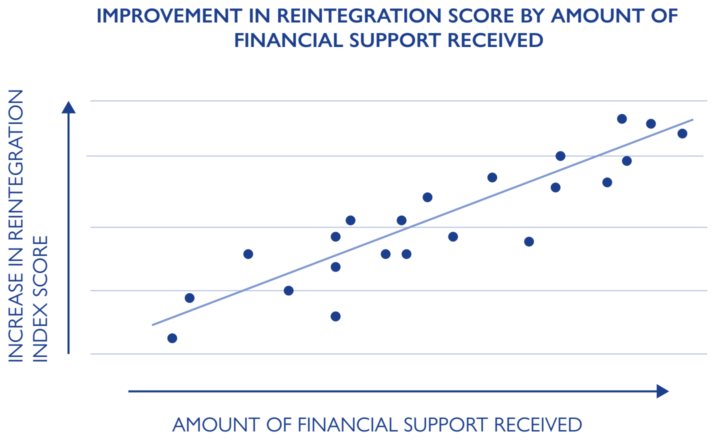
In the second example above, we could use regression analysis to assess whether the effects of gender, age and length of time are important to understand the reintegration index score. The result of the analysis may indicate that age is not a factor that affects the score and we would not include it in our “model”.

Similarly, a regression analysis may indicate that there are significantly different effects on the score depending on which combination of support services a returnee receives. In both cases, regression allows us to assess the relative contribution of a potential explanatory variable to our understanding of the variability of the RSI.

The result is both a mathematical model and the basis of a conceptual model that would help decisions.

Quantify the contribution of explanatory variables that we select into our model.

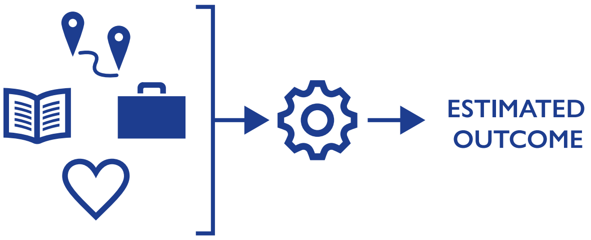
FIGURE 32: EXAMPLE OF A REGRESSION MODEL SHOWING THE EFFECT OF AN EXPLANATORY VARIABLE



This is what we do when we use regression models for the difference in difference approach. The mathematical model that emerges from a regression analysis estimates the size of the "difference in difference" effect (Figure 32). Even better, it can provide the estimate taking into account other factors (not just time and intervention/no intervention) but important characteristics of the returnee that are included into the model (as described above).

Use models to calculate estimated outcomes.

This is when regression models are used to estimate of an outcome. For example, in propensity score matching, we use logistic regression to calculate the propensity scores that eventually allow matching of returnees and non-returnees.



It is worth pointing out the term regression is used for a large family of models that can be used to explore the relationship between outcomes and their determinants. These models can be:

- Relatively simple: one quantitative outcome and one quantitative determinant.
- Complex: an outcome with multiple ordered categorical values versus a set of determinants, some of which can be measured at individual level and some that can only be measured at group level, some of which can be measured without error and some for which we know are subject to different types of error.

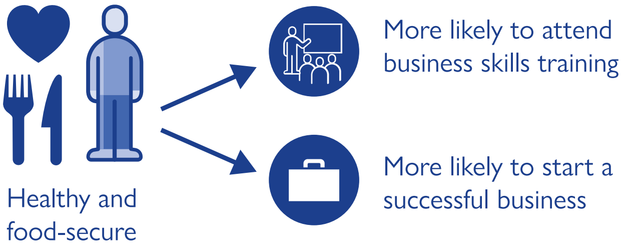
Many people can deal with the first type of regression analysis, and in most cases, we need a competent analyst to deal with the most complex models.

Above all, it is important to know that regression models are statistical models that we use to understand relationships between outcomes and explanatory variables. If a model does not help, or even worse, if it reduces our ability to understand, then it is not fulfilling its purpose and we need to rethink our approach.

INSTRUMENTAL VARIABLES

The use of instrumental variables is a way to counteract selection bias when estimating the impact of an intervention.

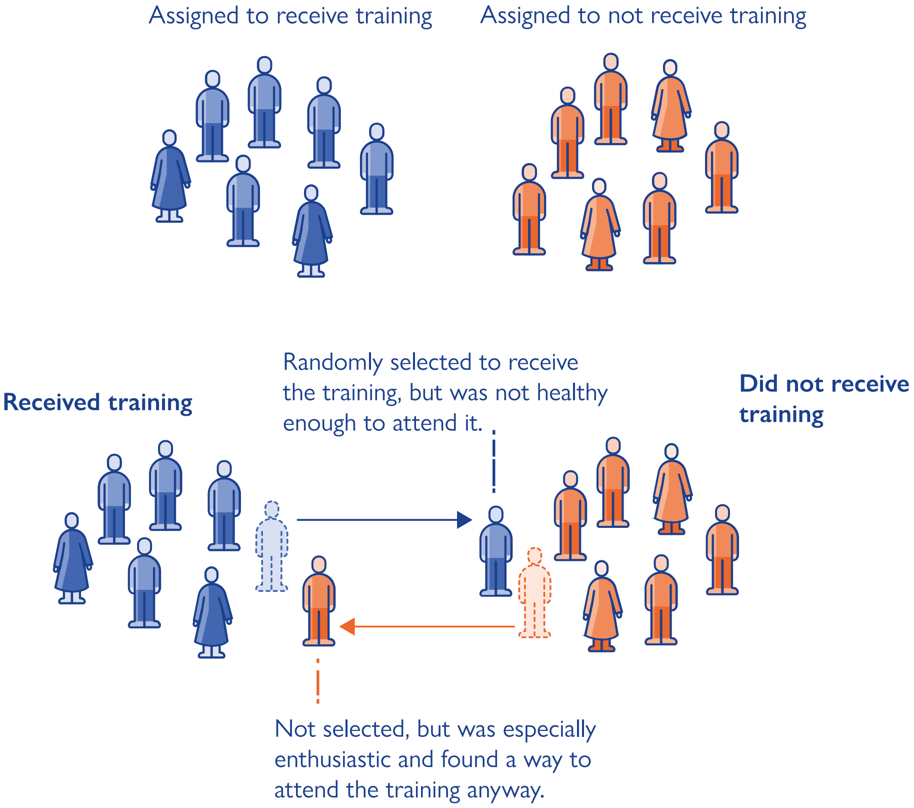
Selection bias is when there are characteristics of units – such as returnees – that make them more likely to receive the treatment (the intervention being evaluated), which also have an effect on the outcome of the treatment. For example, a returnee who is healthy and food-secure is more likely to attend business skills training and also more likely to be able to put the training into practice and start a successful business.


Healthy and food-secure
More likely to attend business skills training
More likely to start a successful business

When using random assignment to create control and treatment groups, this randomness ensures that there is no link between an individual's likelihood of being assigned to receive the treatment and the outcome of the treatment, as none of the characteristics that would make a difference to the effect had any influence on whether the individual was selected. Therefore, when we calculate the effect of the treatment, we can be confident that there were no other influences on that effect.

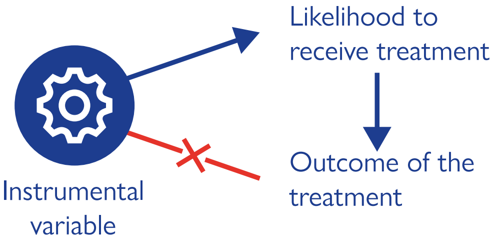
This is great in theory. However, in practice this often does not work as planned. Let's say we assign eligible returnees at random to receive the business skills training. However, there are still characteristics that can affect whether the selected returnees actually receive the training (see Figure 33), and so selection bias has been introduced back into the experiment again.

FIGURE 33: INSTRUMENTAL VARIABLES: SELECTION BIAS CAN OCCUR EVEN WITH RANDOM SELECTION


Assigned to receive training
Assigned to not receive training
Received training
Randomly selected to receive the training, but was not healthy enough to attend it.
Did not receive training
Not selected, but was especially enthusiastic and found a way to attend the training anyway.

We use instrumental variables to resolve this. An instrumental variable is a variable that is strongly correlated with an individual's likelihood to receive the treatment, but has no correlation with the outcome of the treatment other than through whether they receive the treatment (see Figure 34).

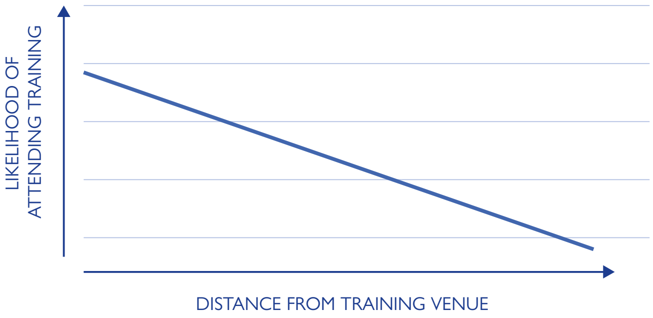
FIGURE 34: CHARACTERISTICS OF AN INSTRUMENTAL VARIABLE



Let's say we use distance from training venue for small business skills training as an instrumental variable. We have seen that returnees who can more conveniently attend training are more likely to attend the training if they are offered and vice versa. However, there is no reason to believe that distance from the training centre would otherwise affect their ability to start a small business.

The use of an instrumental variable works in two stages:

FIGURE 35: EXAMPLE OF THE EFFECT OF AN INSTRUMENTAL VARIABLE ON THE LIKELIHOOD OF RECEIVING TREATMENT



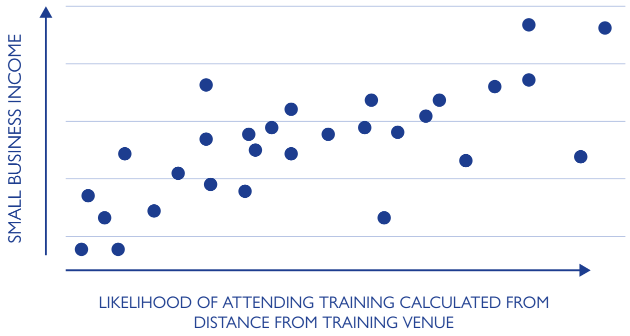
Firstly, we calculate the effect of the instrumental variable on the likelihood of receiving the treatment.

In our example, this means we would look at the difference that distance from the training venue makes to how likely an individual is to attend the training (Figure 35).

Now we can say how likely an individual is to receive the treatment according only to the instrumental variable – so in this case, how likely each person is to attend the training, based on their distance from the training venue and no other factors.

Next, we look at the link between this likelihood we have just calculated and the outcomes of the treatment. In our example, shown in Figure 36, this is the link between the likelihood of receiving the training – based only on distance from the centre – and success in starting a small business.

FIGURE 36: EXAMPLE OF THE RELATIONSHIP BETWEEN AN INSTRUMENTAL VARIABLE AND THE OUTCOME OF THE TREATMENT



As we have already established that there is no direct relationship

between these two things, any correlation between them must be the effect of the treatment – the business skills training.

Therefore, we have an unbiased estimate of the effect of the treatment on the outcome.

Things to consider

- This method only works if the instrumental variables are well selected. They must be a characteristic that is highly correlated with receiving the treatment and which can have no effect on the outcome except through the treatment.
- It can be very difficult to find an instrumental variable that meets this criterion. It will not always be possible.

QUIZ

This quiz will check your understanding of the topics covered in this module. There are seven questions. You must get a score of at least five out of seven to pass.

1. Which of the following are examples of contamination? Select all the answers that apply.

□ When the comparison group is affected by the intervention intended for the treatment group.

□ When the comparison group has a more positive outcome than the group receiving the intervention.

□ When errors are made in data collection.

□ When the study population receives assistance from another programme.

2. Match the quantitative methods for impact evaluations to the disadvantages.

Regression discontinuity design	This design must be decided and planned before programme activities commence.
Matching	Eligibility criteria that are based on single, continuous variables are uncommon.
Randomized controlled trials	Units may have to be dropped from the sample after it has been drawn.
Stepped-wedge design	Requires a longer period of evaluation and might need multiple midlines to be taken.

3. Which two of the following are criteria for instrumental variables? Select all the answers that apply.

□ Must be correlated to the outcome of the intervention.

□ Must be correlated with the likelihood to receive the intervention.

□ Must not be correlated with the likelihood to receive the intervention.

□ Must not be causally linked with the outcome of the intervention.

4. Match the statements to the quantitative method.

Regression based methods	Exploits unplanned events to use comparison groups that would not be practical or ethical.
Natural experiments	Combines the before/after and with/without dimensions of comparison.
Difference in difference	A way to counteract selection bias when estimating the impact of an intervention.
Instrumental variables	Tests which variables are related (or correlated) to the outcome and to what extent.

5. What is a unit of observation? Select one answer.

□ The unit at which analysis is conducted.

□ The level in the hierarchy at which the treatment has taken place.

□ A group which requires a separate estimate of impact.

□ The person or thing at the lowest level of the hierarchy upon which observations and measurements are made.

6. When might it be helpful to use stratification? Select all the answers that apply.

□ If you expect notably different results for certain subpopulations.

□ When there is a need to include certain subpopulations in the study that would otherwise be unlikely to be sampled.

□ When you need to make observations at different levels of the hierarchy.

□ When you want to measure multiple indicators.

7. Which of the following may be important when making decisions about what to measure in an impact evaluation for return and reintegration programming? Select all the answers that apply.

□ Perspectives of the returnees.

□ Evaluation aims.

□ The expected effect of the intervention.

□ Available data.

□ Intended users of the information.

QUIZ ANSWERS

There are seven questions. You must get a score of at least five out of seven to pass.

1. Which of the following are examples of contamination? Select all the answers that apply.

☑ When the comparison group is affected by the intervention intended for the treatment group.

□ When the comparison group has a more positive outcome than the group receiving the intervention.

□ When errors are made in data collection.

☑ When the study population receives assistance from another programme.

2. Match the quantitative methods for impact evaluations to the disadvantages.

Regression discontinuity design	→	This design must be decided and planned before programme activities commence.
Matching	→	Eligibility criteria that are based on single, continuous variables are uncommon.
Randomized controlled trials	→	Units may have to be dropped from the sample after it has been drawn.
Stepped-wedge design	→	Requires a longer period of evaluation and might need multiple midlines to be taken.

3. Which two of the following are criteria for instrumental variables? Select all the answers that apply.

□ Must be correlated to the outcome of the intervention.

☑ Must be correlated with the likelihood to receive the intervention.

□ Must not be correlated with the likelihood to receive the intervention.

☑ Must not be causally linked with the outcome of the intervention.

4. Match the statements to the quantitative method.

Regression based methods	→	Tests which variables are related (or correlated) to the outcome and to what extent.
Natural experiments	→	Exploits unplanned events to use comparison groups that would not be practical or ethical.
Difference in difference	→	Combines the before/after and with/without dimensions of comparison.
Instrumental variables	→	A way to counteract selection bias when estimating the impact of an intervention.

5. What is a unit of observation? Select one answer.

- □ The unit at which analysis is conducted.
- □ The level in the hierarchy at which the treatment has taken place.
- □ A group which requires a separate estimate of impact.
- ☑ The person or thing at the lowest level of the hierarchy upon which observations and measurements are made.

6. When might it be helpful to use stratification? Select all the answers that apply.

- ☑ If you expect notably different results for certain subpopulations.
- ☑ When there is a need to include certain subpopulations in the study that would otherwise be unlikely to be sampled.
- □ When you need to make observations at different levels of the hierarchy.
- □ When you want to measure multiple indicators.

7. Which of the following may be important when making decisions about what to measure in an impact evaluation for return and reintegration programming? Select all the answers that apply.

- ☑ Perspectives of the returnees.
- ☑ Evaluation aims.
- ☑ The expected effect of the intervention.
- ☑ Available data.
- ☑ Intended users of the information.

SUMMARY

In this module, we have seen that:

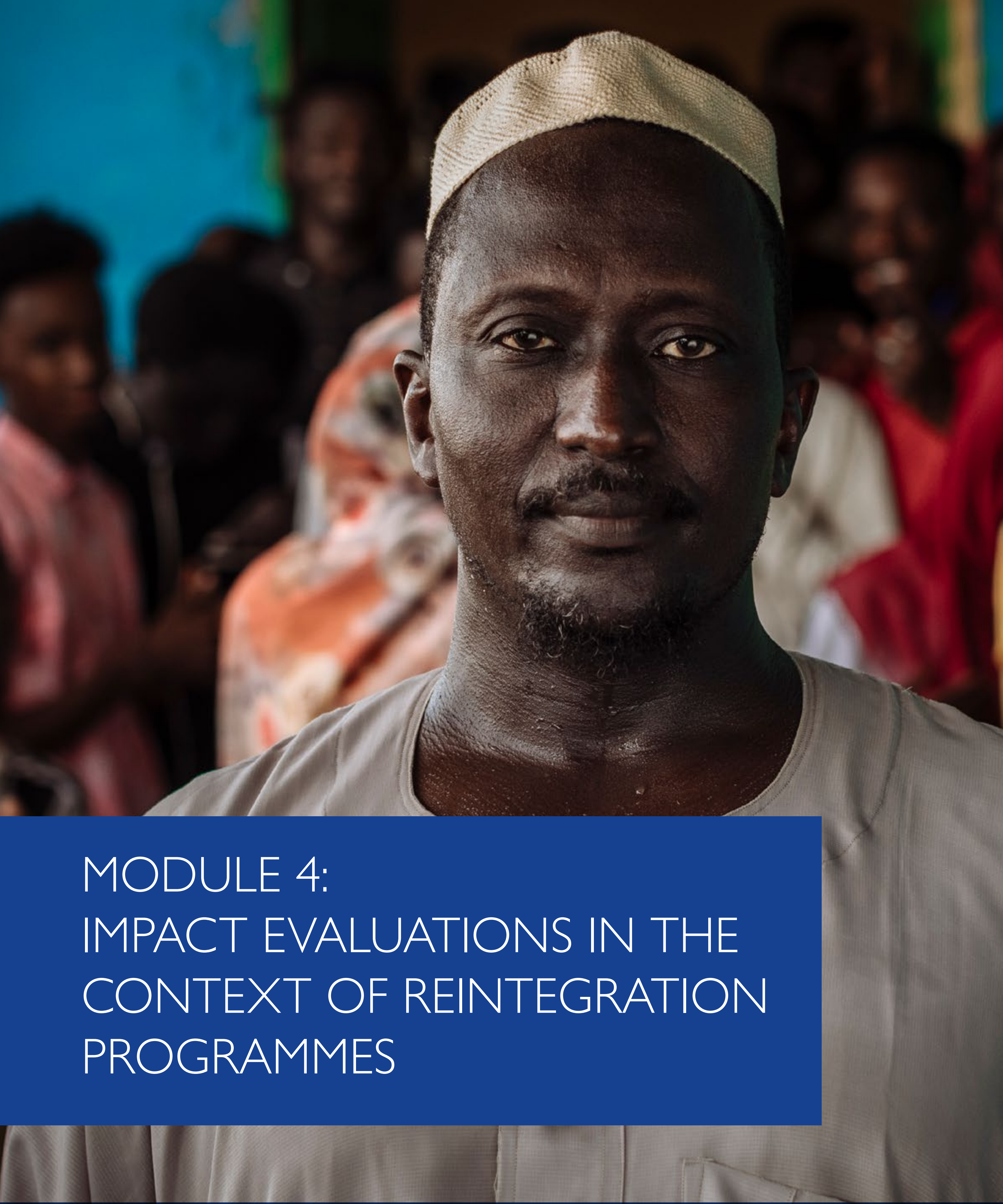
1. Designing a quantitative impact evaluation study involves decisions about what, who and how to measure.

2. Making decisions about the design aspects involves a process of balancing priorities and making compromises between programme aims and what is possible to implement.

3. While a randomized controlled trial would provide the most rigorous evidence for attributable impact, there many be situations where other strategies are more suitable.

4. There is a range of strategies available to make comparisons while avoiding contamination and bias.

MODULE 4:
IMPACT EVALUATIONS IN THE CONTEXT OF REINTEGRATION PROGRAMMES



Ibrahim is the community leader and a teacher at Dar-Elsalam where the EU-IOM Joint Initiative rehabilitated a multipurpose centre. "There were no sports or income-generating activities for the vulnerable families at the centre, which was just an abandoned place. We want our centre to be productive for vulnerable families and youth; we want the centre to provide jobs for them because the country's economic situation has affected them, and we want our centre to provide a variety of activities and different languages." © IOM 2021/Muse MOHAMMED

MODULE 4: IMPACT EVALUATIONS IN THE CONTEXT OF REINTEGRATION PROGRAMMES

INTRODUCTION

This module will elaborate on the challenges, opportunities and possibilities particular to conducting impact evaluations in the context of reintegration programmes, how to measure their impacts, present expert interviews and real-life case studies.

OUTCOMES

At the end of this module, trainees will be able to:

- Summarize some of the challenges that the return and reintegration context creates for conducting impact evaluations.
- Outline examples of existing ways of measuring reintegration and how these can be applied to impact evaluations of return and reintegration programmes.
- Discuss the considerations needed regarding the collection of data for impact evaluations of return and reintegration programming.
- Reflect on the strengths and weaknesses of strategies that are used in impact evaluations of return and reintegration programmes to try to measure impacts and establish attribution/contribution.

Modules 2 and 3 focused on the basic concepts and methods for how impact evaluations are designed and carried out. Having gained familiarity with these principles, the next step is to take into consideration the context and how it intersects with those principles.

This module discusses the aspects that are specific to conducting impact evaluations for return and reintegration programmes and the measurement of impact for these. It provides case studies of how various groups have approached this challenge in the past.

WHAT IS DIFFERENT WHEN DOING AN IMPACT EVALUATION FOR REINTEGRATION PROGRAMMES?

One of the major considerations when planning an impact evaluation for return and reintegration programmes is the situation of the returnees themselves. The circumstances of returning migrants are varied, complex and sensitive in a way rarely seen in other contexts. These are some negative factors that sometimes affect returnees:

- Financial difficulties
- Debt
- Stress
- Trauma (e.g. rejected asylum-seekers, victims of human trafficking or sexual abuse)
- Reliant on humanitarian aid
- Reliant on support from friends and family
- Difficult emotions concerning their migration and return
- Stigma from the communities to which they are returning.

They are likely to be in an unstable position, both geographically and psychosocially. Recent returnees may move around quite frequently (due to social and financial concerns, availability of support and other factors) and are more susceptible to large and sudden changes in social and psychological well-being.

This reintegration context affects many aspects of an impact evaluation design, including:

How we define and measure the effect of the programme

Return and reintegration is a complex thing to define, and measuring it involves combining multiple variables. There is not a unified, agreed-upon approach to this.

Data collection and availability

- At what time(s) should data be collected?
- Practical difficulties that can lead to problems with implementing certain designs (such as those that require a baseline or longitudinal data collection) or may cause biased results.

Establishing a counterfactual or comparison cohorts

There are significant ethical and practical difficulties in creating a control group (e.g. returnees not assisted by the programme). Even identifying comparison groups may be challenging.

This module goes into detail about each of these aspects.

DEFINING AND MEASURING REINTEGRATION

Impact evaluations measure success based on the stated goals of a programme. In a reintegration programme, then, the likely indicator of interest you will want to measure will be "reintegration".

Besides this, you may also be asked to evaluate the "return" component of the process. This is how an evaluation can try to determine if the programme has been effective.

WHAT IS REINTEGRATION?

Take a moment to consider how you would define reintegration and make some notes.

Training of enumerators, South Sudan. © IOM 2018/Rikka TUPAZ

Suggested solution

Here are some definitions of reintegration. Did you come up with something similar?

> *"A process which enables individuals to re-establish the economic, social and psychosocial relationships needed to maintain life, livelihood and dignity and inclusion in civic life"*

IOM. *Glossary on Migration.* International Migration Law, No. 34. (IOM, 2019).
PDF: *https://publications.iom.int/system/files/pdf/iml_34_glossary.pdf#page=188*

> *"Individual has reintegrated into the economic, social and cultural processes of the country of origin and feels that they are in an environment of safety and security upon return"*

K. Koser and K. Kuschminder. *Comparative Research on the Assisted Voluntary Return and Reintegration of Migrants.* (IOM, 2015), p.8.
PDF: *www.iom.int/sites/g/files/tmzbdl486/files/migrated_files/What-We-Do/docs/AVRR-Research-final.pdf#page=8*

Your definition might have been very different but could still be entirely valid. What it means to be "reintegrated" is subjective; organizations, returnees and other individuals will likely have very varied ideas about what "reintegration" means.

MEASURING REINTEGRATION

The next thing to think about is, given these definitions, how can reintegration be measured?

Imagine you plan to conduct a study to measure how much the returning migrant population within a community has reintegrated.

Question

Based on the definitions from a moment ago, how would you measure reintegration? From the options below, which do you think you should measure to determine how successfully a returnee has reintegrated? Choose all the options that apply.

- □ Level of annual income or income diversity
- □ Levels of debt
- □ Food security
- □ Being employed
- □ Their perception of level of self-reliance
- □ Their feelings of well-being and social cohesion within the community
- □ Participation in community groups
- □ Trust in local leadership and decision-making and authorities
- □ Feelings of safety and security
- □ Expert or community leader opinion
- □ Have they remigrated, or are they considering it?

Answer

The answer could in fact be any or all of these, along with many more potential factors.

It is not trivial to go from definitions of reintegration to a measurable indicator. One challenge with measuring reintegration is that it is not a single variable. For some indicators, such as a person's yearly income, or average exam scores for a school, there is usually a fairly clear single variable that can be ascertained.

Reintegration, however, is made up of many different dimensions:

Financial situation

Housing

Employment

Social stability

Health

Perception of their satisfaction and happiness

and more…

- Which of these need to be used in the measure of reintegration?
- And what is the relative importance of each?
- Which are more relevant to evaluating the impact of a particular programme or intervention?
- How do you decide?

DECIDING HOW TO MEASURE REINTEGRATION

When planning an impact evaluation for a reintegration programme, careful thought must be given to establish the most suitable way – or ways – of measuring, based on:

- Project goals
- Returnee perspectives
- Agency priorities
- Data availability
- Uses for information
- Users of information
- Many other factors.

There is currently not a single, agreed-upon way reintegration is measured and calculated.

This does not have to mean planning an impact evaluation necessarily involves inventing a new way of measuring reintegration!

Various groups, such as IOM, the Regional Durable Solutions Secretariat (ReDSS) and the Inter-Agency Standing Committee (IASC), have used a range of approaches and de facto definitions, and these existing ways of measuring can be adopted or adapted for use if suitable.

Those individuals and groups who will make use of the outputs of the impact evaluation need to have agreed on the definition and strategy for measurement that is going to be used before the evaluation is implemented, or the results may not be of use to them.

International Organization for Migration (IOM)
Web page: *www.iom.int*

ReDSS
Web page: *www.regionaldss.org*

Inter-Agency Standing Committee (IASC)
Web page: *https://interagencystandingcommittee.org/the-inter-agency-standing-committee*

INDICES

When aiming to use reintegration as an indicator to measure impact, there are different strategies for accommodating the fact that it is not a single measurable variable.

One of these is to measure reintegration as a composite index; multiple indicators are combined, based on the definition being used, to create a single "index" or score. The result is a single number that summarizes the level of reintegration that has been attained, allowing for reporting and comparison These composite indexes are more than a way of combining information – they define what reintegration means for the purpose of the study or studies that use it and the decisions that may be made as a result.

As we have mentioned previously, it is important to involve potential users and groups involved in or affected by the evaluation, in determining how reintegration will be defined and measured.

Find out more

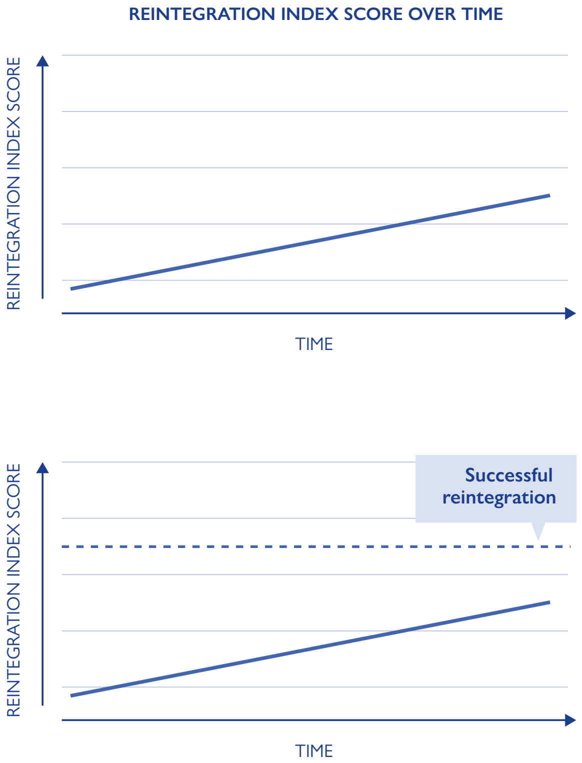
Handbook on Constructing Composite Indicators

PDF: *www.oecd.org/sdd/42495745.pdf*

Throughout this module, we use the term "index" to refer to the composite indexes described here.

As we will see later in this section, there are several existing indices, as different actors have come up with different strategies for measuring reintegration, based on their needs and priorities.

Normative thresholds



Some indices include thresholds.

These are a bit like targets, or benchmarks, that allow measurements of reintegration to be compared against a theoretically constructed idea of "successful reintegration", or "good" or "poor".

This provides an insight into what the number means – a score of 20 per cent is hard to understand in isolation, but given the information that 10 per cent is "extremely poor" and 70 per cent is "successful reintegration", we can draw more useful conclusions.

WEIGHTS

Calculating a reintegration index involves combining information about a number of different factors, but this doesn't just mean adding them up – how they are combined is another crucial aspect of deciding how to measure reintegration.

If certain factors, such as income, are deemed to be more important than others in terms of determining reintegration success, then the calculation that is done to produce a reintegration index should reflect that by "weighting" that factor so it has more impact over the final score than a less important factor.

How is that importance – that weighting – decided? Here are two common ways:

Expert-derived weights

A process of discussion with experts – local authorities, community leaders, experienced reintegration practitioners – leads to a conclusion on the relative importance of these factors. The weights are calculated to reflect this, with higher weighting given to those things that are deemed more important and vice versa.

Data-driven weights

Perhaps you have data on all sorts of factors (emotional well-being, health, employment, income, education level). You can put them together into a statistical model and conduct analysis to see how much of an influence each factor has on the reintegration index created.

Maybe you see that there is clear pattern, where changes in emotional well-being are more correlated with the overall index than the other variables. You would be able to say that the emotional well-being is the most important aspect explaining changes observed in the overall reintegration index.

This kind of analysis would eventually allow the derivation of weights for the calculation of a new reintegration score based on the data you initially analysed.

As with other decisions about measuring reintegration, which way of deriving weights is "better" depends on factors such as stakeholder priorities, uses for information, project goals and more.

Question

Which of the following statements about weights are true? Select all the answers that apply.

- □ Weights are part of the calculation of indices, such as reintegration scores.
- □ Weights are used to calculate which indicators are highly correlated with reintegration success, so they can be used as a proxy for reintegration.
- □ Weights can be determined by analysing data to see how much specific aspects are correlated with reintegration.
- □ It is best for an experienced reintegration practitioner to determine the expert-derived weights for reintegration indices as they are less likely to be biased.

The answers are on the next page.

Answer

☑ **Weights are part of the calculation of indices, such as reintegration scores.**
Correct, weights are used to determine how much of an effect different variables have on the overall score.

□ **Weights are used to calculate which indicators are highly correlated with reintegration success, so they can be used as a proxy for reintegration.**
It is possible to use similar statistical analyses that are used to derive weights to investigate possible proxy indicators, but this is not the purpose of weights. Weights instead are used to adjust the effect particular indicators have on a final calculated index score.

☑ **Weights can be determined by analysing data to see how much specific aspects are correlated with reintegration.** This is correct.

□ **It is best for an experienced reintegration practitioner to determine the expert-derived weights for reintegration indices as they are less likely to be biased.**
This is incorrect. While such a person would have useful input, it is important to incorporate multiple and other perspectives, particularly from the community that is being represented by the measurement.

INTERVIEW TRANSCRIPT: IOM'S MEASURE OF REINTEGRATION

Joy Paone, Capacity-building Project Officer at the IOM Knowledge Management Hub, talks about IOM's measure of reintegration, the Reintegration Sustainability Index:

IOM'S REINTEGRATION SUSTAINABILITY INDEX

Interview with Joy Paone, Project Officer (Capacity-building), IOM Knowledge Management Hub

So, my name is Joy Paone. I work for the EU-IOM Knowledge Management Hub for return and reintegration at IOM's Headquarters in Geneva. I've been with IOM now for over seven years, focusing on reintegration for the last four years. Before, I managed the development of the Reintegration Handbook and have supervized a comparative analysis based on the Reintegration Sustainability Survey. And, in addition to the Knowledge Management Hub's capacity-building activities, I also oversee the hub's monitoring and evaluation component, which aims to support cross-regional harmonization of monitoring and evaluation activities.

So, IOM has been implementing return and reintegration programmes as part of assisted voluntary return and reintegration programmes for over 40 years, and over the years we've seen that the understanding around reintegration has evolved. So, it's gone beyond providing an incentive for migrants to leave voluntarily, towards an understanding that assistance to migrants upon return is really necessary to support their reintegration process and to make it sustainable.

But there's never been a standardized definition for what sustainable reintegration means, let alone a standardized way to measure it, so what we have are many different, scattered programmes in the field of return and integration, each with their own objectives and monitoring approaches. So, IOM tried to fill this gap by creating a definition of sustainable reintegration. and then a way to measure it.

Development of the RSI and RSS

So let me give you a brief history of how the Reintegration Sustainability Index (RSI) and its Survey came about. So, as I just mentioned, the Reintegration Sustainability Index is closely interlinked with IOM's definition for sustainable reintegration. Back in 2017, based on existing research available, looking at factors affecting reintegration and IOM's experience in reintegration programming, IOM developed an institutional definition for sustainable reintegration. But how could we tell that an individual has achieved sustainable reintegration?

So, it was under the UK-funded "MEASURE" project in 2017 that Samuel Hall field-tested a set of indicators, which were mostly telling of the degree of sustainable reintegration among beneficiaries. What came out of this are 15 field-tested indicators and 30 measurement elements, 32 questions, separated into three sections, encompassing economic, social and psychosocial dimensions of reintegration.

How does the RSI reflect IOM's definition of sustainable reintegration?

So, IOM's definition of sustainable reintegration is that "Reintegration can be considered sustainable when returnees have reached levels of economic self-sufficiency, social stability within their communities and psychosocial well-being that allow them to cope with remigration drivers. Having achieved sustainable reintegration, returnees are able to make further migration decisions a matter of choice rather than necessity."

So, first, we see in this definition that reintegration concerns returnees and the communities to which they return. It's also linked to structural factors in the external environment. The Reintegration Sustainability Index focuses on the individual returnees' level of reintegration, not so much on the community and structural level. It does this by asking questions about the returnees' own perceptions.

Secondly, what we see in this definition is that reintegration is a multifaceted phenomenon that refers to economic, social and psychosocial dimensions. It means we have to further define what each dimension means. So, to give an example for economic self-sufficiency, we mean that returnees are exercising a livelihood activity that allows them to support themselves, that the livelihood activity allows them to support their family or household, that employed returnees are satisfied with their employment conditions, that returnees have effective access to opportunities such as training, which can improve their qualifications for employment, and that returnees do not have debts that hamper their self-sufficiency.

The third point that we see in this definition is that remigration doesn't necessarily imply a lack of sustainability. What counts is whether new migration happens as a matter of choice or not. So, in the Reintegration Sustainability Survey, we have questions that look at whether the returnee feels the need to consider further migration – either legal or irregular – as an exit strategy from reintegration challenges, or as a matter of choice.

So, as you can see, the Reintegration Sustainability Index reflects IOM's definition for sustainable reintegration. Depending on the return context, further optional indicators or questions may be added, while still maintaining the others to allow for comparability.

How is the RSI used?

So, these questions are linked to a scoring system where you get three dimensional scores, measuring reintegration at the economic, social and psychosocial dimensions and a composite reintegration score that provides a numerical measure of overall reintegration across dimensions and is a useful tool for evaluation, reporting and analysis.

So, developing the Reintegration Sustainability Index was only one step; we then developed a full monitoring and evaluation framework and guidance for return and reintegration, all based on IOM's definition for sustainable reintegration. The Reintegration Sustainability Survey is one of the tools of this package. So, this M&E package was then institutionalized in IOM's migrant management operational system. As many programmes as possible have been encouraged to use it.

So, with the Reintegration Sustainability Survey as part of the overall monitoring and evaluation framework, standardized data can be collected across reintegration programmes globally. This allows IOM to compare trends in beneficiary reintegration across dimensions, country contexts and over time to support those working in reintegration in understanding the reintegration process of individuals they work with.

It can be applied at different stages of the reintegration process; for example, soon after return as a baseline, as an interim progress assessment and for final monitoring.

It has different intended users. It can be used as a case management tool for case managers, because it helps to understand returnees' reintegration process and needs and adjust the provision of reintegration assistance accordingly. For example, if, at baseline, the returnee has a lower psychosocial score than the other dimensions, then the case manager knows to focus on that when developing their integration plan with the returnee.

It can also be used for monitoring and evaluation purposes. It can be key for programme evaluation by helping identify and address gaps.

KNOWLEDGE CHECK

Question 1

Why is it difficult to decide how to measure reintegration? Select all the answers that apply.

- □ The definition of reintegration depends largely on the situation, priorities and people involved.
- □ Reintegration is too complicated to be quantified as a numerical value.
- □ A returnee's level of reintegration is multidimensional – it is made up of many different criteria.

Question 2

Which of the following statements are true about deciding how to measure reintegration for an impact evaluation? Select all the answers that apply.

- □ It is crucial to use the method your organization has developed so that evaluation results are standardized.
- □ The definition and way of measuring reintegration needs to be agreed on by the intended users of the evaluation outputs.
- □ The measurement of reintegration should be relevant to returnee perceptions of what reintegration looks like.

The answers are on the next page.

KNOWLEDGE CHECK ANSWERS

Question 1 answer

Why is it difficult to decide how to measure reintegration?

☑ **The definition of reintegration depends largely on the situation, priorities and people involved.**
This is true, and so how reintegration is defined and measured should be considered carefully and agreed with relevant stakeholders.

□ **Reintegration is too complicated to be quantified as a numerical value.**
While it is not a simple task, various organizations have developed ways to measure reintegration as a composite indicator.

☑ **A returnee's level of reintegration is multidimensional – it is made up of many different criteria.**
One challenge with measuring reintegration is that it is not a single variable. For some indicators, such as a person's yearly income, or average exam scores for a school, there is usually a fairly clear single variable that can be ascertained.

Reintegration, however, is made up of many different dimensions, including:

- Financial situation;
- Health;
- Employment;
- Housing;
- Social stability;
- Perception of satisfaction and happiness.

These different criteria are sometimes summarized in one composite score.

Question 2 answer

Which of the following statements are true about deciding how to measure reintegration for an impact evaluation?

□ **It is crucial to use the method your organization has developed so that evaluation results are standardized.**
This is not necessarily true. It is more important to take an approach that is suited to the evaluation aims and stakeholder priorities.

☑ **The definition and way of measuring reintegration needs to be agreed on by the intended users of the evaluation outputs.**
This is important, because an evaluation needs to produce information that meets the requirements of the users in order to achieve its aims.

☑ **The measurement of reintegration should be relevant to returnee perceptions of what reintegration looks like.**
Returnees are at the centre of return and reintegration programmes. They are the ones affected by the results of an evaluation, so it is crucial that their views are taken into account.

EXISTING WAYS OF MEASURING REINTEGRATION

Various international agencies and non-governmental organizations have developed approaches for measuring reintegration in the wider context, including internally displaced people (IDPs) and returning refugees.

Here are some examples:

REINTEGRATION SUSTAINABILITY INDEX (RSI)

Description:

IOM's international standard for measurement for returnee reintegration, developed by Samuel Hall in 2017, based on IOM's definition of sustainable reintegration:

> "Reintegration can be considered sustainable when returnees have reached levels of economic self-sufficiency, social stability within their communities, and psychosocial well-being that allow them to cope with (re)migration drivers. Having achieved sustainable reintegration, returnees are able to make further migration decisions a matter of choice, rather than necessity."

Towards an Integrated Approach to Reintegration in the context of return (IOM, 2017) p.5

PDF: *www.iom.int/sites/g/files/tmzbdl486/files/our_work/DMM/AVRR/Towards-an-Integrated-Approach-to-Reintegration.pdf#page=5*

Note that this definition emphasizes the need for reintegration to be "sustainable". The questionnaire includes a couple of questions about whether the returnee feels able to stay and live in the country and why they feel that way.

- Combines 31 indicators of reintegration across economic, social and psychosocial dimensions that were selected from a longer list of possible indicators through a numerical technique called "Principal Component Analysis".
- Returns a score from 0 to 1 with higher scores denoting a better level of reintegration. Scores specific to the economic, social or psychosocial dimension can also be computed.
- Uses expert-derived weights.

Strengths:

- Internationally standardized.
- Computationally straightforward.
- RSIs can be compared over time and across contexts.
- RSI has normative thresholds, indicating what index scores can be interpreted as "poor" (less than 0.33), "borderline" (between 0.33 and 0.66) and "adequate" (above 0.66) levels of reintegration.

Care must be taken when making comparisons. For example, for certain indicators, like health care and drinking water, returnees in urban areas would have a better score than somewhere like rural Somalia or the Sudan. Someone well reintegrated who is located in a rural area might have poorer access to drinking water than someone poorly integrated who lives in an urban area where drinking water is more readily available. Thus, the returnee in the rural area might have a lower RSI score despite being better reintegrated, just because of the inherent differences in the locations.

Looking at just the composite score without considering individual dimension scores may misrepresent returnees' circumstances. For example, someone might score low on the psychosocial dimension, but be doing very well economically and socially. Their overall score will not capture the nuances of their situation.

When using the RSI to make comparisons, it is important to remember contextual factors, such as location-specific differences, and an understanding of the three dimensions that contribute to the score.

Weaknesses:

- The values on the normative threshold are arbitrary and do not account for context-specific conditions.
- Some of the questions are challenging, and calculation of the index does not allow for missing values.
- It was not designed for use with non-migrants and thus presents limitations for studies wanting to use non-migrants as a comparison group.

Setting standards for an integrated approach to reintegration

A report about MEASURE. Gives a good insight into how reintegration can be measured and understood.

PDF: *www.iom.int/sites/g/files/tmzbdl486/files/our_work/DMM/AVRR/IOM_SAMUEL_HALL_MEASURE_REPORT%202017.pdf*

IOM Knowledge Bite #1

Introduces IOM's "Knowledge Bite" series, giving an overview of information about the factors affecting sustainable reintegration outcomes gleaned from analysis of RSS data collected by IOM.

PDF: *https://returnandreintegration.iom.int/sites/g/files/tmzbdl341/files/documents/knowledge_bite_1_-_introduction_0.pdf*

IOM Knowledge Bites series

Factsheets containing insights emerging from RSS data.

Web page: *https://returnandreintegration.iom.int/en/global-search?keyword=knowledge+bites*

IASC Durable Solutions for Internally Displaced People

PDF: *https://interagencystandingcommittee.org/system/files/iasc_framework_on_durable_solutions_for_idps_april_2010.pdf*

Durable solutions indicators and guide

Web page: *www.jips.org/tools-and-guidance/durable-solutions-indicators-guide/*

Description:

- Designed to determine the extent to which a durable solution for forcibly displaced people has been achieved.
- Looks at eight criteria, including: safety/security, standard of living, job opportunities, etc.
- Focuses on IDPs only, does not include explicitly returnees and host communities/non-migrants.
- The framework has been operationalized by an inter-agency process, coordinated by JIPS which provides a library of indicators that could be included. They also provide broad guidance on analysis, which includes the list of indicators that could be used.

JIPS library of indicators
Web page: *https://inform-durablesolutions-idp.org*

JIPS list of indicators
PDF: *www.jips.org/uploads/2018/10/Interagency-Durable-Solutions-Analysis-Guide-March2020.pdf*

Strengths:

- The IASC framework is well received and has been embedded in the development of several reintegration assessment methodologies.
- As it is an inter-agency framework, adoption and use over multiple organizations is easier than for some other ways of measuring.

Weaknesses:

- Work is required to select indicators and transform these into a data-collection tool.
- No methodology for combining indicators into a single index.

Multidimensional Integration Index (MDI) pilot results

Web page: *www.samuelhall.org/publications/unhcr-the-multi-dimensional-integration-index-pilot-results*

Description:

Developed as an inter-agency approach with the Afghan Government, the United Nations and non-governmental organizations. A standardized tool to measure reintegration/integration levels of returnees and IDPs.

- Based on the IASC Framework for IDPs.
- Determines the extent of reintegration of returnees in the context of the community of return.
- Three components: comparison with local populations, range of integration experiences, assessment of self-perceptions of integration.
- Three dimensions of assessment: economic, social and safety realm.
- The index is built on objective indicators, complemented with subjective indicators.
- MDI scores range from 0–1 (or 0–100%) and are presented in a "traffic light" system: scores greater than 0.9 indicate that returnees are hard to distinguish from the host communities, which is considered to be full integration (green); scores of 0.8–0.9 are somewhat distinguishable from host communities (orange); and those below 0.8 are very distinguishable from their local host community across a range of indicators.

Strengths:

- Designed to be standardized internationally and across agencies.
- Question set can be amended and recalibrated to adjust for societal evolutions over time.
- Allows comparison of different types of displaced populations.
- Focuses exclusively on displacement-related vulnerabilities while controlling for external factors such as location, general absence of economic opportunities, security, etc.

Weaknesses:

- The degree of the MDI only indicates similarity to local host populations; it does not give information about overall well-being. Care should be taken when making comparisons.

SELF-RELIANCE INDEX

Refugee Self-Reliance Initiative (SRI)

Web page: *https://reliefweb.int/report/world/self-reliance-index-version-20-indicators-measure-progress-towards-self-reliance*

Description:

- A measurement of self-reliance of refugee households over time, also applicable to migrants.
- The data collection combines discussion with the clients, direct observation (during a home visit if one takes place), knowledge of local conditions and any prior knowledge of the household's circumstances, to arrive at an evaluation.
- Combines 12 components relating to self-reliance: housing, food, education, health care, safety, employment, financial resources, assistance, debt, savings and social capital.

Strengths:

- Non-statistical method for creating single index value ranging from 1–5.
- Small number of questions provide a rapid assessment of self-reliance component of reintegration.

Weaknesses:

- No justification given for the methodology of taking the individual domain scores to create the SRI.
- Although questions are simple, they can require quite detailed probing to get the final value for each component.
- Self-reliance is just one component of reintegration.

LOCAL REINTEGRATION ASSESSMENT, LORA – IOM SOMALIA

PDF: *https://regionaldss.org/wp-content/uploads/2020/12/Danwadaag_Measuring-the-End-of-Displacement-Emerging-Learning-from-Somalia.pdf*

Description:

- The assessment evaluates the extent to which IDPs and returnees feel integrated, the inequalities between IDPs and their host communities and the level of displacement-affected communities' self-reliance.
- Collects data from host communities as well as IDPs and returnees.
- Builds on both the IASC framework described earlier and ReDSS Solutions Framework.
- Contextualized to Somalia, the questionnaire tool obtains 32 variables that contribute to assessment of these three aspects of local reintegration.
- Includes single question on perceived feelings of integration.
- The output includes a binary "integrated" or "not integrated" outcome – based on self-reported perception of integration – and results for the 32 variables. The results for the variables can be explored to see which of them correlate with perceived reintegration.

Strengths:

- Provides opportunity for a returnee's own perception of level of reintegration to be used as an outcome together with actual measures of inequalities and self-reliance.
- Questionnaire is of reasonable length (20–30 minutes).

Weaknesses:

- The single question on perceived level of integration is not comparable between returnees, IDPs and their host communities as it may be assumed that the definition of integration by a host community member may be quite different.
- Possibly vulnerable to bias as displacement-affected communities may understate their perception of integration in the hope of qualifying for further integration assistance.
- Multiple statistical analyses required to analyse the three aspects of local reintegration with no overarching model to bring these together.

LATENT LOCAL REINTEGRATION INDEX

MESH SHARP Programme Evaluation Methodology Notes June 2020

PDF: *https://static1.squarespace.com/static/5d0dee49c9ddd900015bd2e7/t/5f80896fadf5786e6c32f125/1602259312492/FCDO--SHARP+Programme+Evaluation+Overview.pdf*

MESH Danwadaag analysis-annual report 2019/2020

PDF: *https://static1.squarespace.com/static/5d0dee49c9ddd900015bd2e7/t/5f80829948634d787e87ca34/1602257564877/FCDO+Somalia--MESH--Programme+Evaluation+Annual+Report--Danwadaag.pdf*

Description:

- Uses the same framework and questionnaire tool as LORA to provide a combination of proxy outcome indicators of local reintegration (i.e. how a returnee would look if reintegrated) and variables to measure local reintegration (i.e. drivers of integration).
- Uses a multiple indicator, multiple cause (MIMIC) model to bring these aspects together to create a latent (unmeasurable directly) local reintegration index.
- Produces a single value index score based on data-driven weights that are calculated within each run of the model. The weighting is not consistent across multiple runs of the model, and the index score does not have a standardized format (for example, between 0 and 1), meaning scores from separate model runs cannot be compared.
- Applicable across the whole displacement-affected community (IDPs, returnees and host communities).

Strengths:

- Accommodates multiple measurable integration outcome proxies with multiple integration drivers.
- Can easily be adapted to different or additional drivers and/or outcome proxies to better reflect local context.
- Provides single index of integration that can be compared across locations and displacement-affected communities (e.g. returnees versus host communities).

Weaknesses:

- No normative thresholds for the index values.
- Cannot directly compare scores across time.
- Statistically complex, therefore harder to explain and gain credibility with a larger audience.

KNOWLEDGE CHECK

Depending on the aims of an impact evaluation, different aspects may be important when choosing or developing a way to measure reintegration.

Imagine you are planning an impact evaluation of a reintegration assistance programme that has begun rolling out a strategy for providing monetary support to returnees using mobile money transfers. This implementation has begun in some but not all programme locations. There is an interest in seeing the difference in the programme's impact on beneficiaries' level of reintegration between locations that have and have not begun the mobile money strategy.

Budget and time available for data collection are particularly limited.

Intended users of the evaluation outputs include stakeholders across multiple organizations and countries who are interested in the success of the mobile money strategy. There is a hope that the evidence produced by the evaluation will encourage wider adoption of the strategy and support fundraising to enable it.

Question

Which three of the following features of possible ways of measuring reintegration would be most important in this scenario, given the information provided? Choose three answers.

□ Generation of a single reintegration index.

□ The measurement is widely understood and accepted.

□ The measurement includes a normative threshold to measure against.

□ Ease of comparison e.g. over time, different locations.

□ Ease of enumeration.

Answer

□ **Generation of a single reintegration index.**
This could be helpful for simplicity of communication and comparison, but it's not the most vital priority.

☑ **The measurement is widely understood and accepted.**
Correct. Given the intention for the information to be used internationally, by multiple organizations, this is important.

□ **The measurement includes a normative threshold to measure against.**
While this might be useful, the important comparison for this evaluation's aims is between locations with and without the mobile money strategy.

☑ **Ease of comparison e.g. over time, different locations.**
Yes, being able to make comparisons between the locations and over time will enable the evaluation to understand the difference in impact the mobile money strategy makes.

☑ **Ease of enumeration.**
Correct. Resources for data collection are a concern, so the enumeration needs to be as quick and simple as possible.

MEASURES OF REINTEGRATION

Reintegration measure	Provides a single index	Widely understood and accepted	Includes a normative threshold	Easy to make comparisons	Easy to enumerate	Applicable to...	Allows for localization adjustments
Reintegration Sustainability Index (RSS)	✓	✓	✓	✓ Some caveats		Returnees	✓
Operationalised IASC Framework on Durable Solutions for Internally Displaced Persons		✓				Internally Displaced Persons (but can be broadened to consider whole displacement affected communities)	✓
Multidimensional Integration Index (MDI)	✓	✓	✓	✓		Displaced populations, including returnees and IDPs	✓
Self-reliance Index (SRI)	✓			✓	✓	Refugees (but can be adapted)	
Local Reintegration Assessment, LORA – IOM Somalia, 2020					✓	Whole displacement-affected community: IDPs, returnees and host communities	
Latent Local Reintegration Index	✓			✓ Except over time		Whole displacement-affected community: IDPs, returnees and host communities	✓

INTERVIEW TRANSCRIPT: DEVELOPING THE RSI

Interview with Stefanie Barratt, Data Standards and Analytics Pillar Lead at Samuel Hall, talking about Samuel Hall's work with IOM on the development of the RSI composite index.

DEVELOPMENT OF THE REINTEGRATION SUSTAINABILITY INDEX

Interview with Stefanie Barratt, Data Standards and Analytics Pillar Lead, Samuel Hall

My name is Stefanie Barratt. I'm an economist by training. I've been working with Samuel Hall for eight years. Samuel Hall is a think-tank and social enterprise, focused on migration studies. We've over the years worked a lot on return and reintegration specifically.

So what do people mean when they talk about reintegration? Do they all mean the same thing? And the best thing to ensure that, you know, people speak the same language when it comes to concepts as important as this one is for them to use a similar metric – similar indicators – to track progress.

How does the RSI reflect IOM's definition of sustainable reintegration?

The RSI is built on the IOM definition. It defines "sustainable reintegration" as the moment where returnees have reached levels of economic self-sufficiency, social stability within their communities and psychosocial well-being, and these levels of sufficiency and well-being allow them to cope with remigration drivers. And that means that a remigration decision becomes a matter of choice rather than necessity. So, there's a lot already in there to work with when you're building a metric.

So, from the onset, you know you need the multidimensional approach to reintegration. What are the dimensions? Well, they're economic, social and psychosocial; it's right there in the definition. But there's more.

In the definition, we talk about social stability within their communities, so that means that interventions need to look at the individual, the community, the structural level as well. And that the reintegration metric cannot just focus on the individual, but needs to also talk a little bit about the context.

The definition also looks at reintegration as a process, at the end of which a goal is achieved. So, that means we can't just go in and measure one time; the tool needs to be usable over time to determine trends and progress. We know that the reintegration process isn't linear; it's well established by a whole body of literature that there are ups and downs post-return in your well-being. So, the tool needs to be able to track these evolutions and allow IOM and its partners to address the low points and build on the high points to render integration sustainable.

In this definition, integration is conceptualized as something that will eventually be achieved. The metrics should probably be able to inform a graduation approach: a time-bound and sequenced series of programming that we carry out, and eventually we're able to say, "we have succeeded; the migrant is now reintegrated".

How the RSI was developed

So, with these guidelines in mind, we set out to develop the metric and that of course first starts with the indicator selection. So, to choose the indicators, we started off with the literature review and we reviewed over, I think it was 150 papers – IOM publications but also academic articles – and of course we went back to Samuel Hall's own work over 10 years of working on return and reintegration in different contexts.

We conducted almost 100 key informant interviews over the course of this project, with IOM staff in country and headquarters, and experts – academics. We went and spoke with returnees using a qualitative and quantitative tool, so we could understand what reintegration meant in these individual contexts, but also more universally across contexts.

So, based on this, and with our three dimensions in mind, we selected indicators, tested indicators and then, after the data collection, had technical validation workshops with IOM and local partners to share and discuss key findings.

And now we had a tool and we had these indicators, but the question is: how do you go from that – that filled-in survey, if you will, to a score? Actually, to four scores: three in the individual dimensions and one overall reintegration score. So, the method we used was principal component analysis (PCA). It's often used to derive scores; it's based on the principle that in a data set with many variables, a lot of the information is redundant. You can largely infer the value of one variable from a combination of the other variables in the data set for the same individual. So, that means a data set can be compressed. And the point of PCA is to transform a large set of variables into a smaller set of variables that still contains most of the information in the larger data set.

We ran the PCA calculations to look at the relationship between the individual variables and to see which ones were correlated across the sample. And so, based on this matrix the method builds new variables called principal components that are uncorrelated, but still contain most of the information within the individual variables. So, you end up, in the end, with as many principal components as you had variables to begin with, but it's usually the first one that explains the bulk of the variation in the sample.

We have this first principal component that explains the variation in the data set to a large extent and we scale that to fit a zero-to-one index. In the end, we ended up with our set of variables, our indicators, reduced to one single digest. So, via this method, you ended up with four scores as part of the global reintegration scoring system: the economic, the social, the psychosocial and the overall reintegrated score. In each dimension, the sum of these weighted indicators falls between zero and one, and one represents full reintegration – so it's the perfect score – and zero represents a total lack of reintegration.

Limitations of the RSI

The RSI currently looks at reintegration as an absolute; it doesn't take into account the local population. If a community lacks access to formal health care, then the fact that the migrant or displaced household or returnee also lacks such access is perhaps not an indicator of integration or lack of integration. On the other hand, if most households in the community enjoy access to the grid – access to electricity – and the returnee households most often do not, then then that might be a good indicator of integration.

When the RSI was designed, it was always to be used in combination with other qualitative tools to ensure that programming is adjusted to the personal circumstances of the returnee. The metric allows you to track reintegration, but it doesn't explain anything. It doesn't tell a story, so it doesn't do justice to the individual situation of every returnee in their individual context. So, scores always need to be triangulated, and qualitative tools are definitely needed to inform a reintegration plan.

Advantages of the RSI

So, this is the first time that there's been a universal metric for this thorny concept of reintegration. All country offices of IOM now speak of the same thing when they talk of reintegration in the context of AVRR. And, perhaps more broadly, the index has also advanced the conversation, so more analysis is being done on this data, and it's not just used for case management. And so, IOM is publishing a learning series on it, and the community of practice can still learn a lot from the RSI data.

DATA-COLLECTION CHALLENGES

The complexities of the return and reintegration context can make the process of collecting data challenging. The challenges we will be looking at in this section are: timing data collection, a lack of baseline data, mobile population, risk of bias and contamination.

TIMING DATA COLLECTION

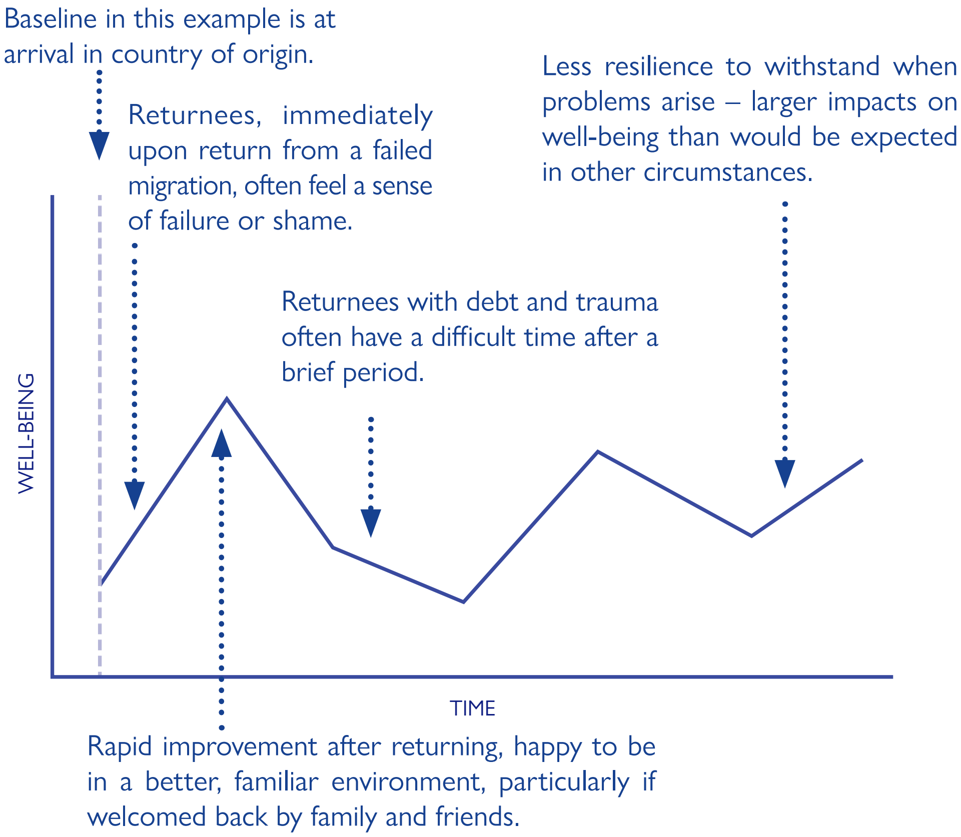
The timing of data collection can be crucial for getting useful results for impact evaluations of return and reintegration programmes. Returnees are prone to rapid changes in their situation and well-being, and the expected timelines of effects of programme interventions are varied.

Deciding when to collect data requires a good understanding of the context and causal mechanisms of your planned programme intervention. You also need to be aware of the project timeline requirements and any potential delays to intervention implementation and delivery.

Rapid changes

When returnees come back, their situation is unstable geographically, but also psychosocially. There may be big "swings" in how well they are doing, as shown in Figure 37.

FIGURE 37: EXAMPLE OF RAPID CHANGES IN A RETURNEE'S WELL-BEING

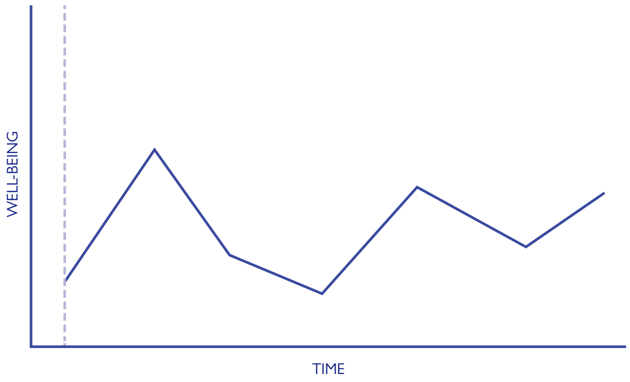


Note that changes in well-being are sometimes heavily related to whether the programme delivers support reliably and on time.

When do you take measurements?

Many indicators that we look at to understand reintegration success will be erratic. This can make it difficult to get a measurement of these indicators that will provide useful information about the impact of programme interventions.

When planning your evaluation and establishing the evaluation questions, it is important to recognize that reintegration is not static or a straight-line increase over time.



Rapid changes: mitigation

There has been little research on the issue of these rapid changes, so existing strategies for resolving the issue are limited.

One possibility is to undertake rapid, mini "snapshot" surveys, which collect only a small number of key indicators. This can help get an idea of what is happening between full surveys.

Another idea being explored is the possibility of collecting some retrospective data, asking respondents to look back some time after the baseline – a year later, for example.

The hypothesis is that, with hindsight, returnees might be able to give a more objective report of their status. Data are being collected to test this at the time of writing.

Note that recall data comes with risk: memories can be unreliable, and events that happened later can influence how people perceive their past.

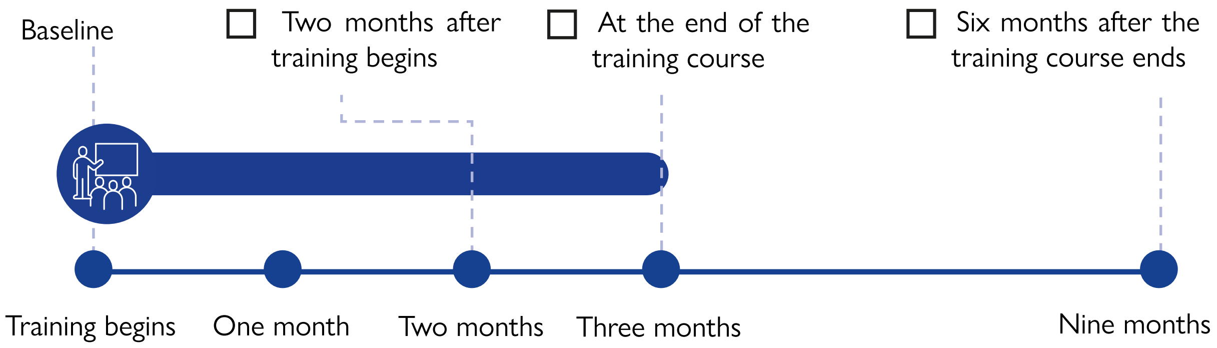
Time to maturity of impact

Some intervention effects can be expected to take some time to "mature". Let's consider an example.

A programme is providing a three-month long business skills training course. The evaluation team plans to take a baseline and endline measurement of beneficiaries' household income to measure the effect of the training.

Question

Considering for now only the time to maturity of this impact, and not other programme considerations that may affect the decision, when do you think the endline should be conducted?



□ Two months after training begins

□ At the end of the training course

□ Six months after the training course ends

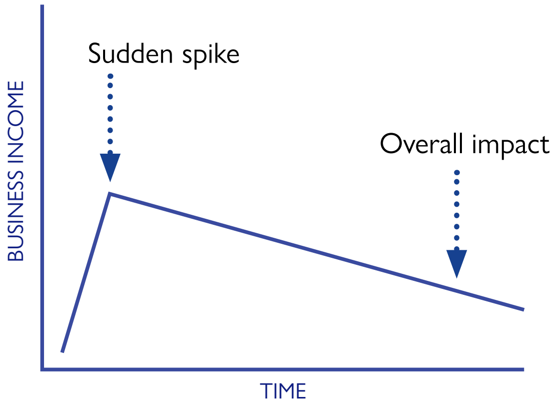
Answer

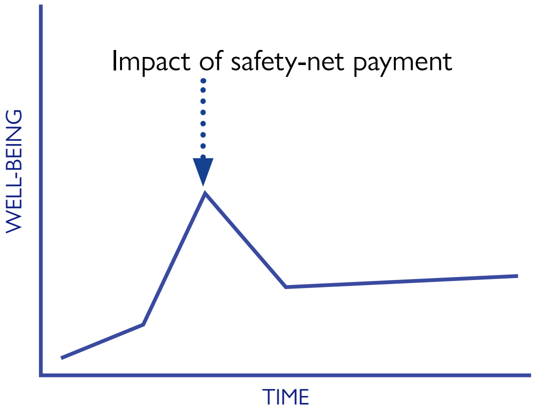
☑ Six months after the training course ends

Consider how much time it will take for training activities to complete and for beneficiaries to experience a change in household income due to their new or improved business. They cannot expect to see improvement in household income as soon as the training has completed!

Other interventions might cause a sudden spike in an indicator that stabilizes over a period to what can be considered the overall or sustained impact.

For example, an intervention provides business start-up grants at the end of skills training. In the short term, the start-up grant enabled the returnee to establish the business and to buy and sell initial stock. But over time, they were not able to sustain the same level of business profit.





Alternatively, some interventions may only lead to short-term impacts. For example, the impact of an intervention such as safety-net cash payments may only be observable in the short term – perhaps a few weeks after the intervention is implemented.

Time to maturity of impact: mitigation

It can be hard to judge when to measure. There may not be a single “correct” answer, but rather a “best guess”. Multiple postintervention measurements may be necessary to establish sustainability of the impact.

The key is to think ahead about what the impact is expected to be and when it might be possible to observe the indications of it. Other considerations, such as programme logistics and practicality of data collection, may also affect this decision.

LACK OF BASELINE DATA

The aim of reintegration might be to get returning migrants at least back to a situation similar to premigration, preferably better. In this case, the logical baseline would be before migration.

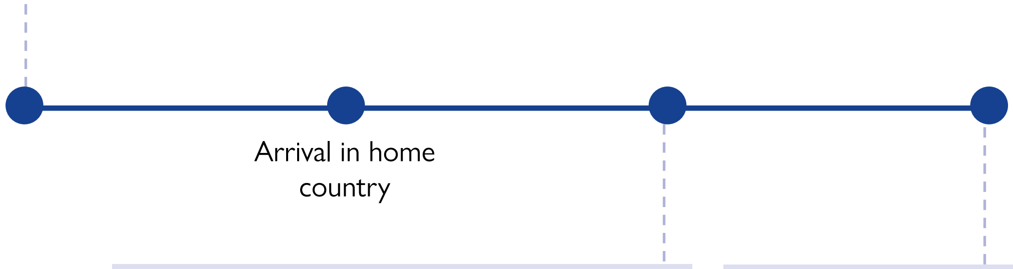
This would be almost impossible to collect and extremely variable with returnees migrating to and from different countries, for varying durations, having different push-pull factors to explain the reasons for their migrations, etc.

Lack of baseline data: mitigation

There are other "baselines" we can consider and for which we can obtain data:

If we need to use the baseline to compare the situation before the returnee joined the programme and at the end of or after the programme we could conduct a baseline immediately prior to their return to their home country.

- This might be particularly relevant if the definition of the treatment includes support to return.
- But what would that be measuring? We cannot measure reintegration before the migrant returns!
- Baseline data prereturn may still be useful for the programme, for example, in terms of looking at the efficiency or targeting of programme participants.
- It may also be useful if we have a comparison group of returnees not registered into the programme or if we have an impact indicator which looks at the overall change from prereturn to end of programme (e.g. monthly household income).



If we are evaluating impact in terms of a level of reintegration, we can only conduct the baseline after the returnee arrives back in their home country.

While this baseline misses evaluation of the initial return activities, it will provide a reintegration measure to provide a comparison of before and after.

As mentioned previously, it is possible to collect information retrospectively to provide some baseline data.

- This can be a slightly problematic strategy due to the potential for inaccuracies.
- However, it could be a suitable way of getting some kinds of information, such as previous salary or employment.

Regardless of the point at which the baseline is conducted, bear in mind that the data collection for the group receiving the treatment and the comparison group should take place at as closely together in time as possible.

MOBILE POPULATION

When returnees come back, they tend to move around a lot due to social stigma, financial difficulties and other challenges. They are likely to change phone numbers and relocate soon after arrival.

This is a challenge with making before/after comparisons, particularly if, for example, you want to conduct a longitudinal study, which looks at the same respondents on multiple occasions. This causes problems with retention. As well as being hard to keep track of, returnees might also disappear from the study entirely – they may deliberately exit the programme, or be impossible to find.

Mobile population: mitigation

There are a number of strategies to reconnect with returnees:

Use multiple means of contact, including social media profiles (such as Facebook). These are less likely to change than phone numbers.

Identify and make use of informal network**s** formed among returnees to assist in reconnecting with returnees.

RISK OF BIAS

There are various things that may cause biases in the data collected.

- **Reliance on programme.**
  A returnee who is reliant on programme support might, for example, feel that if they report that they are doing well then they won't receive further aid.

- **Motivation to "strategically" try to secure more support.**
  There may be returnees who feel they should give "good" answers as a sort of bribe to ensure future support from the programme; or returnees who give "bad" answers to appear worse what than they are in reality, in order to secure additional support.

- **Social and emotional influences.**
  Returnees might exaggerate positive responses to please the interviewer, or conversely express anger or dissatisfaction towards the programme by exaggerating negative responses.

- **Social stigma.**
  Concerns about the perceptions or retaliation of others may affect the answers given.

- **Emotional influences.**
  A returnee might not want to admit if they are struggling socially or financially due to feelings of shame associated with this.

- **Practicalities of contacting returnees to collect data.**
  For example, the returnees who stay in the same place, are contactable and happy to be interviewed and are easier to follow up with, are also likely to be the more successful returnees.

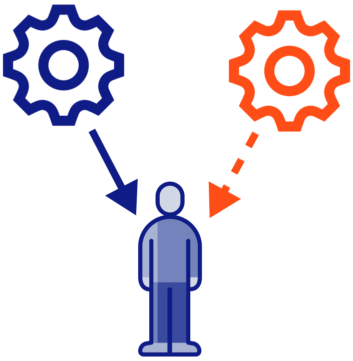
- **Enumeration and project staff.**
  If project staff is conducting the interviews or is involved in the processing of the data, there may be a conflict of interest, and it is possible for results to be falsified.

Risk of bias: mitigation

There are several things that can be done to reduce bias:

- Ensure good training of enumerators; this might help them with how to ask sensitive questions and encourage honest responses.
- Utilize relationships that are already built when possible.
- Careful wording of questions on the survey.
- Clear communication and explanations.
- Consider carefully who conducts interviews, manages data and supervizes data collection. Look out for potential conflicts of interest.

CONTAMINATION



Beneficiaries may be exposed to different forms of assistance and programmes besides the one being evaluated.

Many of the countries with return and reintegration programmes have many other challenges which means there are likely to be other programmes operating that affect returnees and their communities. This can make it difficult to isolate the impact of the programme or intervention being evaluated.

For example, in Somalia, there is a Durable Solutions consortia focusing on IDPs, returnees and non-migrant or host communities in poor urban environments. Besides the members of this consortia there are other donors operating humanitarian, resilience and governance programmes in these same poor urban locations.

Contamination: mitigation

Contamination can be detected by:

- Conducting key informant interviews and collecting qualitative information on what other programmes are operating in the area and their interventions.
- It may be possible to obtain data from the other agencies working in the area.
- Quantitative surveys could include questions asking respondents if they have received benefits from other programmes. However, respondents may not know from which programme they have received interventions.

KNOWLEDGE CHECK

Which of the following statements are true? Select all that apply.

□ The timing of data collection should always be determined by the time to maturity of expected impacts.

□ Information provided by returnees may be biased by their reliance on the programme.

□ Returnees should only receive support from one programme to avoid contamination.

□ If a returnee’s situation is changing rapidly, there is a risk of missing or misinterpreting changes and trends in the data collected.

□ If data are collected too early, the impact will be underestimated.

KNOWLEDGE CHECK ANSWERS

Which of the following statements are true?

□ **The timing of data collection should always be determined by the time to maturity of expected impacts.**
While this is certainly an important factor, there are other considerations involved in the decision, such as programme logistics or length of programme or project.

☑ **Information provided by returnees may be biased by their reliance on the programme.**
This is true – for example, returnees might feel they need to give positive responses to please the programme staff or exaggerate their problems to ensure continued support.

□ **Returnees should only receive support from one programme to avoid contamination.**
We cannot and should not prevent returnees from receiving assistance for which they are eligible. Instead, it is important to be aware of such contamination risks and factor them into data collection and analysis.

☑ **If a returnee's situation is changing rapidly, there is a risk of missing or misinterpreting changes and trends in the data collected.**
Returnees' situations tend to be prone to large and rapid changes. This can make it challenging to capture information and understand trends.

□ **If data are collected too early, the impact will be underestimated.**
This is not always true. It is also possible to overestimate effects of interventions that cause a large change at first before levelling out to the real impact.

INTERVIEW TRANSCRIPT: SELF-SELECTION BIAS

In this interview, Martin Schmitt, Regional M&E Officer at the IOM Regional Office in San José, Costa Rica, talks about the issue of self-selection bias in impact evaluations for return and reintegration programmes.

SELF-SELECTION BIAS IN IMPACT EVALUATIONS

Interview with Martin Schmitt, IOM Regional M&E Officer, San José, Costa Rica

My name is Martin Schmitt. I'm currently the Regional Monitoring and Evaluation Officer at the Regional Office for Central America, North America and the Caribbean in San Jose, Costa Rica.

I think one of the one of the main challenges, not only for impact evaluations but I think for studies in general is the self-selection bias. Self-selection bias has to be seen as a two-stage process, especially in return.

So, there is, one: self-selection into programming. So, there are only certain persons that are taking part in, or that are eligible to take part in, the programme. And, by this, you have to be careful when you are drawing conclusions that it's not about all migrants living in a certain country, but it's only about the persons that are really taking part in your programme. And this is also important to mention this then to the general public, to make this really understandable, that it's just a certain group that you are talking about.

And then the self-selection into your study. And there might be certain characteristics why persons are not taking part into your study. So, maybe persons who are very critical about IOM or the UN system don't want to take part because they don't have trust; maybe they have other characteristics in common that then could bias your results, and you're not having a representative sample.

So, basically, the self-selection, what it does is it compromises your sample. And, by this, you are not able to draw representative conclusions. And these are points that you have to consider, not only when setting up the study, but also when afterwards analysing your results.

I think the one of the most important aspects, in terms of self-selection, is to identify who are the persons who are dropping out. Because it's very hard to avoid some degree of self-selection, but it's important that you know who are the ones.

Controlling for self-selection bias

A good way to identify who is self-selecting themselves into the study is to connect your study data with the programmatic data. So, usually when there is a return, then the organization is collecting some basic demographic data. This, of course, does not cover all aspects like attitudes and so on, but it might be a very good indicator of how your sample is composed; for example, do you have the same age distribution? Do you have participants – especially important, for example, gender might be very critical, that you usually, in some countries, you might get issues to get sufficient women participating, for example, or elderly persons who do not want to or do not know how to use the computer. And if you then compare this or to connect this with the data, and then you can see – is there any difference? And then use statistical methods like weighting, for example, to then give those under-represented groups more power.

Another good option to control for potential biases that you have in the data is to make use of alternative data sources, like, for example, expert interviews. So, it's always good to complement your quantitative survey with qualitative aspects. These might be focus groups, might be expert interviews. So, by this, you get much more context, and you understand also much better the impact in the data that you have in your quantitative survey. And this, at times, also helps you to understand whether you really have a bias in your data or if it's some phenomenon that you couldn't explain up to now, but now you have an explanation.

One example that we had, we had very contradictory results in terms of the relationship of returnees to their families, especially in the Western African context, where other studies showed that there is often a reluctance to return because they have fear of reactions of the family. And our data didn't show this at all. And then, after talking to experts in the country, they told us yes, because those persons that are returning with the programme are probably those that have better relationships to their families and the other ones are not taking part in the programme at all. This then explained the self-selection into our programme, which helped us to understand our data and to not make generalizations.

I think the really most important point is to be able to reflect critically your own data, to really reflect on who are the persons whom you are not reaching with your study. Especially due to the sensitivity of the topic, return and reintegration, especially in some regions, it's very important to don't make generalizations of things that you cannot actually do because your data doesn't allow for that.

HOW DO WE MAKE COMPARISONS WHEN DETERMINING REINTEGRATION OUTCOMES?

CREATING COUNTERFACTUALS IN THE REINTEGRATION CONTEXT

Let's recap a bit about comparisons from Modules 2 and 3:

- Evaluations make comparisons to determine what impacts have occurred and seek to establish whether the intervention was the cause of those changes.
- There are various strategies for creating counterfactuals, including comparison groups and quasi-experimental methods such as matching.
- The best estimate of a counterfactual comes from finding a group that is as similar as possible to those receiving the treatment (the intervention or programme assistance being evaluated).

IOM provides non-food item kits, medical assistance, water and sanitation facilities and onward transportation assistance to vulnerable and stranded returnees in Renk. © IOM 2012/Samantha DONKIN

An ideal comparison group for evaluating the impact of return and reintegration programming would likely be returnees who were eligible for programme assistance but who do not receive it. However, return and reintegration programmes often assist migrants in vulnerable situations and the number of individuals who do not depend on the assistance provided (or who can simply afford not to receive it) is likely minimal. There are also ethical issues to be sensitive of in this context.

Based on the experience of the EU-IOM Joint Initiative Programme for Migrant Protection and Reintegration, the individuals who drop out from the programme are likely less vulnerable than the average; moreover, they tend to be difficult to find – for example, if they do not wish to be involved with the programme, or disappear shortly after return. They are not always impossible to locate, but in practice this can be difficult or expensive, and other strategies are often used instead.

Choosing an appropriate strategy can be challenging, and it involves taking into consideration:

- The available data
- The budget
- The specific circumstances of the target population
- How beneficiaries are selected (when and how they become eligible to receive assistance)
- The details of programme delivery
- What information outputs are required from the evaluation
- The overall aim of the evaluation.

There is often a compromise required in deciding the best option, based on its strengths and weaknesses and suitability to the context.

For example, it might be sensible in some cases to choose an option that does not give strong proof for causal attribution but allows a clear measurement of the size of the impact. This might be the case if there is already a strong theoretical basis (see page 55) for claiming that the programme activities have at least contributed to the impact, even if they cannot be claimed to be the only cause.

This section contains information and examples exploring how the principles of comparison we have looked at previously are applied in practice in impact evaluations for return and reintegration programmes.

INTERVIEW TRANSCRIPT: MAKING COMPARISONS

In this interview, Andrew Pinney, co-director of Statistics for Sustainable Development, talks about the realities of establishing comparison or "calibration" groups in the IMPACT study.

MAKING COMPARISONS IN AN IMPACT EVALUATION OF A RETURN AND REINTEGRATION PROGRAMME

Interview with Andrew Pinney, Director, Statistics for Sustainable Development

I'm Andrew Pinney. I'm a co-director of Statistics for Sustainable Development, and I focus on the portfolio related to impact evaluations and occasionally targeting as well. So, in recent years we have been involved in reintegration and how to measure reintegration.

One of the big challenges with reintegration measurement is you can't start at a baseline, so in a sense we pick up a process that's already well under way, so in a sense we don't have a baseline. So, already this starting point is really challenging compared to any other evaluation setting that I've been involved in.

In the work we're doing on the Joint Initiative programme, which is a very significantly funded programme by the European Union to help reintegration of returnees when they get back to their country of origin, we were tasked with looking at the IOM standard way of measuring reintregration, which is a set of 29 indicators across three dimensions.

It's a very heterogeneous group that you're applying this index across. They tend to be – the populations we see tend to be more male, tend to be from the 20s through to the 30s, but we have older and we have females. And the range of education attainment before they leave is also wide. So that is a very, very diverse group; much more diverse than I'd say a normal impact evaluation where you're focusing on a geographic space with a geographically focused project and people tend to be similar, in a similar livelihood. This is far from the case.

So, how would you have any sort of comparison group that makes sense in that situation? Where you want to see progress from when they come back from the airport to some sort of point where you think programme support will have matured into the type of improvements and changes in lives for those for those returnees.

We've got a group that is – we don't call it a counterfactual on purpose, because it can never be a counterfactual, because we can't get back to the baseline – and so we came up with this term, "calibration". You know, we want to calibrate the progress or otherwise the lack of progress that these returnees make over those first 12, 18 months against a group, and this is this calibration group of non-migrants. And we do that by matching them on age, sex and educational attainment.

So, the RSS is obviously designed to measure reintegration, and suddenly we decide to apply the RSS to the non-migrants. So, we had to change the word reintegration to integration, so that's not a huge reach.

But the area that was most difficult was the psychosocial questions. There are a number of questions about trauma, psychological issues, nightmares. You can see those questions are very relevant for a recent returnee that's come back from – it's a distressed – I like to think of it as a "distressed return". So these questions are obviously trying to capture that and capture how they recover from those states of distress at baseline as they move forward. So, that is difficult, and we found that we had to write quite a big introduction: "These questions are unusual. You will not have been asked them before. The reason we are asking them is because we are looking at a programme of reintegration, so please try and answer them the best you can…"

So, that is the weakness of the non-migrant calibration group, is that they are responding to these unusual questions.

Was the RSI ever designed to be asked to non-migrants? Absolutely not. So, we are extending the envelope of reasonable application of this tool.

The recent work we've done with IOM has been looking at the use and the value of the RSI in three countries: Ethiopia, Somalia and the Sudan. And these four methods have worked very well in Ethiopia. When I say very well – as expected, where at baseline the returnees are worse off than the non-migrants, and they gradually improve or they improve, so by the time, at endline, quite often the returnees are sort of statistically indistinguishable from the non-migrants. Very much the sort of theory of change being manifest.

And then we look at the Sudan data, and at no point were the non-migrants better off than the returnees, even at baseline. So, we need to unpack why and how the selection of this non-migrant calibration group was so unusual, because we've tried this now twice in Somalia, once in Ethiopia – so three countries, we get that expected pattern where the non-migrants by and large are better off than the returnees or the IDPs at the outset, and then the returnees or IDPs improve and move towards the value of our calibration group, but in the Sudan, this completely didn't happen. In fact, it's the reverse.

So, is it a selection bias? Or is it just that those that migrated were significantly different from everybody else in their community, that they had an entrepreneurial push and, even if they failed, they come back and somehow they are better off, or better able to reintegrate? We hope when we get teams go to the field and start to talk to groups of, if we can find them, these non-migrants and returnees together and say, "you guys seem really different!"

We tried our best to do this one-to-one matching to try and create a calibration group that was as unbiased as we could imagine, without having to go into sort of elements or factors that were affected by migration. And we used a snowball sampling approach, because we weren't able to get to communities easily. We'd actually phone up the returnees that completed the survey – often the endline survey – and said "look, would you mind helping us in nominating a non-migrant yourself? We want you to nominate a non-migrant along the categories I just mentioned: same sex, similar age, similar education."

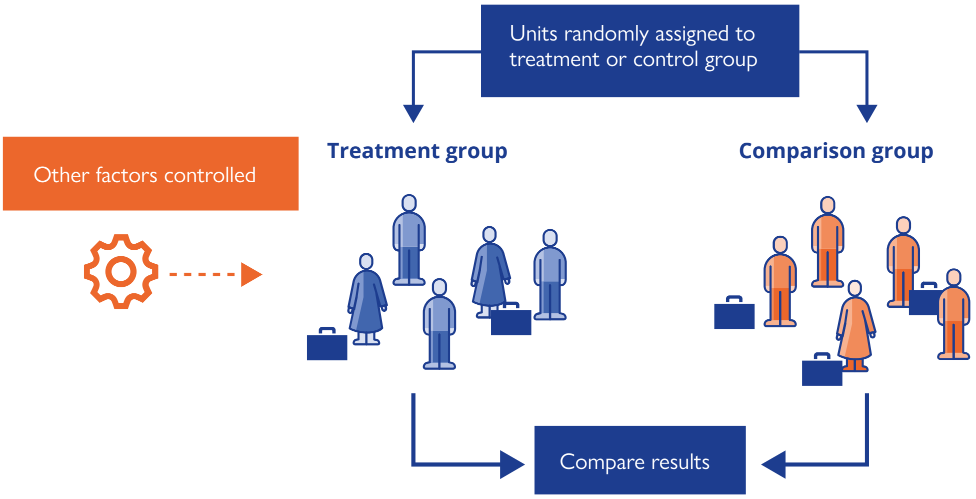
One of the things we learned pretty quickly in the pilot everybody was nominating family members. And we said that – we changed the advice, changed the requirement that they shouldn't be part of the immediate family, because they would be coupled. If a returnee comes back and is in a family, then his or her success is likely to be coupled with the success of the larger household. So, that was an initial error.

But it did work, and we weren't sure whether it was going to work at all, but it has.

CONTROL GROUPS

As we have seen, return and reintegration programmes work in situations where returnees are often dependent on the programme's assistance. The use of a control group – randomly selected from eligible beneficiaries to not receive the treatment (Figure 38) – can be challenging in this context.

FIGURE 38: RANDOMIZED CONTROL GROUP



Ethical issues

It would in most situations be unethical to withhold support from returnees for the sake of creating a control group.

Practical issues

It can be hard to find a pure control group. We have seen that returnees who do not receive programme assistance are usually those who cannot be located. Even if they could be contacted, these returnees have not been randomly selected for the control group; they chose not to participate in the programme and so there are issues of selection bias (see page 86).

For example, returnees who chose not to receive assistance are likely to be less vulnerable than those receiving it.

You could also consider returnees who returned independently of any programme as a control group but similarly, how do you identify them?

Aside from ethical concerns, it would be challenging to find participants for a return and reintegration assistance programme that comes with a random chance of being in the control group and thus not receiving support.

Applicability issues

At the point when an impact evaluation is being conducted, it is often likely that a more useful comparison would be between those different subcomponents of the programme or methods of implementation, rather than trying to discover if the overall programme is better than no assistance.

In this situation, there is unlikely to be a clear control group that could be identified but instead options for comparison groups.

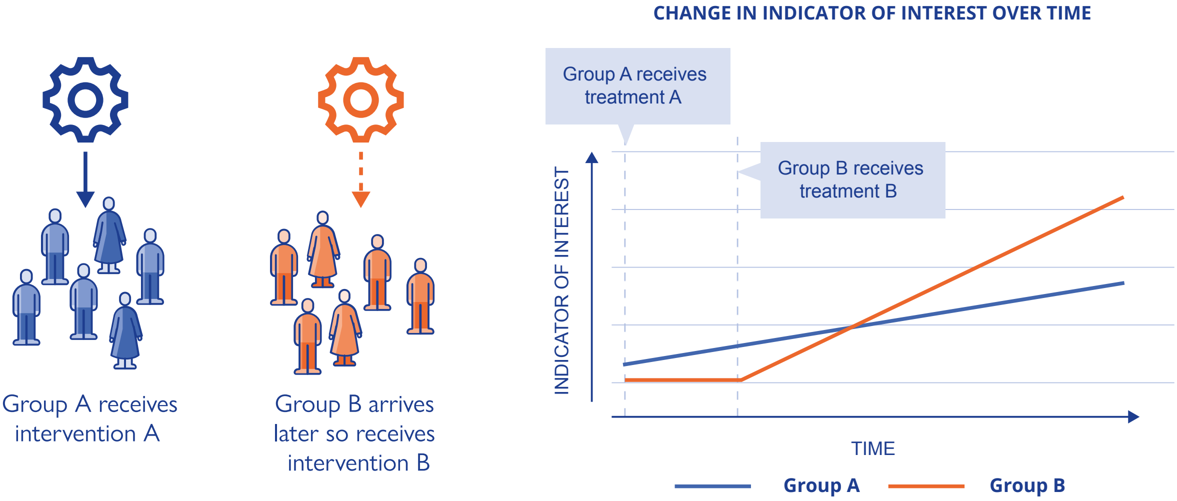
COMPARISON GROUPS

There are various possible options for comparison groups:

- Returnees who received different assistance
- Non-migrants in the communities
- What factors most affect the indices?

Returnees who received different assistance

FIGURE 39: USING RETURNEES WHO HAPPEN TO RECEIVE DIFFERENT ASSISTANCE TO CREATE A COMPARISON GROUP


Group A receives intervention A
Group B arrives later so receives intervention B
CHANGE IN INDICATOR OF INTEREST OVER TIME
Group A receives treatment A
Group B receives treatment B
INDICATOR OF INTEREST
TIME
Group A
Group B

This is useful for evaluations that aim to compare one intervention approach or method to another; for example, to see if a new strategy of implementation has a preferable impact to the standard approach.

While it would be challenging in the implementation of programme interventions to assign different assistance to different groups in a structured and random way, this type of comparison group could be practical if groups came through a different programme or arrived at a different time and thus received different assistance (Figure 39).

These groups could also become emergent during the programme and statistical analysis can be used to try to understand the differences among them.

Things to consider

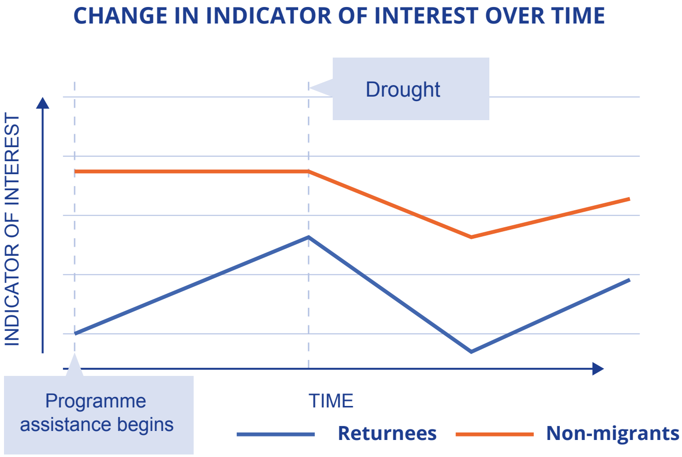
- This method is only applicable if it suits the aims of the evaluation – for example, if the information needed is about whether one approach is better than another.
- It also requires certain practical conditions: if there are different interventions being given within a reasonable time frame of each other and if it is possible to obtain data on the identified group.
- This method is also only applicable if the comparison group can be considered to be similar to the group receiving the treatment. For example, if using a comparison group of returnees who are returning from a different country to the treatment group, then we need to look at whether their characteristics might be fundamentally different; in a situation where returnees from the comparison group country reached their intended destination but those from the treatment group country got stranded in a transit country, then the two groups have had very different experiences and may not be useful to compare.

Non-migrants in the communities

In this case, returnees would be compared with members of the community they have returned to who did not migrate.

This approach can provide a point of comparison to determine if returnees are integrating well, by seeing if returnees are becoming more similar to the host community.

FIGURE 40: EXAMPLE OF COMPARING RETURNEES WITH NON-MIGRANTS



Using non-migrants makes it possible to identify events that impact the whole community and might cause the impact of the treatment on returnees to be incorrectly measured.

For example, as illustrated in Figure 40, if a drought affected the area, the negative impact of this might mask any positive impacts from the treatment. Looking at the difference between returnees and non-migrants rather than the overall score would help you isolate the impact of the treatment from the impact of the drought.

Things to consider

- This comparison could be helpful for return and reintegration programmes that include interventions which target the whole community affected by returnees.
- Using non-migrants as a comparison group may be of limited use for attribution; for example, returnees might make significant progress in outcomes, but comparing with non-migrants won't help rule out if the improvement was due to other causes specifically affecting returnees.
- It can be difficult to determine how appropriate non-migrants are as a comparison. There may be characteristics or factors that differ between returnees and non-migrants that have a significant effect on outcomes that are not well understood or accounted for. For example, non-migrants have not had the migration experience and are not going through a reintegration process. This makes them clearly different from returnees in a way that can raise concerns about their use as a comparison group. In some cases, even when events impact the whole community, non-migrants may be more resilient to them.

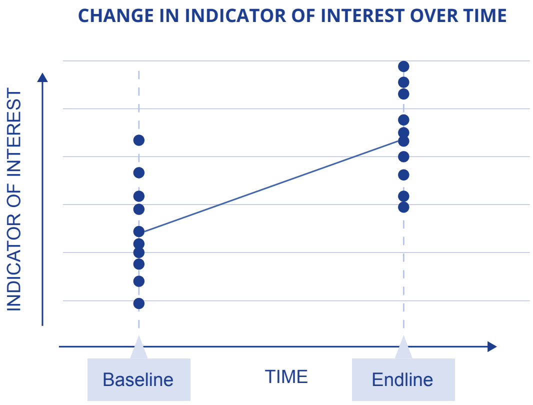
What factors most affect the indices?

Imagine you have collected baseline and endline data from returnees who have received programme assistance and calculated a reintegration index score for each of them. They won't all be the same. The results might look like Figure 41.

Some of the variation will be expected differences between individuals, and we look at the averages to estimate the effect. However, there may also be patterns that are correlated with certain factors.

The programme activities include a range of assistance and interventions. Returnees will have received different elements of this assistance. Some got psychosocial counselling, some didn't, some got stipends, business grants and so on. Not everybody got everything.

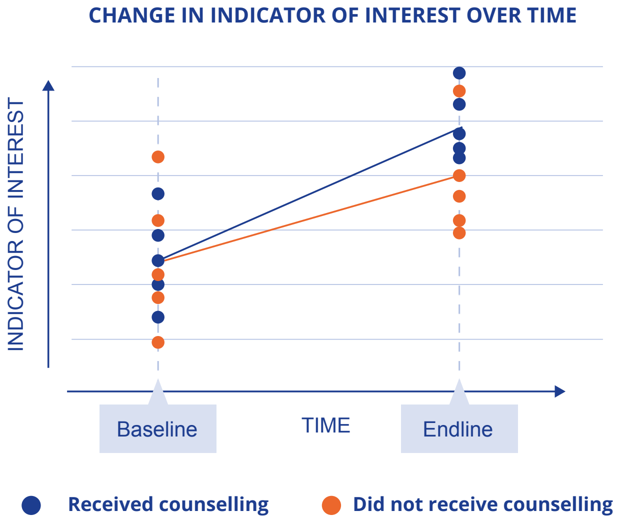
FIGURE 41: MEASURING THE CHANGE IN AN INDICATOR OVER TIME



Assistance was assigned based on certain criteria, this could be needs based and therefore not random; or it could be that it wasn't possible logistically to implement all interventions in all locations, all of the time, in which case we may be able to consider this a "natural experiment".

You could make comparisons by looking at the outcomes based on who received what. This effectively creates a comparison group. For example, we can separate the data into groups for those who received psychosocial counselling and those who didn't (Figure 42).

FIGURE 42: EXAMPLE OF COMPARING THE CHANGE IN AN INDICATOR OVER TIME FOR THOSE WHO RECEIVED A PARTICULAR INTERVENTION



Looking at this, we have an indication that the counselling may be having a positive impact on returnees. We could try this for each of the different possible interventions to see which have a strong correlation. However, there would be more thorough analysis needed:

- Are the same returnees receiving other assistance that might explain the changes? You would need to look at the other possible comparisons within the data set, as well as considering external confounding factors.
- What are the characteristics of the returnees in each group? Perhaps the returnees who didn't receive it were in a specific location that had some other circumstances that caused negative impacts.

Of course, this is a simplified example. In real life, there would be many interventions and other factors, and the amount of data would be considerably higher than what is shown in the plots.

Things to consider

- This method could, to an extent, be used in a similar way to regular comparison groups, although it is less deliberate and the interpretation of the data is more complicated.
- Although insufficient for proper attribution, this approach could be suitable for assessing the relative contributions of different aspects of a programme.

MEASURING IMPACT WITHOUT COMPARISON GROUPS

One option is to accept that there is no way to establish a comparison group that is similar in all ways except for receiving the intervention and focus on making before/after comparisons. This is a big compromise in terms of what evidence the evaluation can produce for attribution, and some people would argue that this is not an impact evaluation. However, depending on the information requirements and other information available, it may be suitable and/or necessary.

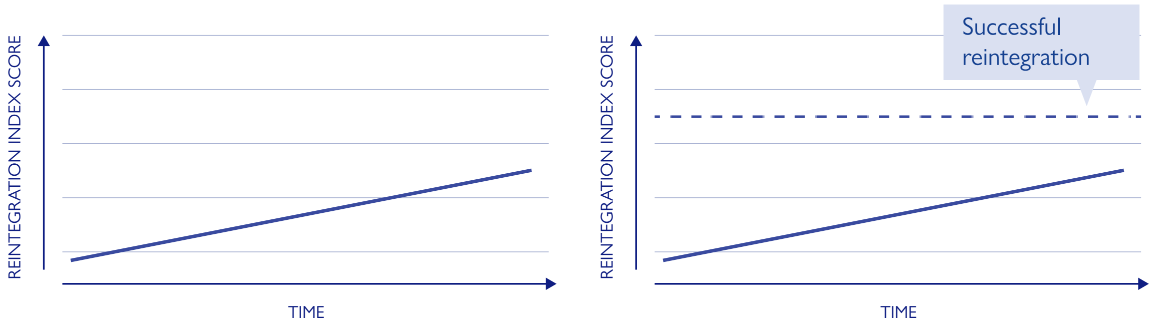
If you have a strong programme theory and understanding of other potential causes of impact, this could be a good option. However, it is hard to control for other contaminating factors. This approach could also be combined with qualitative methods (see Module 5).

Measuring impact using indices

The indices that we looked at in the previous section are usually determined by a mix of expert input and statistical evidence to provide a measure of reintegration. We saw that some of these indices and ways of measuring include thresholds, also defined by these expert and statistical insights, which give indications for how to interpret scores.

While this may resemble a simple before/after comparison, comparing against a normative standard provides a reference point so we can understand the change in relation to a set target. For example, while seeing an improvement in reintegration is useful, it means something very different depending on the distance from, and progress made towards, a threshold target defining successful reintegration:

REINTEGRATION INDEX SCORE OVER TIME



Care must be taken when adopting this approach to ensure that the results are correctly interpreted.

If we were monitoring the progress of reintegration in a treatment group alone and we saw no change over time, then without a comparison group, the likely interpretation would be that the treatment was not working. However, if there was a comparison group not benefiting from that treatment, it might be apparent that respondents in this group had experienced a decline in their reintegration score (while the treated had remained unchanged). This would lead to a completely different interpretation of the impact of the programme intervention.

Some indices do not include thresholds and, with these, the approach would be to look at the trend over time, or to combine the use of an index score with other methods of comparison.

Things to consider

- There is discussion among experts about the validity of using thresholds, and some suggest that they are arbitrary. However, the main focus is that you can not only understand if progress is being made, but also see this with the context of a target.
- The main concern with comparing to a threshold instead of a comparison group is that this method cannot be used to establish attribution, or even contribution. The comparison can provide information about changes and progress towards a goal, but not about what would have happened without the treatment.

INTERVIEW TRANSCRIPT: IMPACT EVALUATION IN RETURN AND REINTEGRATION CONTEXTS

In this interview, Jasper Tjaden talks about the realities of conducting an impact evaluation in return and reintegration context and gives advice on creating control or comparison groups and planning data collection.

IMPACT EVALUATION FOR RETURN AND REINTEGRATION PROGRAMMES

Interview with Professor Jasper Tjaden, Professor for Applied Social Research and Public Policy, University of Potsdam

Finding a counterfactual

The key in designing a good impact evaluation study is, of course, finding that counterfactual – finding that control group that you want to compare your intervention groups… the people participating in the project or programme, can compare to. Because you want to basically know, "what if the people that I am supporting, or implementing projects with, what if they hadn't benefited from that project?". That's what we call the counterfactual. And if those people – if the difference in the outcomes between people that have benefited from the project versus those that have not is positive, we can attribute this as the true impact of a project or a programme, to put it very simply.

Randomized controlled trials

One approach that's often considered the gold standard is the randomized controlled trial or randomized controlled experiment where you have a group of people, you randomly choose the ones that benefit from your programme or your project and some that do not, and then after the programme is over you compare the two. Well, what are the challenges that apply here?

Well, first of all, randomization often is not possible in the real world. If you work with implementing agencies, it could simply be not practical to randomly allocate beneficiaries – people – to different programmes, maybe because they've already enjoyed the programmes or already participated in the programme. So, you can't randomize after the fact.

The other issue with randomization often is ethical; many partnering organizations have ethical concerns that by simply throwing the dice, deciding who benefits from something and who doesn't is unethical. This ethical concern also applies to the area of return and reintegration, where it would be difficult in practice to randomly allocate reintegration assistance to certain returnees but not to others. So that's a fundamental challenge with RCTs – randomized controlled trials – in practice

Other ways to create a comparison group

One way to avoid this issue of randomization is, for example, to have what's called a "stepped-wedge" trial or a "phased approach" where certain beneficiaries in the beginning of the programme are randomly assigned to receiving the project, then after the project is over. the ones who were in the control group in the first stage – so did not receive anything – are then eligible to receive the project. This makes sure that everyone that you are in touch with actually does benefit from the project, so there is not that ethical concern of "do people miss out and others win?" Everyone gets it, but they get it at different times. This of course works if there is not a real-time sensitivity, which again is difficult with returnees because probably the sooner they receive reintegration assistance, the better, and a difference of half a year or so, or a year, can really make a big difference here.

Now, with pure randomization, finding the control group is easy. You have 100 people, you flip the coin, 50 of them get it, 50 don't, let's say – a simple example. Well, if randomization… in the practice, to give you a practical example of this, we did one RCT in Senegal, measuring the impact of an information campaign for potential migrants. So, we went to several districts in Dakar, and we randomly invited people on the street, through random routes that we walked through those districts, we randomly, on the spot, invited potential migrants to attend either a movie screaming about migration or a movie screening about something completely unrelated to migration – this is our control group. That was sort of a very neat design, because you have this pure randomization of people into treatment.

Now, in other instances, this isn't possible. To give you a practical example from Guinea, one study I was involved in was again on measuring the impact of an information campaign targeting potential migrants. We were working with a real campaign that we didn't have any control over, and this campaign was a mobile cinema that basically was a caravan going from village to village, showing movies, conducting a theatre, conducting discussions with community members. And in the perfect world, we would have a hundred villages, and we would randomly choose which villages the caravan goes to. This would be a good randomization approach. This was impossible, because the people in the campaign basically said, "are you crazy? No, we're not going to do this. We're not going to go this crazy route. Some of these villages aren't even accessible by road, you know. It wouldn't make any sense logistically to go there, and also with some villages, we don't have agreement by the local government that we can visit these villages", so on and so on. And these sort of practical impediments to randomization are very common, regardless of the area.

So, what we did is we followed the caravan to the villages that they went to. We collected data there, so now we had the challenge of actually generating a suitable control group. What we did is basically choose villages that are fairly close to the villages where the project was, but not too close – if there's too much contact, because we don't want that overlap. So, they are far away enough but they're very similar: same region, same ethnic composition and so forth. And we collected data there, and then we compared the treatment villages versus those kind of hand-picked control villages, and we made certain adjustments to make sure that they're actually similar. And then we applied what's called a "difference in difference" design, where we compared both of these villages before the project and after the project and sort of put the two differences between the groups in relation to each other and that was our main impact. There are certain assumptions that you have to make when using this approach, but it's one way to work around this issue of not having perfect randomization.

The next best approach, if a full experiment – an RCT – is not possible, are often quasi-experimental approaches, and there are various designs here. However, the main challenge there is they're often quite technical and really depend on the context that you're in. There are difference in difference designs, there are instrumental variable designs, regression discontinuity designs… they're all great, but they really depend on the specific context. What does this mean for implementers? Often you're interested, not in one project at one time, but you're interested in actually applying a methodology in various locations, at various times, maybe in various countries. This is really hard to do with these types of methodologies, because they really depend on context a lot, and to use these methodologies, a lot of – let's say – requirements have to be met. So this question of "what is really the best control group here?" is a tricky one for return and reintegration programmes.

Obviously, returnees, you can compare them to other returnees, maybe, that haven't benefited from reintegration, or they may return from different countries or whatever. Or you compare them to people who didn't leave, but that is problematic because they're not very similar.

Data collection

Another more general challenge with implementing these approaches is that, next to the demand in technical expertise and actually conducting them, is resources needed to implement data collection. So, the actual research design, once you have people who know what they're doing, is not that complicated. However, the data collection often is quite costly and labour intensive, because you're not only collecting data with people that actually are in your programmes and projects – you may be already in touch with – but there's also data collection with people, beneficiaries, that do not benefit from your projects and function as a kind of control or comparison group. It's often harder to collect data on these people and it basically doubles your data-collection efforts.

Well, let's say 80 per cent of the work is collecting good data, and that's very tricky. And it doesn't matter whether you're doing a fancy impact evaluation study or other types of survey-based but, often, the environments that IOM works in and the environments where return and reintegration takes place are often low-income places – difficult places to collect data in. And that's where the big challenge is: collecting high-quality data. It requires a lot of knowledge of the local settings; it requires a lot of resources; it requires good teams on the ground and it requires quality checks. So, I think implementing a really thorough data-collection plan and allocating enough resources to data collection is really key.

Sometimes impact evaluations are "top heavy", meaning a lot of money goes to principal investigators that sit somewhere in universities – like myself! A lot of money should go to people collecting data on the ground who really know what they're doing. I think that's probably the biggest recommendation I would give – practical recommendation for impact evaluations. You need to have a good very good survey firm, or a lot of local capacity to collect data properly.

So, a couple of things that can go wrong especially during data collection. First of all, you're collecting data on some of the people participating in your programme, but only certain types and not others. It's a problem because then when you later talk about impacts of your project, you're only talking about the impact on certain types of people that participated in the study and those are often not representative of your larger pool of beneficiaries that you're reaching.

Another obstacle is the questionnaire is way too long, people get tired of answering it, they stop the interviews midway, they don't even want to take the interview in the first place. And a related issue is that, for impact evaluations, you need to often have data for several moments in time. It could be that people answer first time you reach out to them, but then you reach out three months later and you can't find them anymore, or they don't want to be part of your study anymore. This is called attrition and it's a big problem for impact evaluations. That's something that can go wrong.

Many things can go wrong during data collection, for example, people go out and actually have a list of people that they're supposed to interview, but they interview other people for some reason. Because they can't find them and then they just interview other people. All of this can happen, all of this can happen!

Measurement issues. Measurement error is when the question that you put in the questionnaire, you later realize it doesn't really accurately reflect what you're trying to measure. Because maybe the questionnaire design process was too quick and there was a lot of time pressure, and then you thought, "oh, this would be a great question". You haven't really validated it, cross-checked and all of that; it was just rushed, and then later you want to measure the impact and you realize that the questions that you actually have in the questionnaire don't really get to what you consider the impact would be of your project. So issues around measurement.

Then, of course, there are issues later more on the analysis side, where people use the wrong methods to analyse the data they collected, but this gets very technical and that's something that is easily solvable. You can just change the methods with data collection. You can't change the data you have. You can't go back to the field and collect new data. You could, but it's very costly, so you know the most important thing is designing a good study and collecting great data.

KNOWLEDGE CHECK

Match the scenarios with the most suitable methods for comparison:

Programme activities are changing with time or by location, either by design or randomly, and information is sought on the difference between new and old interventions.

The priority is on measuring the extent of the change to reintegration score, and it would be possible to supplement the findings with qualitative methods.

There is already a strong evidence base for the current understanding of the causal mechanisms of the programme activities. It is important to control for events that might affect the whole community in a location.

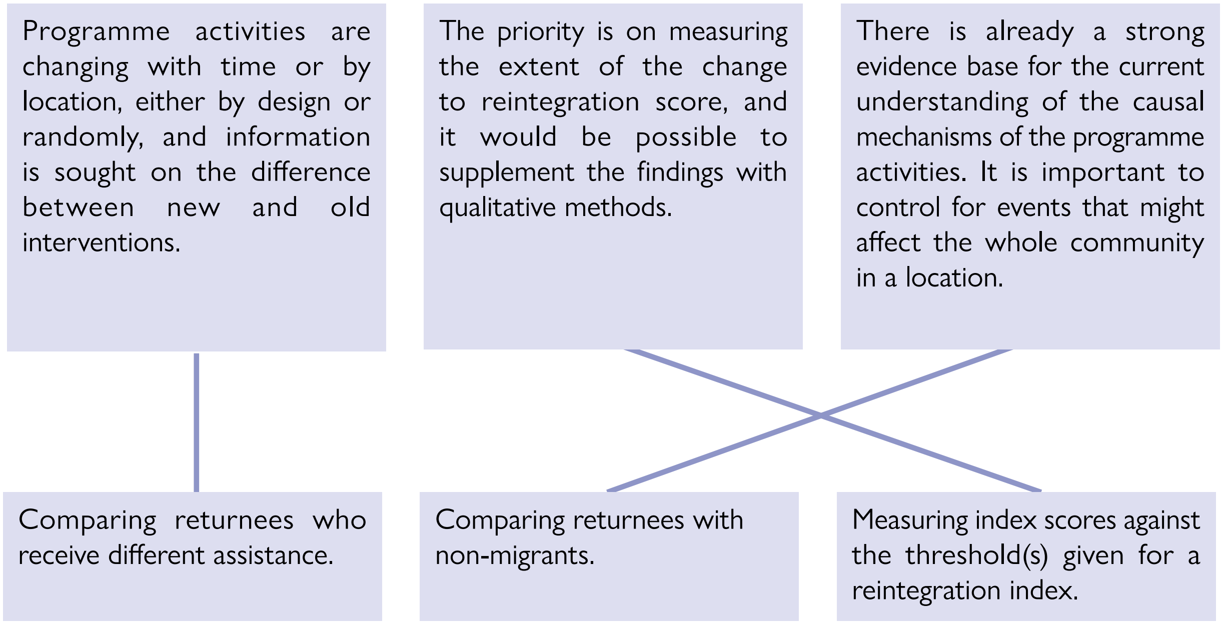
Comparing returnees who receive different assistance.

Comparing returnees with non-migrants.

Measuring index scores against the threshold(s) given for a reintegration index.

KNOWLEDGE CHECK ANSWER

Match the scenarios with the most suitable methods for comparison:



Programme activities are changing with time or by location, either by design or randomly, and information is sought on the difference between new and old interventions.

The priority is on measuring the extent of the change to reintegration score, and it would be possible to supplement the findings with qualitative methods.

There is already a strong evidence base for the current understanding of the causal mechanisms of the programme activities. It is important to control for events that might affect the whole community in a location.

Comparing returnees who receive different assistance.

Comparing returnees with non-migrants.

Measuring index scores against the threshold(s) given for a reintegration index.

NATURAL EXPERIMENTS

Let's recap what we learned in Module 3.

Question 1

What is a natural experiment? Choose the definition you think is correct.

- □ Using comparison groups based on the environmental and agricultural factors that affect returnees.
- □ Taking advantage of unplanned changes to make comparisons that might not otherwise be possible.
- □ Comparing the impact of the treatment group to a group that receives no programme intervention.

The correct answer is below. You may wish to review the Module 3 section about natural experiments on page 95.

Question 1 answer

What is a natural experiment? Choose the definition you think is correct.

- □ Using comparison groups based on the environmental and agricultural factors that affect returnees.
- ☑ Taking advantage of unplanned changes to make comparisons that might not otherwise be possible.
- □ Comparing the impact of the treatment group to a group that receives no programme intervention.

We can see an example of natural experiments being used in the IMPACT study of the EU-IOM Joint Initiative. The evaluation design combined multiple approaches, including planned natural experiments to supplement the main quantitative methods.

Natural experiment: delays to assistance

IMPACT will assess the impact of delays in providing in-kind assistance to returnees. Some returnees have unfortunately waited longer than others to receive the planned assistance.

By collecting and analysing data from returnees who have received assistance and comparing it with those who, due to delays, have not received assistance, a natural experiment is possible. The delay naturally creates an effective comparison group for a with/without comparison. This natural experiment can:

- Clarify the impact of the delays.
- Measure the impact of the assistance.

Natural experiment: Mobile Money procurement strategy pilot

Starting in 2019, the Sudan country programme began piloting a new approach to providing assistance.

To reduce delays and engage returnees in the procurement process, the programme switched from their previous method of procuring in-kind assistance through vendors to enlisting the returnees to obtain quotations locally and transferring the money to the selected company by mobile phone transfer. There was also later a shift to providing assistance in cash form.

The IMPACT study aims to frame a natural experiment based on a comparison of returnees receiving reintegration assistance before and after the transition to the new procurement process – allowing an understanding of the impacts of the new process compared to the old one, which would likely be useful to IOM and its stakeholders in the Sudan, the rest of the EU-IOM Joint Initiative (HoA) programme and return and reintegration programmes elsewhere.

Question 2

Think about the strengths and weaknesses of the natural experiments you just read about. Make some notes, then see our suggestions below.

Ahmed, a returnee from Libya, was in detention for two years before he decided to return home. "I was too proud to come back after having failed to reach Europe, but my friends and family were very supportive. I now have a small business and hope to grow it." © IOM 2020/Muse MOHAMMED

Question 2 suggested solution

Strengths:

The natural experiments allow comparisons that would otherwise not be practical (e.g. where there is no identifiable comparison or control group).

Weaknesses:

There are potentially some significant differences between the comparison group and those receiving the treatment:

- Returnees recruited earlier into the programme, who received assistance using the old procurement method, could have different characteristics to the returnees recruited later, who experience the new procurement method. This could happen if the programme changed its recruitment criteria.
- The new procurement method was introduced due to challenging conditions in the Sudan linked to the Sudanese revolution, so the situation is quite different for the returnees before and after the change to the new method.
- Delays and slower roll-out of new implementations could be correlated with other factors that affect reintegration.
- Those who have not received assistance are not likely to be motivated to cooperate with data collection.

INTERVIEW TRANSCRIPT: A NATURAL EXPERIMENT

In this interview, Andrew Pinney, director of Statistics for Sustainable Development, shares an example of a natural experiment that was incorporated into the IMPACT study.

USING A NATURAL EXPERIMENT IN AN IMPACT EVALUATION OF A RETURN AND REINTEGRATION PROGRAMME

Interview with Andrew Pinney, Director, Statistics for Sustainable Development

We would expect returnees to probably get more integrated with time, unless they're having such a lousy time they decide to get up and go – you know, remigrate again. And that inevitably will happen for some. So, there is a time effect.

So, one of the more nuanced questions in the impact evaluation, is not "do we see progress for returnees?" but the question is, "do we see progress with returnees that is faster than we might expect if they weren't supported?"

Now it happens – not by design, but by sort of programme realities – in Ethiopia, we now have quite a big group, about 1,600 returnees, that have come through the programme that never got a micro-business grant for various reasons, or any type of significant support other than that initial stipend when they arrived at the airport. And there's 2,500 that did get the support. Now, out of the three countries, we only have that in Ethiopia. But suddenly this has become an emergent calibration group or contrast group. Now, these are a much more interesting group because they're all returnees, and suddenly we've got a with and without.

The question there will be: if everybody's becoming integrated at a certain point, the ones getting the micro-business support or the education support, did they get to a sort of stable level of reintegration – as measured by, maybe, our calibration group? Did they get there earlier? Did they get there six months earlier? Did they get there a year earlier? Or did it actually make no difference? And what we're seeing, very reassuringly, in the case so far of Ethiopia, when we compare the treated to the untreated, we see their baseline RSI score numerically exactly the same, or so close, totally indistinguishable, and then by endline the treated – the ones that did get the assistance – are very significantly better than the untreated returnees – the returnees that did not get assistance – so they are both improving, but those who get the micro-business support improve much faster.

QUIZ

This quiz will check your understanding of the topics covered in this module. There are seven questions. You must get a score of at least five out of seven to pass.

1. Which of the following statements about impact evaluations for return and reintegration programmes is true? Select all the answers that apply.

   □ It is impossible to use a control group.
   □ It is often necessary to make compromises to find a sensible solution.
   □ Unstable circumstances for returnees can create difficulties with data collection.
   □ Measuring reintegration involves lots of very long survey questionnaires because there are so many components.

2. Select the correct option to fill in the gap:

   It can be difficult to conduct a longitudinal study in return and reintegration contexts because returnees often _____________.

   □ Relocate.
   □ Receive assistance from other programmes.
   □ Have debt.

3. Select the correct option to fill in the gap:

   The work of other programmes and organizations in the same location as the programme being evaluated can lead to __________________.

   □ Negative impact.
   □ Unequal provision of assistance.
   □ Contamination.

4. Select the correct option to fill in the gap:

   A(n) __________________ is an indicator derived from a combination of multiple indicators.

   □ Proxy indicator.
   □ Instrumental variable.
   □ Index.
   □ Normative threshold.

5. Select the correct option to fill in the gap:

   Some indices include a _____________ which defines how different scores can be interpreted.

   □ Programme theory.
   □ Composite indicator.
   □ Normative threshold.

6. Which of the following best describes how an approach can be chosen for making comparisons in impact evaluations for return and reintegration programmes? Select one answer.

   □ It is always best to use returnees who did not join the programme as a control group if it is possible.
   □ Attribution is impossible without a control group.
   □ The most suitable approach for comparisons depends on the evaluation aims.
   □ Using non-migrants as a comparison group does not provide helpful information.

7. Which of the following are common challenges of identifying comparison groups for impact evaluations in return and reintegration contexts? Select all that apply.

   □ It is often difficult and expensive to locate returnees who did not receive programme assistance for use as a comparison group.
   □ It is impossible to measure impact without a control group.
   □ It can be logistically difficult to find or establish groups that will receive different interventions for comparison.
   □ There can be many factors that have a significant effect on reintegration success that can be difficult to predict or control for.

QUIZ ANSWERS

This quiz will check your understanding of the topics covered in this module. There are seven questions. You must get a score of at least five out of seven to pass.

1. Which of the following statements about impact evaluations for return and reintegration programmes is true? Select all the answers that apply.

   □ It is impossible to use a control group.

   ☑ It is often necessary to make compromises to find a sensible solution.

   ☑ Unstable circumstances for returnees can create difficulties with data collection.

   □ Measuring reintegration involves lots of very long survey questionnaires because there are so many components.

2. Select the correct option to fill in the gap:

   It can be difficult to conduct a longitudinal study in return and reintegration contexts because returnees often _____________.

   ☑ Relocate.

   □ Receive assistance from other programmes.

   □ Have debt.

3. Select the correct option to fill in the gap:

   The work of other programmes and organizations in the same location as the programme being evaluated can lead to ___________________.

   □ Negative impact.

   □ Unequal provision of assistance.

   ☑ Contamination.

4. Select the correct option to fill in the gap:

   A(n) ___________________ is an indicator derived from a combination of multiple indicators.

   □ Proxy indicator.

   □ Instrumental variable.

   ☑ Index.

   □ Normative threshold.

5. Select the correct option to fill in the gap:

   Some indices include a ______________ which defines how different scores can be interpreted.

   □ Programme theory.
   □ Composite indicator.
   ☑ Normative threshold.

6. Which of the following best describes how an approach can be chosen for making comparisons in impact evaluations for return and reintegration programmes? Select one answer.

   □ It is always best to use returnees who did not join the programme as a control group if it is possible.
   □ Attribution is impossible without a control group.
   ☑ The most suitable approach for comparisons depends on the evaluation aims.
   □ Using non-migrants as a comparison group does not provide helpful information.

7. Which of the following are common challenges of identifying comparison groups for impact evaluations in return and reintegration contexts? Select all that apply.

   ☑ It is often difficult and expensive to locate returnees who did not receive programme assistance for use as a comparison group.
   □ It is impossible to measure impact without a control group.
   ☑ It can be logistically difficult to find or establish groups that will receive different interventions for comparison.
   ☑ There can be many factors that have a significant effect on reintegration success that can be difficult to predict or control for.

SUMMARY

In this module, we have seen that:

1. If an impact evaluation intends to measure "reintegration", careful thought needs to be given to deciding how reintegration will be defined and measured.

2. There are several existing ways of measuring reintegration that have been developed for specific purposes and which may be suitable to use or adapt for an impact evaluation.

3. Data collection for impact evaluation of return and reintegration programmes can present challenges in terms of timing, feasibility of acquiring the intended data and contextual factors that may bias or contaminate the data collected.

4. In a return and reintegration context, finding or creating a suitable comparison group can be challenging, and sometimes compromises need to be made between feasibility, ethics and rigorous evaluation.

Assisted voluntary return and reintegration is among the services offered through Migration Response Centres in the Sudan. This makes it possible for migrants to voluntarily return to their communities of origin with support from IOM which ensures that they travel safely and with dignity. © IOM 2021/Muse MOHAMMED

MODULE 5: QUALITATIVE METHODS

MODULE 5: QUALITATIVE METHODS

INTRODUCTION

This module provides an overview of qualitative impact evaluation methods and explains how quantitative and qualitative methods can be combined to increase the robustness of the evaluation.

OUTCOMES

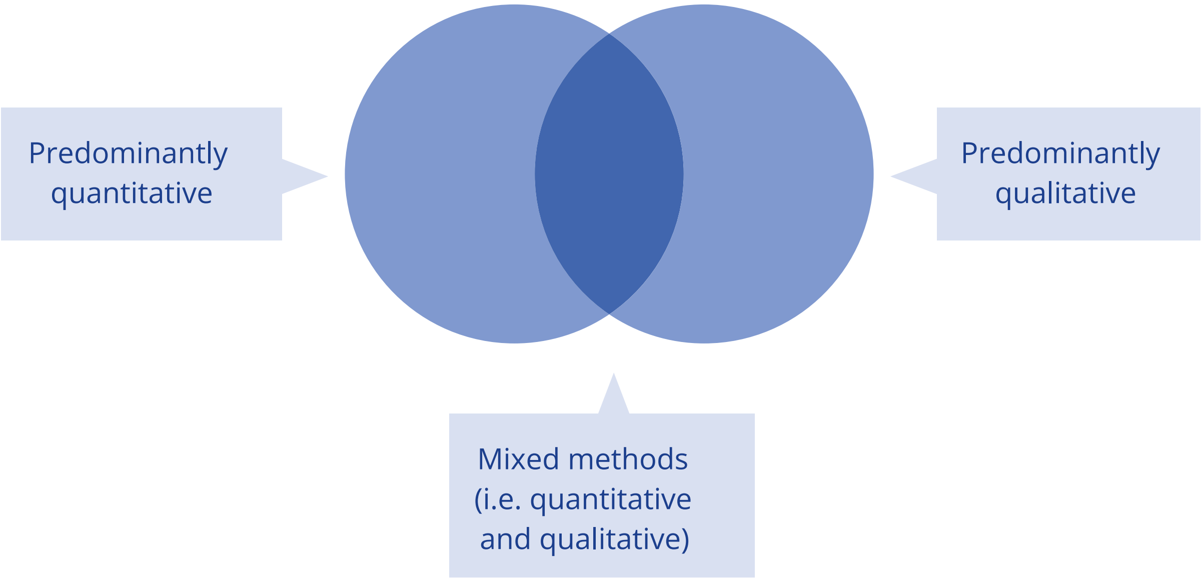
At the end of this module, trainees will be able to:

- Outline the role of qualitative methods in impact evaluation.
- Explain the potential benefits of using qualitative methods to complement and enhance quantitative approaches at different stages in the design and implementation of impact evaluation.
- Give examples of how qualitative and quantitative data can be combined effectively in the context of impact evaluation for return and reintegration programmes.

INTRODUCTION

This course has largely focused on quantitative methods in impact evaluations, emphasizing the need to measure impact. However, qualitative methods are also used in, and may be a large component of, impact evaluations.

Impact evaluation approaches can be:


Predominantly quantitative
Predominantly qualitative
Mixed methods (i.e. quantitative and qualitative)

Depending on the characteristics of the evaluation, a combined approach can be very effective. Most impact evaluations use mixed-methods to some extent.

Find out more

See the Better Evaluation website to find out more about impact evaluation approaches.

Web page: *www.betterevaluation.org/en/approaches*

WHAT ARE QUALITATIVE METHODS?

Qualitative methods facilitate the study of issues in depth and detail, based on data that are descriptive in nature, rather than data that can be measured or counted. They can produce important insights that are not easily quantified, including exploring the social, emotional and cultural drivers that can help explain why changes have occurred. This is very relevant to the aims of impact evaluation.

Qualitative and quantitative methods contrast in the way that data collection is planned and conducted.

- Quantitative methods require defining exactly what is to be measured ahead of time and designing tools – such as questionnaires – with constraints so that the answers are given in a standardized way.
- Qualitative methods, on the other hand, tend to avoid this predetermination and will use open-ended questions and discussions to invite varied and potentially unexpected responses.

Quantitative and qualitative questions

Here is what a question about eviction might look like on a quantitative and a qualitative survey:

Quantitative

- **Were you evicted in the past 12 months?** Yes / no.
- **If yes, for what reasons were you evicted?**
    - Inability to keep up rent payments
    - Disagreement with landlord
    - Breach of policy
    - Other

Qualitative

- **Over the past year how would you describe the stability of your residential location?**

Text box – the respondent puts an answer in their own words.

This means that qualitative methods open the possibility for deepening our understanding of impacts and impact mechanisms, with limited constraints imposed on the way data are collected.

Information obtained from using qualitative methods is commonly used to:

- Extract themes, patterns and concepts;
- Gain insights and understandings of what happens, how it happens and why it happens.

Qualitative methods may also be:

- Used to help design quantitative tools;
- Embedded within quantitative tools;
- Transformed into quantitative data to be utilized in analysis.

We will discuss these uses in further detail later in this module.

TYPES OF QUALITATIVE METHODS

In practice, the term "qualitative methods" is used to refer to two aspects of impact evaluation work: data-collection methods and data-analysis methods.

DATA-COLLECTION METHODS: WAYS TO GATHER THE DATA

Interviews

A conversation between an interviewer and a respondent, in which the interviewer seeks to obtain information by asking questions. This is a good way to gain an understanding of individuals' opinions, experiences and perspectives.

Interviews can be quantitative or qualitative; qualitative interviews are less structured, tending towards open-ended questions rather than the rigid questionnaires with preset answers used in quantitative surveys.

Pros of qualitative interviews:

- They can obtain detailed and unanticipated responses.
- They allow for follow-up questions in response to previous answers much more easily than in quantitative interviews.
- It is possible to target respondents with specific, relevant knowledge and insights.

Cons of qualitative interviews:

- They have potential for bias, for example in the selection of respondents, or through the interviewer unintentionally influencing the responses.
- It is time-consuming to conduct each individual qualitative interview.
- Producing transcripts and translations for qualitative interviews is both time-consuming and difficult, requiring skilled staff and making it a potentially very expensive option.
- Conducting a qualitative interview requires skills, experience and also a good understanding of the objectives of the evaluation and the programme's theory of change.

Example types of qualitative interviews:

Semi-structured interviews (sometimes also referred to as key informant interviews)

- An interview in which there is a list of topics or questions to cover, but which allows for a spontaneous, conversational approach. They have a loose agenda, but interviewers have the freedom to improvize, for example to probe further into interesting points raised by the respondent.
- Interviewees are usually deliberately selected according to certain criteria, with the aim of getting specific contexts or viewpoints; for example, interviewing a woman in a community leadership position.

Life story interviews

- An interview in which the respondent is invited to share their "life story", however they choose to present it – reflective of what they remember and what they want others to know.
- This is a very open style of interview, suited to gaining a deeper understanding of respondents' perspectives and context, rather than information on a particular question or topic. Life stories will vary significantly in their tone, detail, veracity and focus.

Qualitative data in quantitative surveys

Very often, qualitative questions are included in quantitative surveys, generally as follow-up questions. These would generally be open-ended, so that respondents can give answers in their own words rather than select from predetermined options. For example, a survey could ask a returnee how integrated they feel (using a scale from "not at all" to "very much"), followed by a question on why they gave this score, as a free-text answer.

Pros of qualitative data in quantitative surveys:

- Depending on the mode of delivery, these can be used to collect data from a larger number of respondents than it would be possible to hold one-on-one interviews with.
- Obtaining multiple answers to the same, defined questions may be useful for identifying patterns and differences.

Cons of qualitative data in quantitative surveys:

- Qualitative survey data are more complex and time-consuming to analyse compared to quantitative.
- Certain formats – such as digital – may cause certain respondents to be less likely to participate. For example, online surveys would be challenging for partially literate respondents, and phone surveys may be difficult to arrange with respondents with only limited access to a phone.
- As with interviews, it is possible for the enumerator to unintentionally influence the answers given.
- Inflexible written questions could inadvertently limit answers.

Group discussions

An interview-like discussion with a group, rather than one-on-one. As with interviews, these can vary in the amount of structure given.

Pros of group discussions:

- Group discussions can be useful when we want to capture the diversity or understand the consensus of community-level knowledge on a topic. For example, if we want to know about the functioning of government services to returnees, we could ask individuals, but it would be more efficient and be easier to capture the diversity if conducted as a group conversation.
- There may be information that does not vary within a community but requires discussion to clarify. For example, if we want to look at the influence of the agricultural cropping-calendar on availability of jobs (as farm labourers) for returnees in rural areas, this information would most effectively be collected from a group discussion rather than individual interviews.
- The interactions between group members allow the diversity of viewpoints to emerge and can produce additional information; for example, through the discussions that occur when there is a disagreement.

Cons of group discussions:

- If the composition of the group is not right, some participants may be less open, sincere or willing to share certain information with other people present.
- Requires good planning and facilitation to be successful.

Example type of group interviews:

Focus group discussions

- A group discussion with the aim of focusing on specific issues.
- The group is encouraged to interact and discuss issues with each other rather than stick to answering the interviewer's questions.

Observations (participant observation)

Data are collected by watching people in a relevant setting – such as their regular day-to-day life or at a particular event – and making notes, taking photos, video or audio recordings. This type of data collection could be particularly useful to combine with other methods. It can be an effective way of understanding the contexts, behaviours and lifestyle of the observed participants, and insights can be gained from any difference between participants' reported experience and what is observed.

Pros of observations:

- Can reveal information that may not emerge through questioning.
- Can provide more details than might be described in an interview or questionnaire format.

Cons of observations:

- Recording the data is challenging, as it would be hard to know what to focus on and not always possible to note things down in the moment it is happening.
- Time-consuming.
- Potential for bias in that observations are filtered through the interpretation of the enumerator. Taking observations without making judgements or interpretations is a skill that requires practice and training.

DATA ANALYSIS METHODS: WAYS OF ANALYSING QUALITATIVE DATA

Qualitative content analysis

This method looks for patterns in the content, such as the frequency that a particular word is mentioned. Can be conducted using software, for example NVivo or R.

🔗 NVivo: *https://lumivero.com/products/nvivo/*

🔗 R: *www.r-project.org/*

Pros of qualitative content analysis:

- Can be automated to an extent and thus provide a relatively efficient way of extracting key ideas from qualitative data.

Cons of qualitative content analysis:

- By attempting to extract meaning from patterns rather than examining the details of individual statements, there is a risk of oversimplifying or losing important ideas.

Narrative analysis

This method is focused on the stories respondents tell and what insight can be gained from close examination of what is said and how. To give a simplified example, maybe one returnee describes their experience with a strong negative focus on their failed migration, and another's answers focus more on plans for the future.

This can reveal a lot about their emotional state and perceptions of their reintegration process.

Pros of narrative analysis:

- This method makes it possible to extract unique insights, beyond what is directly stated.

Cons of narrative analysis:

- It can be difficult to verify the results.
- The process of interpretation can be very subjective, thus open to bias.

Discourse analysis

This is an analysis of the language used in social contexts. For example, it might involve observing returnees interacting at a market, or listening to conversations during a focus group discussion. Examining how people talk to each other can reveal a lot about interpersonal relationships and the social and cultural dynamics.

Pros of discourse analysis:

- Can reveal a lot about the social factors that govern interactions.

Cons of discourse analysis:

- Requires a very clear aim for the analysis.

Thematic analysis

This method looks for patterns in data by grouping them into themes, such as when returnees' interview answers to questions about their experiences include phrases that describe positive or negative emotion, mention of violence, financial concerns and so on. The full set of answers from respondents can be scanned for phrases that fit into these categories and this can be reviewed to understand patterns and recurring ideas.

This is possible to automate using software such as NVivo or R.

Pros of thematic analysis:

- Useful for finding out about experiences and opinions.
- A useful process near the beginning of a study to gain new insights that could be further explored later.

Cons of thematic analysis:

- The choice of themes is very subjective and can shape the interpretations of the data set.
- It is possible to miss key points from individual respondents by looking at the overall themes.

You can read about an example of a qualitative study that used thematic analysis here:

Web page: *www.tandfonline.com/doi/full/10.1080/13669877.2018.1517376*

KNOWLEDGE CHECK

Connect each data-collection method to the correct category:

Instrumental variables

Focus group discussions

Thematic analysis

Propensity score matching

Surveys

Semi-structured interviews

Quantitative methods

Qualitative methods

Methods used for both

KNOWLEDGE CHECK ANSWER

Connect each data-collection method to the correct category:

Qualitative methods

- Thematic analysis
- Focus group discussions
- Semi-structured interviews

Quantitative methods

- Propensity score matching
- Instrumental variables

Methods used for both

- Surveys

INTERVIEW TRANSCRIPT: USING QUALITATIVE METHODS FOR IMPACT EVALUATION

In this interview, Professor Papa Sakho gives first-hand reflections and advice on the strengths and weaknesses of qualitative methods for impact evaluation of return and reintegration programmes.

ADVANTAGES AND DISADVANTAGES OF USING QUALITATIVE METHODS FOR IMPACT EVALUATION

Interview with Professor Papa Sakho, Cheikh Anta Diop University, Dakar

Hello, I am Professor Sakho, professor of geography at the Cheikh Anta Diop University in Dakar. I first started working on urban issues, then on mobility and for the last twenty years, I have been conducting research on internal and international migration.

The advantage of the qualitative method is that it allows us to weigh the feelings and representations that people have, because in the case of return and reintegration, this is the most important aspect, because we are talking to communities that have cultures and attitudes towards social practices that can only be acquired by using the qualitative method, i.e. by communicating with the communities concerned.

Bias in qualitative methods

In impact evaluation, the quantitative method is generally used much more because the qualitative method often contains biases. Those who conduct the surveys, the sponsors, tend to want to positivize the approach and the objectives, whereas we should be aiming for objectivity.

The qualitative method makes it possible to assess the state of the community. However, it still presents a major difficulty which is linked to the position of the interviewee to the interviewer. When we go into the field on several occasions, the interviewee often imagines the interviewer, saying to himself that the answers that will be given will have to be formulated in such a way that they can go in the direction of the person making the evaluation. They are human, thoughtful people who believe that by going the way of the donor they may be able to bring benefits. A recent example with doctoral students interviewing returnees showed that there is a gap between the answers given by these young interviewees and what they actually think. By working with the interviewee, observing him and immersing herself in his environment every day, the doctoral student realized that what is said in the interview is not in his everyday practice.

The solution is immersion because in the qualitative method, you can't just stop and interview. You have to immerse yourself, observe and talk with people who are in the environment of the interviewee. This way, you get a glimpse of reality and a better appreciation of it.

Generally, in order to avoid these biases, when we were in the field in southern Senegal, we interviewed people who were relatively far from these leaders and the programme. In this context, we tried to move away from the project and look for targets that were relatively far from the leaders and the project managers. The results were interesting in terms of response.

Whatever the attitude of the beneficiaries, it would be useful to decide objectively on the purpose of this evaluation. This will help clarify positions and target those who are willing to respond to conduct the project evaluation.

SAMPLE SIZES IN QUALITATIVE DATA

Qualitative methods tend to be significantly more time-consuming and complex than quantitative. This reduces the size of the sample that can be used, making it much smaller than for quantitative methods.

Question

How do you think this might affect an impact evaluation's design and conclusions? Select all the answers that apply.

- □ It is impossible to evidence claims about the general population using solely qualitative data.
- □ There is a risk of misinterpreting an individual's unique experience as common to the population.
- □ It is not possible to implement the same sampling strategy (such as stratification) as would be used for quantitative methods.
- □ It is reasonable to use a sample that does not represent the whole population, as long as this is made clear when sharing conclusions.

Answer

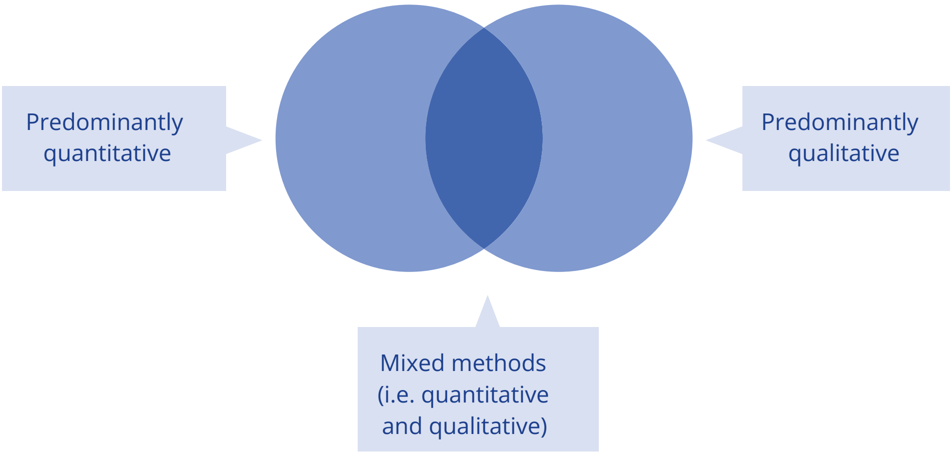
How do you think this might affect an impact evaluation's design and conclusions?

- □ **It is impossible to evidence claims about the general population using solely qualitative data.**
  It is not impossible, but it is difficult. The smaller sample size limits the capacity of the impact evaluator to easily aggregate data across many cases and to present a broad and generalizable description of the impact. This is a reason why it can be so helpful to use a combination of quantitative and qualitative methods, to take advantage of the strengths of each. For example, qualitative methods might reveal insights and raise questions, and then quantitative methods could be used to test and produce evidence to answer these questions.
- ☑ **There is a risk of misinterpreting an individual's unique experience as common to the population.**
  Due to the small sample size, it is difficult to use qualitative methods to produce results that are generalizable to the whole population. There is a risk of drawing conclusions that do not reflect the majority of the population based on what a small number of people report. It is important to be aware of this risk and take care when interpreting qualitative data.
- □ **It is not possible to implement the same sampling strategy (such as stratification) as would be used for quantitative methods.**
  Although sample sizes are reduced, it is still important to work within the sampling structure that has been planned for the study. The reasons behind the use of strategies like stratification are still important. Review Module 3 or see the IOM Monitoring and Evaluation Guidelines for more information on stratification.
- □ **It is reasonable to use a sample that does not represent the whole population, as long as this is made clear when sharing conclusions.**
  It is sometimes helpful to use qualitative methodology to target certain respondents, such as those who responded in a particular way to previous surveys, to probe further into certain ideas that have emerged. The fact that this sample only represents a subset of the overall population should be made clear and reflected in the conclusions.

With a small sample it becomes even more important to think carefully about its composition. If the qualitative data collection precedes any other, then initial design stratification should be a guide. On the other hand, if the qualitative data collection comes after quantitative data analysis, there is an opportunity for deliberate sampling, based upon the findings of the quantitative data. The qualitative enquiry may want to sample areas that have very different data responses on a certain indicator; e.g. where quantitative data indicates communities where a particular outcome is progressing well, contrasted with communities where that same outcome is not progressing well.

When conducting qualitative studies, the aim is not usually to make precise estimates about the population, but rather to capture the diversity of experiences and viewpoints. An increased sample size allows a broader range of viewpoints to be understood, but eventually can lead to a "saturation point" where more responses simply repeat similar answers and do not provide new insights. Therefore, given the aims of a qualitative study, increased sample sizes would often add significant expense without providing useful benefits.

MIXED-METHODS


Predominantly quantitative
Predominantly qualitative
Mixed methods (i.e. quantitative and qualitative)

There are often benefits to using mixed-methods in impact evaluation. While quantitative data may be able to measure the change in an impact indicator, qualitative data can provide insights that help explain the causes of the change, or put this change in context.

The inclusion of qualitative methods can be very important if, for example, it is not possible to use quantitative methods to establish causal attribution (such as where there is no comparison or control group).

The next section will explore how mixed-methods can be used.

For impact evaluation approaches using these mixed-methods, experts are uncommon, as evaluators often specialize in one or the other approach, and qualitative methods require significantly different skills to quantitative evaluation.

The most effective way to integrate qualitative and quantitative methods in an impact evaluation utilizing mixed-methods is to incorporate both into the overall design from the start.

.

The key to any impact evaluation approach, irrespective of the methods used, remains the articulation and use of the programme theory, theories of change and/or causal pathways and evaluation questions to guide the design, as discussed in earlier modules.

INTERVIEW TRANSCRIPT: ADVANTAGES OF MIXED-METHODS FOR IMPACT EVALUATION

In the following interview, Professor Papa Sakho gives some thoughts on the benefits of combining quantitative and qualitative methods for impact evaluation of return and reintegration programmes.

MIXED-METHODS: ADVANTAGES OF COMBINING QUALITATIVE AND QUANTITATIVE METHODS

Interview with Professor Papa Sakho, Cheikh Anta Diop University, Dakar

To measure the impacts of return and reintegration policies in particular and in migration, generally, quantitative methods are crucial to measure individuals, costs, financial benefits flows but above all it is important to assess the interpretations of the targeted beneficiaries.

In impact evaluations, the qualitative method must play an important role because quantitative methods usually stop at studying flows. When using the latter, it might be interesting to evaluate the targets reached in the project, but in the long run also to study the long-term quantitative data of these reintegration flows. At this point, the qualitative approach comes into play, which allows us to provide an explanation of these quantitative results acquired over the long term.

The quantitative method makes it possible to quantify, to categorize, but this gives "snapshots" of a moment in time and does not make it possible to measure the feelings which are behind the acts studied quantitatively.

The migrant is an individual and the host community is composed of individuals. It is important to be able to evaluate the perception and interpretations that these individuals and populations have of the policies that are being implemented. Beyond the face-to-face survey, there is an important part of participatory observation because some of what the individual says is not always identical to what we observe in the field. Observing these populations yourself and integrating their experiences into the analysis criteria is therefore crucial.

I will take the example of the return and reintegration programmes for migrants repatriated from Libya in the region of Kolda, in southern Senegal. As soon as these people arrive, they are given a certain amount of money to do market gardening or other small handicraft activities. What we see from a quantitative point of view is that money has been invested, targets have been identified and work has been produced based on these resources. However, a more in-depth observation, such as a perception survey, shows that these same targets say that this assistance was used to produce work, while at the same time specifying that this does not prevent them from migrating again because the investment is not large enough to meet the expectations for a decent life.

COMBINING QUALITATIVE AND QUANTITATIVE METHODS

Qualitative methods can be used throughout the evaluation:

Before data collection

Qualitative methods can be used early in the design process for an evaluation, when defining what is to be measured and planning for quantitative data collection, to help gain an understanding of what change is expected, what is the measure of this (such as the effect of the intervention on returnees' level of reintegration), how to measure it and what questions to ask.

This can help make sure that the information collected in later data-collection activities is relevant and useful for its intended purpose.

> Example
>
> In the SHARP Danwadaag Programme, focus group discussions were conducted, asking people in the community what integration looks like: "If this community was fully integrated, what would that look like to you?"
>
> They came out with a definition, covering things like "everybody would have the same access to jobs, everybody would have freedom to move around". The team used this qualitative information to design questions to ask people how they felt about the identified elements – "do you feel safe going to buy food?", etc. – in order to measure the things that the qualitative methods identified as important to defining local reintegration.

During data collection for the evaluation

It is possible to undertake quantitative and qualitative methods at the same time, often referred to as "in parallel". For example, it would be possible to conduct focus group discussions, a qualitative methodology, alongside individual returnee surveys, a quantitative methodology, to bring out more opinion-based and community-level consensus on the progress and experience returnees are reporting, with the interventions they are receiving.

This could also be conducted later in the process, but there could be advantages to carrying it out simultaneously with the quantitative data collection:

- The qualitative methods can produce information that cannot be obtained quantitatively, which can enrich the quantitative data you collect.
- Asking for respondents to recall their opinions or experiences later on can be suboptimal; recollection can be inaccurate and influenced by what has happened since.
- There is the potential to follow up on qualitative results that are collected in parallel with quantitative data collection later in the process.

Post-quantitative data collection and analysis stage

When conducting analysis of quantitative data, patterns can emerge that are hard to explain purely by looking at these data. Subsequent targeted qualitative research can deepen understanding of these patterns through "triangulation".

While this sometimes is thought of a process of verifying or rejecting results from quantitative data, the use of these mixed-methods for impact evaluation should be seen as a way to enrich our understanding of the cause–effect system of interest.

A pattern is observed, and perhaps programme members and experts discuss and agree why this could have happened. However, rather than simply making conclusions about a community in their absence, it would make sense to confirm these conclusions by talking to the communities and asking about why they think the changes have happened. This allows a process of increased understanding of results, to see if what the respondents say confirms, contradicts or qualifies what the quantitative analysis has suggested.

What is triangulation?

> "Triangulation facilitates validation of data through cross verification from more than two sources. It tests the consistency of findings obtained through different instruments and increases the chance to control, or at least assess, some of the threats or multiple causes influencing our results.
>
> Triangulation is not just about validation but about deepening and widening one's understanding. It can be used to produce innovation in conceptual framing. It can lead to multiperspective meta-interpretations. Triangulation is an attempt to map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one standpoint."

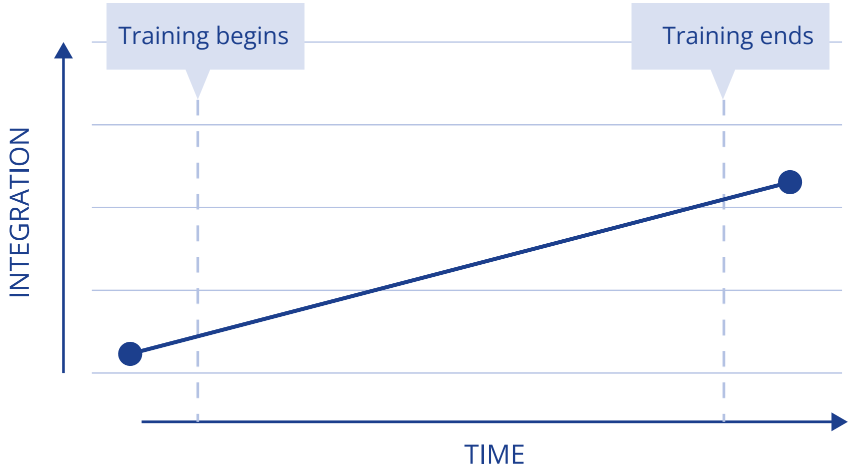
Web page: *www.betterevaluation.org/en/evaluation-options/triangulation*

COMBINING QUALITATIVE AND QUANTITATIVE METHODS

Let's look at an example. As part of a return and reintegration programme, a group of returnees have participated in training on eviction prevention.

Following the training, they complete quantitative surveys; the results of these show that the returnees who participated in the training felt more integrated overall afterwards.

FEELINGS OF INTEGRATION BEFORE AND AFTER EVICTION PREVENTION TRAINING



Quantitative analysis identifies this correlation and hypothesizes that they feel more integrated because of the training. Specifically, the hypothesis is that the skills they learned led them to feel more integrated. The quantitative analysis also showed a pattern between whether or not a returnee has been evicted and whether they feel integrated.

However, we cannot know, at this point, if it was the training, the post-training eviction experience and/or something completely unrelated to the training that caused them to feel more integrated.

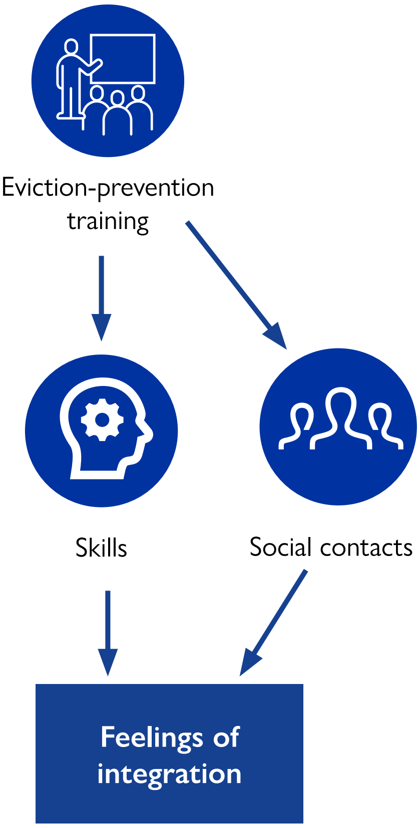
Following this, a qualitative process of discussion of these results involved meeting with the respondents to ask about their answers and the reasons why their feeling of integration has improved between survey rounds.

Perhaps their answers confirm the results of the quantitative analysis, and responses largely say that the eviction training was a major factor in this improvement as it reduced their chances of being evicted.

However, perhaps respondents say something completely different – for example, they made new social contacts during the training and now have a base of friends and acquaintances, so they feel more integrated.

Although the training has also influenced them feeling integrated, the explanation of "why" and the subsequent feedback to stakeholders will be quite different.

This example shows how qualitative methods can allow you to either verify the results, realize you have been mistaken, or – more likely – better understand the nuances of the issues that affect reintegration.



KNOWLEDGE CHECK

Question 1

As with quantitative methods, it is important to be mindful of the risk of bias.

In the context of the previous example about evaluation of eviction prevention training, which of these do you think would be the best way to ask questions when investigating whether the training was the cause of the improvement in participants' feeling integrated?

Choose one answer:

- □ Did the eviction training make you feel better integrated?
- □ Why do you think you reported that you feel more integrated now than you did previously?
- □ What effect did the eviction training have on how integrated you feel?

Question 2

Which of the following are advantages of using mixed-methods in an impact evaluation? Select all the answers that apply.

- □ Quantitative methodology brings out commonality, allowing the identification of patterns, while qualitative methodology allows more detailed probing of these patterns.
- □ Qualitative methods can save time and money compared to using large scale quantitative studies.
- □ Qualitative methods can inform the design decisions involved in planning a quantitative study.
- □ Information gained from qualitative methods can be used to verify or clarify conclusions drawn from quantitative analysis.

The answers are on the next page.

Question 1 answer

In the context of the previous example about evaluation of eviction prevention training, which of these do you think would be the best way to ask questions when investigating whether the training was the cause of the improvement in participants' feeling integrated?

- □ **Did the eviction training make you feel better integrated?**
  This question is leading the respondent towards a particular response and thus would risk producing biased results.
- ☑ **Why do you think you reported that you feel more integrated now than you did previously?**
  This is the best option in the list. Open questions like this reduce the chance of positively biasing answers but still require careful probing during the subsequent discussions. For example, even if the respondent mentions the eviction training, then the interviewer would want to try and tease out whether they felt the training gave them skills to prevent being evicted and therefore they felt more integrated – or if there were other aspects of the training which made them feel more integrated.
- □ **What effect did the eviction training have on how integrated you feel?**
  This is less leading than the first option but still encourages the respondent to consider the eviction training as the cause of their feelings of integration.

Question 2 answer

Which of the following are advantages of using mixed-methods in an impact evaluation?

- ☑ **Quantitative methodology brings out commonality, allowing the identification of patterns, while qualitative methodology allows more detailed probing of these patterns.**
  This is often the root of the added value of incorporating mixed-methods during and after a quantitative study.
- □ **Qualitative methods can save time and money compared to using large scale quantitative studies.**
  Qualitative methods are time-consuming. They can be cheaper in that they use less enumerators and a smaller sample size, but they require skilled interviewers, which would usually be more costly than quantitative survey enumerators. Although they can be very valuable, they should not be considered an easy or cost-saving option.
- ☑ **Qualitative methods can inform the design decisions involved in planning a quantitative study.**
  Using qualitative methods can provide information to help define what is expected to change and how this can be defined and measured.
- □ **Information gained from qualitative methods can be used to verify or clarify conclusions drawn from quantitative analysis.**
  As seen in the eviction-prevention training example, qualitative methods can provide information that provides a better understanding of nuanced situations such as reintegration.

CASE STUDY

An impact evaluation is being planned of a programme that aims to support sustainable reintegration. The programme provides a range of services that target the whole community that is affected by the migration and return. The aims are to evaluate how the activities contribute to improved integration for the returnees and members of the community and discover which activities have the biggest effect and sustained long-term impact on reintegration.

The team employed qualitative methods in the design phase; they conducted a series of focus group discussions with returnees to establish an agreed-upon definition of successful reintegration and used this to inform the design of their questionnaires for data collection. Following this use of qualitative approaches at the start of the process, the rest of the study was conducted primarily using only quantitative methods.

Consider the following questions:

- How did the inclusion of qualitative methods improve the quantitative design?
- Could the use of qualitative methods elsewhere in the evaluation process have been useful? How?
- Why might the project have chosen not to have additional qualitative methods in their evaluation?
- If you were designing a similar evaluation, what do you think you would do?

Take some time to note down your thoughts.

INTERVIEW TRANSCRIPT: USING MIXED-METHODS IN IMPACT EVALUATION

In this interview, Jasper Tjaden, Professor for Applied Social Research and Public Policy at the Economic and Social Science Department of the University of Potsdam, and formerly of IOM and the World Bank, gives an insight into the importance and potential of mixed-methods in impact evaluations.

USING MIXED-METHODS IN IMPACT EVALUATION

Interview with Professor Jasper Tjaden, Professor for Applied Social Research and Public Policy, University of Potsdam

In my opinion, mixed-methods are always great; combining quantitative and qualitative approaches… they both provide a lot of value and they have some great synergies and reinforce each other.

For example, with any impact evaluation you want to have a really good understanding of the theory of change: what is it that you're actually trying to achieve? And what is everything that needs to happen for this objective to be achieved? This involves a lot of assumptions. Qualitative research could be a great way of checking those assumptions.

Because project implementers always go to the field and they have certain assumptions about what beneficiaries want, who they are, what their preferences are, what their potential behaviour might be, but often there's very limited information on that ground, so qualitative research a good way of getting to know your target audience, getting to know beneficiaries more properly to actually fine-tune the assumptions that you have in your theory of change. For me, this is one really good way to use qualitative data collection.

Another good way of using qualitative data collection is to understand why certain things have worked or have not worked after the project is over. Sometimes we might find a strong effect through quantitative methods and say, "oh the impact of this is huge", but we don't really know why and how it works. This is called the mechanism. So to find out how this impact kind of unfolded and what the mechanism was, qualitative research is a really great way to get this rich kind of big information.

Equally, it's helpful to understand why something did not work. Often, you try out something, you run a study and you find actually there are no effects. This is possible; it happens quite often. And everyone has to be okay with that, by the way. But if that happens, then qualitative research is a great way of understanding why maybe it hasn't worked. Talking with certain beneficiaries or with people involved and trying to get trying to get information on what happened that may have prevented the effect from unfolding, or where your assumptions about how the project would work may have been wrong.

When qualitative methods can be used in an impact evaluation process

You can use qualitative data collection, certainly before the project to check whether, as I said, whether the theory of change is actually accurate, that you're planning to implement before you implement it.

You can also use qualitative data collection to check whether beneficiaries understand the questions that you have in your questionnaire; whether they mean the same thing that you think it means. You can even use it way earlier, when you're designing the project that you're actually trying to implement, and often the impact evaluation goes hand in hand with… the people on the implementation side, thinking about what they actually implemented. And what is being implemented often changes on the spot in the field.

So, the impact evaluation team often works very closely with implementers, and there some qualitative data collection in the very beginning before the project has started can be very useful to fine-tune the intervention, but also fine-tune the data collection and the questionnaire and so forth – that goes hand in hand.

MIXED-METHODS APPROACHES IN IMPACT EVALUATION

"Approaches" for impact evaluation refers to the overall strategy – the combination of options that are used in an impact evaluation design.

There are a wide range of impact evaluation approaches, some of which can be implemented as purely qualitative (such as a qualitative impact assessment protocol) or quantitative (such as a randomized control trial), but the majority involve, and usually benefit from, a combination of both (i.e. mixed-methods) (examples include contribution analysis and realist evaluation).

Mixed-methods have the advantage of providing a richer understanding through the combination of different kinds of data around differing aspects of the evaluation.

While approaches for impact evaluation can differ significantly, they have the same aim as described in earlier modules, i.e. to establish causal attribution of the programme on changes observed. Similarly, they all require clear programme theory, theories of change and/or causal pathways and evaluation questions, although some approaches are more theory-based than others.

You can read more about evaluation approaches and see a list with definitions on the Better Evaluation website.

Web page: *www.betterevaluation.org/methods-approaches/approaches*

QUIZ

This quiz will check your understanding of the topics covered in this module.

There are six questions. You must get a score of at least five out of six to pass.

1. Which of the following best describes qualitative methods? Select one answer.

- □ Qualitative methods require definition of what is to be measured ahead of time to design data collection and analysis.
- □ Qualitative methods use open-ended questions to facilitate study of issues in depth and detail, based on data that are descriptive in nature, rather than data that can be measured or counted.
- □ Qualitative methods deal with feelings and social issues.
- □ Methods that are used to verify information gained from quantitative methods.

2. Which of the following are qualitative methods for data collection? Select all the answers that apply.

- □ Qualitative content analysis.
- □ Participant observation.
- □ Survey.
- □ Difference in difference.
- □ Semi-structured interview.

3. Which of the following describe roles that qualitative methods play in mixed-methods approaches? Select all the answers that apply.

- □ Provide data about emotions, personal perspectives and the drivers behind changes that might not be easy to identify using quantitative methods.
- □ Inform the planning of quantitative data collection.
- □ Deepen understanding of the patterns observed in quantitative data.
- □ Provide an unbiased insight into respondents' experiences.

4. When can you use qualitative methods in an impact evaluation? Select one answer.

- □ To inform planning for evaluation data collection.
- □ During quantitative data collection.
- □ After analysis of quantitative data.
- □ All of the above.

5. Which of the following are potential benefits of using qualitative methods as part of a mixed-methods evaluation? Select all the answers that apply.

- □ Verifying conclusions made from quantitative analysis.
- □ Explaining patterns in the quantitative data.
- □ Qualitative methods help establish what contributed to changes that are observed.
- □ Sample sizes are not important for qualitative methods.

6. Which of the following could be a source of bias in qualitative methods? Select all the answers that apply.

- □ The enumerator.
- □ How a question is phrased.
- □ Asking open-ended questions.
- □ Using an online survey.
- □ Small sample size.

QUIZ ANSWERS

There are **s**ix questions. You must get a score of at least five out of six to pass.

1. Which of the following best describes qualitative methods? Select one answer.

- ☐ Qualitative methods require definition of what is to be measured ahead of time to design data collection and analysis.
- ☑ Qualitative methods use open-ended questions to facilitate study of issues in depth and detail, based on data that are descriptive in nature, rather than data that can be measured or counted.
- ☐ Qualitative methods deal with feelings and social issues.
- ☐ Methods that are used to verify information gained from quantitative methods.

2. Which of the following are qualitative methods for data collection? Select all the answers that apply.

- ☐ Qualitative content analysis.
- ☑ Participant observation.
- ☑ Survey.
- ☐ Difference in difference.
- ☑ Semi-structured interview.

3. Which of the following describe roles that qualitative methods play in mixed-methods approaches? Select all the answers that apply.

- ☑ Provide data about emotions, personal perspectives and the drivers behind changes that might not be easy to identify using quantitative methods.
- ☑ Inform the planning of quantitative data collection.
- ☑ Deepen understanding of the patterns observed in quantitative data.
- ☐ Provide an unbiased insight into respondents' experiences.

4. When can you use qualitative methods in an impact evaluation? Select one answer.

- ☐ To inform planning for evaluation data collection.
- ☐ During quantitative data collection.
- ☐ After analysis of quantitative data.
- ☑ All of the above.

5. Which of the following are potential benefits of using qualitative methods as part of a mixed-methods evaluation? Select all the answers that apply.

- ☑ Verifying conclusions made from quantitative analysis.
- ☑ Explaining patterns in the quantitative data.
- ☑ Qualitative methods help establish what contributed to changes that are observed.
- ☐ Sample sizes are not important for qualitative methods.

6. Which of the following could be a source of bias in qualitative methods? Select all the answers that apply.

- ☑ The enumerator.
- ☑ How a question is phrased.
- □ Asking open-ended questions.
- ☑ Using an online survey.
- ☑ Small sample size.

SUMMARY

In this module, we have seen that:

1. Qualitative methods use open-ended questioning to obtain detailed, descriptive information about opinions, experiences and perspectives.
2. Qualitative methods can enrich information gained from quantitative methods.
3. Qualitative methods can be introduced to a mixed-methods evaluation at any stage in the evaluation process.
4. Mixed-methods approaches take advantage of the strengths of both quantitative and qualitative methods.

MODULE 6: WHICH IMPACT EVALUATION DESIGN, WHEN?

Returnee woman living in an IDP camp explains to an IOM staff her journey from Yemen.
© IOM 2021/Rikka TUPAZ

MODULE 6: WHICH IMPACT EVALUATION DESIGN, WHEN?

INTRODUCTION

This module will present decision trees to select appropriate design options for the quantitative impact evaluation, providing an overview or checklist. Trainees will be invited to navigate these, applying understanding gained in the previous modules, to generate a suggested set of impact evaluation designs by context.

OUTCOMES

At the end of this module, trainees will be able to:

- Summarize the core decisions that form the basis for designing an impact evaluation.
- Explain how the decisions were reached in presented examples.
- Navigate an impact evaluation decision tree.

WHAT IS A DECISION TREE?

The decision trees describe common paths of questions and possible conclusions that an evaluator or evaluation team would consider when planning an impact evaluation process. They summarize the process of designing an impact evaluation and can act as flow charts or checklists to guide this process.

Module 3 (see page 69) introduced several different methodologies that can be used in an impact evaluation. If you are not familiar with these methods, we recommend reviewing Module 3.

Depending on the characteristics of the evaluation, a combined approach can be very effective. Most impact evaluations used mixed-methods to some extent.

How do you choose which methods are appropriate to use?

There is no single way to do an impact evaluation and designing one means making a series of decisions about what is appropriate for the context and to meet requirements.

The decision trees provide a route through this decision-making process. The trees cover three key decision processes to guide the design of an impact evaluation:

Decisions about what should be measured in the context of reintegration.

How to measure the difference that the treatment has made, i.e. the size and direction of impact.

Decisions about the returnees on whom the impact is to be measured.

PREREQUISITES AND CAUTIONS

As we saw in previous modules, planning an impact evaluation relies upon certain groundwork being in place; the project must have:

- A programme theory that articulates how the intervention will produce the intended impact.
- Evaluation questions based on clear definitions of exactly what treatment will be evaluated and what the intended outcome is.

This series of decision trees assumes that the above prerequisites for planning an impact evaluation are in place.

Points to note

- The decision trees presented here have a bias towards quantitative methods.
- As this course content is focused solely on the specifics of impact evaluation for return and reintegration, two assumptions have been made:
  - The trees focus on impact evaluations for interventions that are designed to achieve reintegration. We know that this is not always the case, but decisions on how to select indicators that measure other types of impact are outside the scope of the course, and we recommend reviewing the IOM Monitoring and Evaluation Guidelines for guidance on selecting indicators:
    - PDF: https://publications.iom.int/system/files/pdf/IOM-Monitoring-and-Evaluation-Guidelines_1.pdf#page=78
  - The trees assume that the impact of the treatment will be measured on individual returnees. In some cases, impact is expected on groups of individuals, organizations, communities, etc. However, the trees are focused on evaluations that measure reintegration, and measures of reintegration are made at the individual level.
- By necessity, there is some simplification of the process, but the decision trees should provide a foundation of the key elements needed in a design.

The trees will lead to answers that will support the design of an impact evaluation.

It is likely that some decisions will require consultation with experts, establishing availability of data or consultations with partners. Therefore, the user may expect deviations and some increase in the level of complexity before reaching conclusions about the design of a specific impact evaluation process.

THE DECISION TREES

Here are the decision trees for each key decision process (Figure 43, Figure 44 and Figure 45). In the following section, we will walk through the processes with an example to examine how the decision-making process works.

FIGURE 43: DECISION TREE – WHAT SHOULD BE MEASURED?

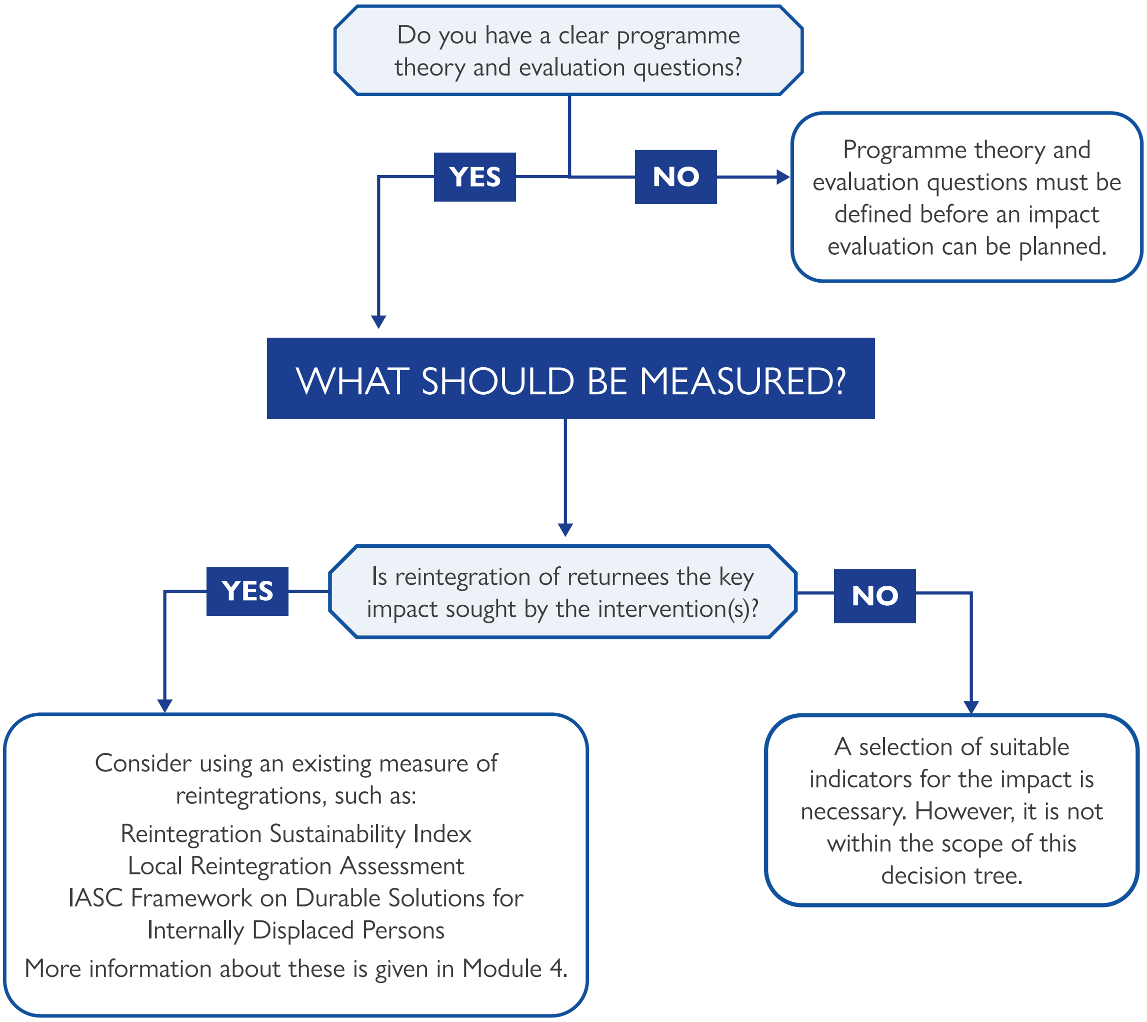
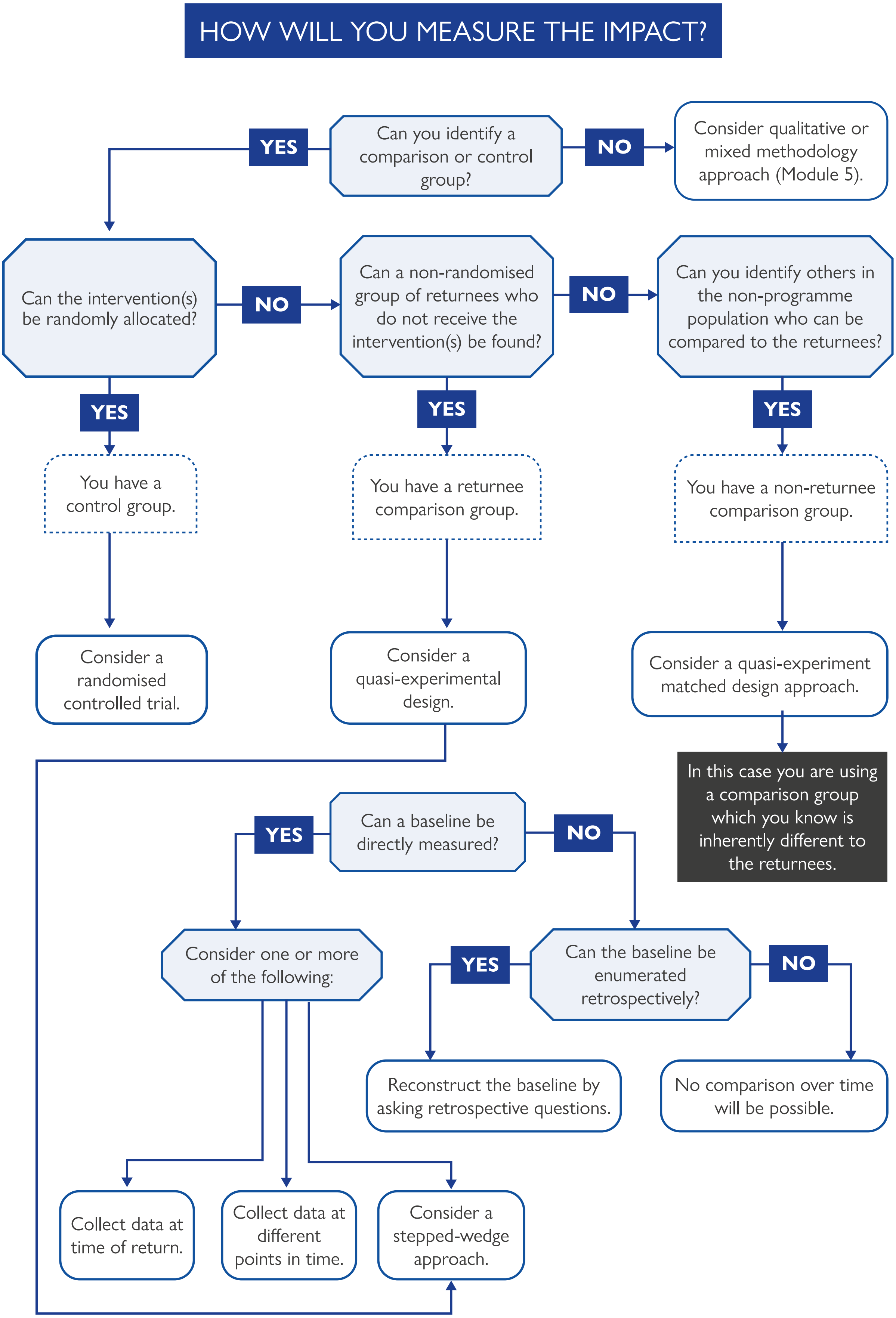
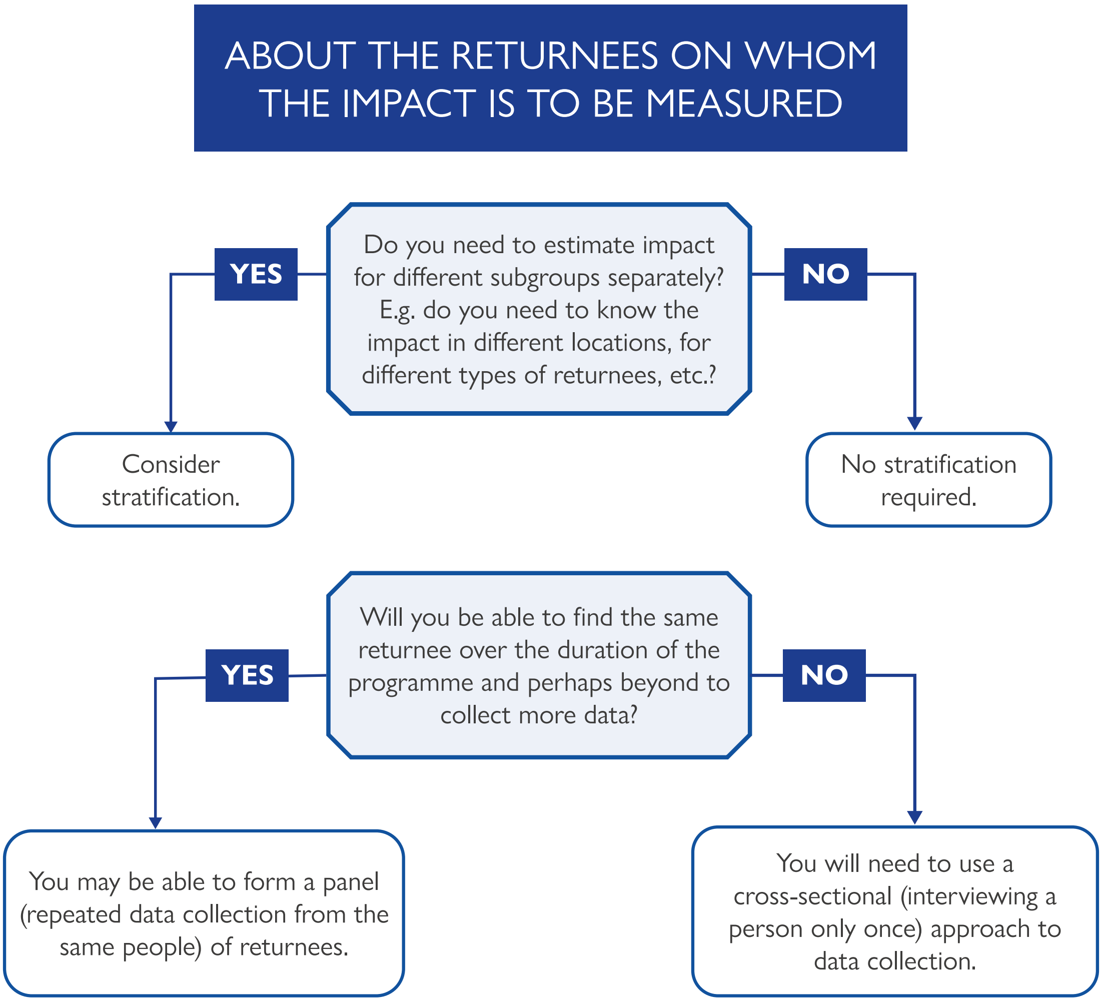

Do you have a clear programme theory and evaluation questions?
YES
NO
Programme theory and evaluation questions must be defined before an impact evaluation can be planned.
WHAT SHOULD BE MEASURED?
Is reintegration of returnees the key impact sought by the intervention(s)?
YES
NO
Consider using an existing measure of reintegrations, such as:
Reintegration Sustainability Index
Local Reintegration Assessment
IASC Framework on Durable Solutions for Internally Displaced Persons
More information about these is given in Module 4.
A selection of suitable indicators for the impact is necessary. However, it is not within the scope of this decision tree.

FIGURE 44: DECISION TREE – HOW WILL YOU MEASURE THE IMPACT?



HOW WILL YOU MEASURE THE IMPACT?
Can you identify a comparison or control group?
YES
NO
Consider qualitative or mixed methodology approach (Module 5).
Can the intervention(s) be randomly allocated?
NO
Can a non-randomised group of returnees who do not receive the intervention(s) be found?
NO
Can you identify others in the non-programme population who can be compared to the returnees?
YES
YES
YES
You have a control group.
You have a returnee comparison group.
You have a non-returnee comparison group.
Consider a randomised controlled trial.
Consider a quasi-experimental design.
Consider a quasi-experiment matched design approach.
In this case you are using a comparison group which you know is inherently different to the returnees.
YES
Can a baseline be directly measured?
NO
Consider one or more of the following:
YES
Can the baseline be enumerated retrospectively?
NO
Reconstruct the baseline by asking retrospective questions.
No comparison over time will be possible.
Collect data at time of return.
Collect data at different points in time.
Consider a stepped-wedge approach.

FIGURE 45: DECISION TREE – ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED


ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED
Do you need to estimate impact for different subgroups separately? E.g. do you need to know the impact in different locations, for different types of returnees, etc.?
YES
NO
Consider stratification.
No stratification required.
Will you be able to find the same returnee over the duration of the programme and perhaps beyond to collect more data?
YES
NO
You may be able to form a panel (repeated data collection from the same people) of returnees.
You will need to use a cross-sectional (interviewing a person only once) approach to data collection.

USING THE DECISION TREES

CASE STUDY

Let's explore the decision tree by looking at a familiar example of a return and reintegration programme: the EU-IOM Joint Initiative for Migrant Protection and Reintegration.

> "The EU-IOM Joint Initiative aims at enabling returnees to restart their lives in their countries of origin, grounded in IOM's integrated approach to reintegration. Reintegration assistance under the EU-IOM Joint Initiative supports migrants and their communities, has the potential to complement local development, and mitigates some of the drivers of irregular migration."
>
> "The reintegration support aims to address returnees' economic, social and psychosocial needs and to foster inclusion of communities of return in reintegration planning and support whenever possible. The EU-IOM Joint Initiative does not foresee standard reintegration packages. Instead, reintegration counsellors and returnees jointly define individual reintegration plans, which are tailored to the returnees' needs and vulnerabilities as well as their opportunities and motivations. The support may be provided to individuals, groups or communities."

Web page: *www.migrationjointinitiative.org/reintegration*

The interventions available include:

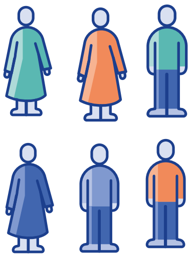
- Assistance with immediate needs upon arrival – from reception and transportation to support with urgent medical or legal needs
- A range of training, education, housing and childcare support
- Grants to establish a micro-business
- Counselling
- Legal aid
- Community projects to foster reintegration on a longer-term basis.

Assisted voluntary return and reintegration is among the services offered through Migration Response Centres in the Sudan. © IOM 2021/ Muse MOHAMMED

The programme includes an impact evaluation, the "Impact Evaluation of the EU-IOM Joint Initiative Programme for Migrant Protection and Reintegration (Horn of Africa)", (also known as IMPACT), which aims to provide a robust assessment of the impact of IOM's reintegration assistance on the sustainable reintegration of supported migrant returnees. IMPACT focuses on Ethiopia, Somalia and the Sudan, where the number of programme beneficiaries is the highest.

There is a need for evidence about the reintegration impacts produced by the individualized support provided to the returnees who are enrolled in the programme. This should provide evidence and learning to inform decision-making for the current and future programming.

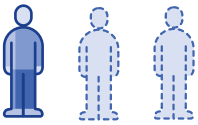
The IMPACT study is also acting as a test case for using the Reintegration Sustainability Survey as part of an impact evaluation. This survey is meant to be conducted with returnees within three months after return (baseline) and then a second time after one year (endline).



The evaluation attempts to categorize and compare different types of returnees over time. For example: they may be categorized by level of traumatization during migration, well-being before last migration, support levels provided since return, length of time before receiving support after return.

The evaluation includes multiple countries; there are differences in how the programme is implemented in these countries, and evaluation results will need to be reported for each one.

Some challenges are foreseen in the implementation of the IMPACT study, including the issue of a very mobile population: returnees can change mobile numbers and addresses frequently, have limited connectivity and sometimes remigrate. This makes it difficult to follow returnees and get in touch to collect data.



There are also other organizations working in the same area; similar support offered by other programmes can make it difficult to isolate the effect of IOM’s activities.

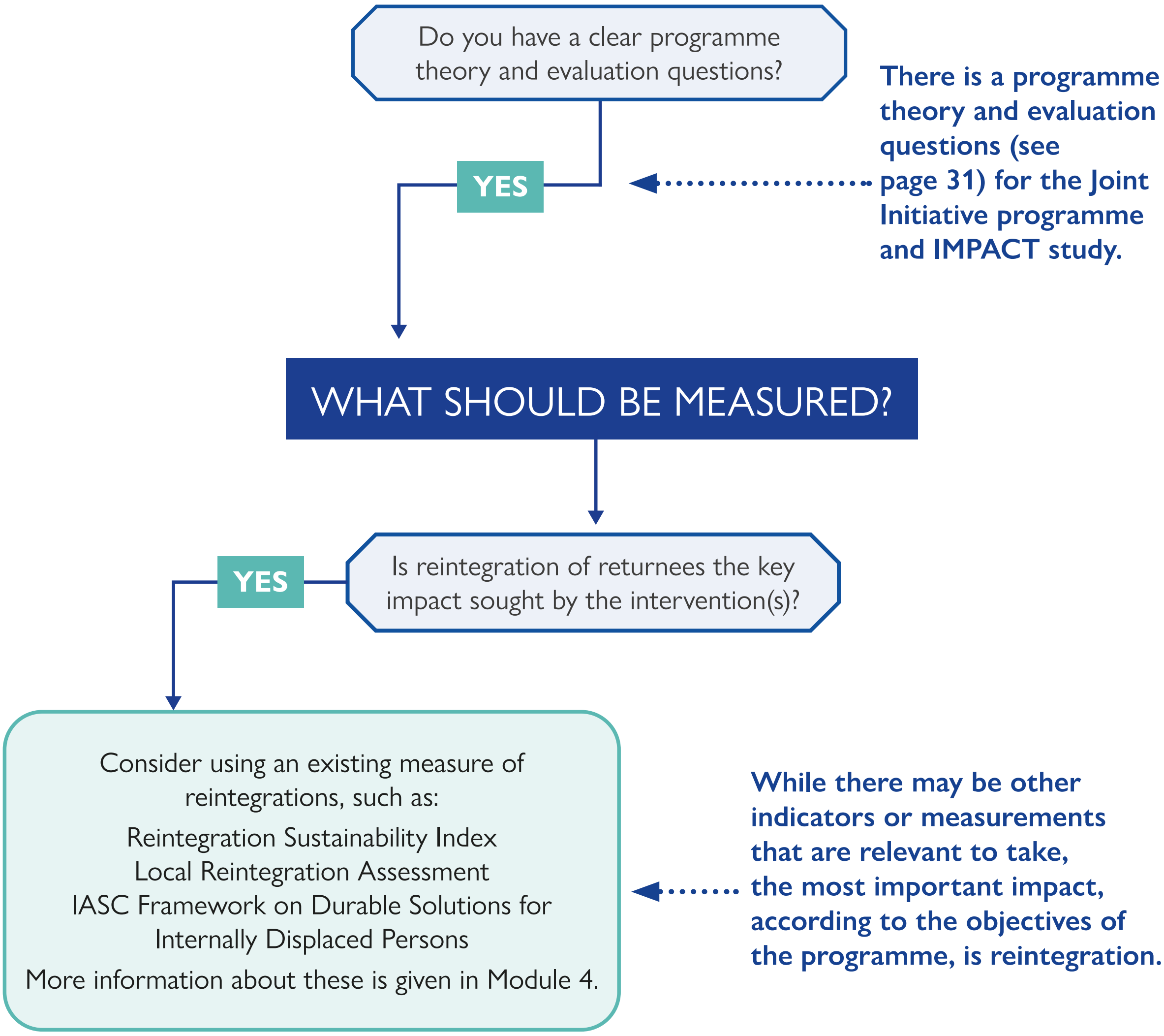
COMPLETED DECISION TREES

The following pages show the completed decision trees for the design of the IMPACT study with notes and annotations about the choices and the process that led to the final implementation. It also includes questions inviting you to decide what you would choose from the available options, or to consider the reasoning that led to the actual decisions for the design of the study. It also includes the “IMPACT team response” – notes from the perspective of the evaluation team responsible for the study.

For a lot of the decisions involved in the process, there is not simply a “correct” answer. There is more than one suitable solution to the challenge of designing any impact evaluation; the choices between competing valid options can be subjective, and other factors, such as cost, may need to be considered.

Although we will share the design decisions the IMPACT evaluation team arrived at, we encourage you to think about whether you would have done things differently.

FIGURE 46: DECISION TREE FOR IMPACT STUDY DESIGN – WHAT SHOULD BE MEASURED?


Do you have a clear programme theory and evaluation questions?
YES
WHAT SHOULD BE MEASURED?
Is reintegration of returnees the key impact sought by the intervention(s)?
YES
Consider using an existing measure of reintegrations, such as:
Reintegration Sustainability Index
Local Reintegration Assessment
IASC Framework on Durable Solutions for Internally Displaced Persons
More information about these is given in Module 4.

There is a programme theory and evaluation questions (see page 31) for the Joint Initiative programme and IMPACT study.

While there may be other indicators or measurements that are relevant to take, the most important impact, according to the objectives of the programme, is reintegration.

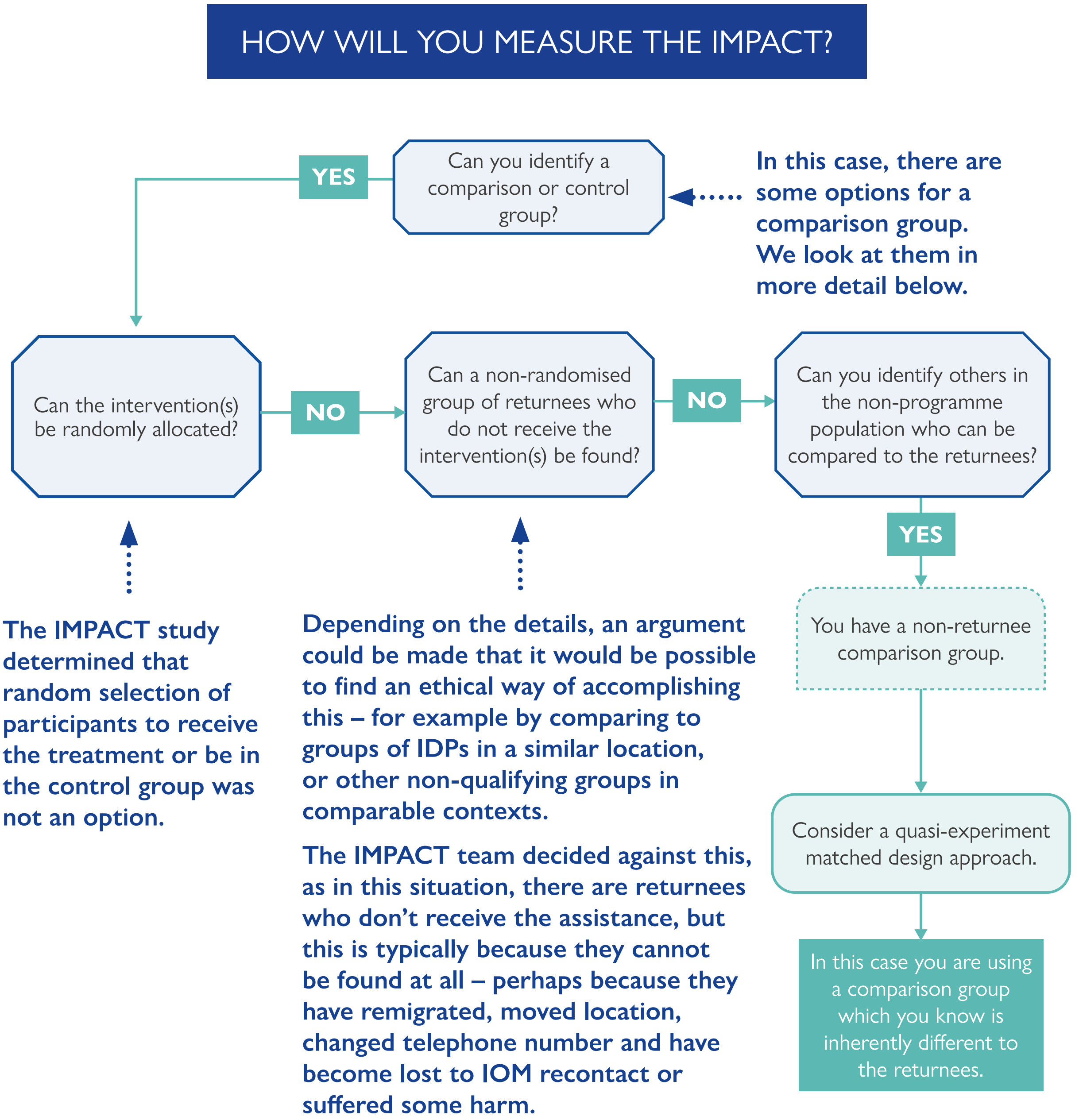
IMPACT team response:

The IMPACT study aims to understand, specifically, the impact on reintegration produced by the programme's tailored assistance. The overall intended impact of the treatment is to increase the level of sustainable reintegration of the beneficiaries. Therefore, this is the key impact that must be measured.

The IMPACT study will use the Reintegration Sustainability Survey (RSS); in fact, one of the objectives of the study is to function as a test case for the RSS as part of an impact evaluation.

In other scenarios, selection of an appropriate way of measuring reintegration may depend on a number of factors, including your programme or organization's definition of reintegration and the specific aims and priorities of the treatment or evaluation (see page 113).

FIGURE 47: DECISION TREE FOR IMPACT STUDY DESIGN – HOW WILL YOU MEASURE THE IMPACT? (PART 1)



HOW WILL YOU MEASURE THE IMPACT?

Can you identify a comparison or control group?

In this case, there are some options for a comparison group. We look at them in more detail below.

YES

Can the intervention(s) be randomly allocated?

NO

Can a non-randomised group of returnees who do not receive the intervention(s) be found?

NO

Can you identify others in the non-programme population who can be compared to the returnees?

YES

You have a non-returnee comparison group.

Consider a quasi-experiment matched design approach.

In this case you are using a comparison group which you know is inherently different to the returnees.

The IMPACT study determined that random selection of participants to receive the treatment or be in the control group was not an option.

Depending on the details, an argument could be made that it would be possible to find an ethical way of accomplishing this – for example by comparing to groups of IDPs in a similar location, or other non-qualifying groups in comparable contexts.

The IMPACT team decided against this, as in this situation, there are returnees who don’t receive the assistance, but this is typically because they cannot be found at all – perhaps because they have remigrated, moved location, changed telephone number and have become lost to IOM recontact or suffered some harm.

IMPACT team response:

The IMPACT study decided to use non-migrant residents with a matched design (see page 89) as a comparison group.

In the documentation, they refer to this as a “calibration group” in acknowledgement of the fact that this type of comparison group is fundamentally different to the returnees. As they have not recently migrated and returned, they cannot give an estimate of what would have happened to the returnees without the assistance. However, by matching returnees with non-migrants with similar characteristics, some useful information can still be gained using this comparison.

Question 1

The IMPACT study determined that random selection of participants to receive the treatment or be in the control group was not an option. Why do you think this was? Choose the option you think is the most important factor in making this decision.

- □ Ethical issues with random allocation.
- □ Practical difficulties due to the logistics of delivering interventions.
- □ Ethically, the participants would need to be informed that they will be randomly assigned to treatment or control. It would be difficult to motivate participation in a reintegration programme if there is a chance of not receiving the support.

Question 2

The decision to use non-migrant residents as a "calibration group" is a compromise that had to be made in the IMPACT study design for reasons of ethics and feasibility. What are the consequences of this? Select all the answers that apply.

- □ Evidence for causal attribution or contribution provided by this study will be weaker.
- □ It will be more difficult to understand the impact of the programme activities.
- □ Evaluators will have to be careful to try to counteract inherent biases in the study's findings.
- □ There may be a need to supplement the information provided by the study with other methods in order to reach the objectives of the impact evaluation.

The answers are on the next page.

Question 1 answer

The IMPACT study determined that random selection of participants to receive the treatment or be in the control group was not an option. Why do you think this was?

□ **Ethical issues with random allocation.**
This was the primary reason for the IMPACT team's choice not to use a randomized returnee control group.

☑ **Practical difficulties due to the logistics of delivering interventions.**
This was not the primary concern for the IMPACT team's choice, but it is a relevant concern in some situations.

□ **Ethically, the participants would need to be informed that they will be randomly assigned to treatment or control. It would be difficult to motivate participation in a reintegration programme if there is a chance of not receiving the support.**
This is something to consider whenever deciding if a control group is a valid option. In this instance, it was not the main reason the team decided against it.

Question 2 answer

The decision to use non-migrant residents as a "calibration group" is a compromise that had to be made in the IMPACT study design for reasons of ethics and feasibility. What are the consequences of this?

☑ **Evidence for causal attribution or contribution provided by this study will be weaker.**
This is correct. Without a comparison group that is very similar to the treatment group, evidence for attribution will be limited.

☑ **It will be more difficult to understand the impact of the programme activities.**
This is correct. Using only a "calibration group" limits the ability of the study to isolate which changes observed are due to the treatment.

☑ **Evaluators will have to be careful to try to counteract inherent biases in the study's findings.**
This use of a calibration group is prone to potential biases.

□ **There may be a need to supplement the information provided by the study with other methods in order to reach the objectives of the impact evaluation.**
Complementing the quantitative study with qualitative methods, for example, could help avoid or reduce the difficulties described above.

IMPACT team response:

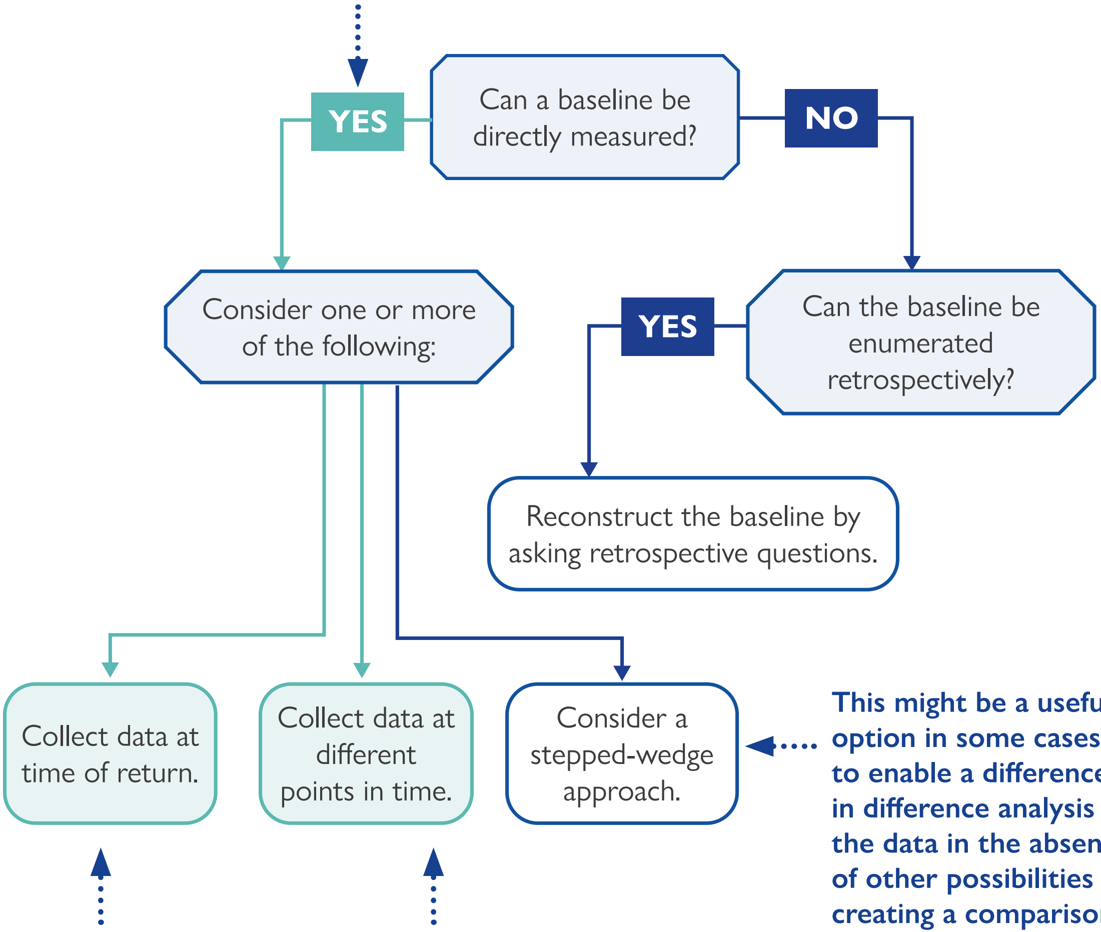
In the IMPACT study, the evaluation team acknowledged that it would not be feasible to create a comparison group that was very similar to the returnees receiving the treatment and that this would reduce the ability of the study to provide a "robust assessment of the impact of IOM reintegration assistance".

The solution was to adopt a hybrid design, or "mixed-methods" approach, which combined a quantitative approach (the design we have arrived at using this decision tree) with natural experiments (see page 95) and qualitative research.

In this way, the study aims to take advantage of the strengths of the different methods and counteract their weaknesses.

FIGURE 48: DECISION TREE FOR IMPACT STUDY DESIGN – HOW WILL YOU MEASURE THE IMPACT? (PART 2)

To measure impact of the Joint Initiative reintegration assistance, the relevant time for the baseline data collection is prior to provision of assistance, or on the day assistance is delivered. At this stage, at the beginning of participation in the programme, the intention was to contact beneficiaries and carry out the RSS baseline survey.


Can a baseline be directly measured?
YES
NO
Consider one or more of the following:
YES
Can the baseline be enumerated retrospectively?
Reconstruct the baseline by asking retrospective questions.
Collect data at time of return.
Collect data at different points in time.
Consider a stepped-wedge approach.

This might be a useful option in some cases to enable a difference in difference analysis of the data in the absence of other possibilities for creating a comparison group. However, this option involves delaying provision of support for some beneficiaries.

This would be the sensible time to collect baseline data, as we saw in the previous step. Programme assistance begins with immediate assistance upon arrival, so the baseline should be conducted before this comes into effect.

As the study aims to measure the impact of the treatment, some kind of before/after measurement would be strongly advisable, so this is a sensible option.

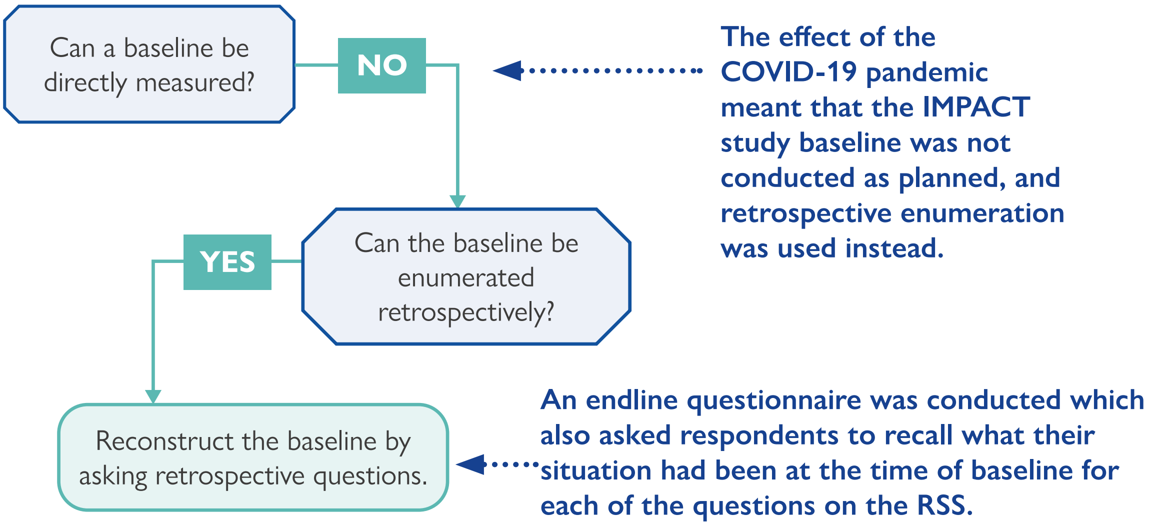
IMPACT team response:

While a stepped-wedge approach may have helped solve the challenge of creating a comparison group to permit a difference in difference approach, the decision was made not to artificially delay provision of support.

The IMPACT study design aims to collect data between 12–18 months after return to use as an endline. Although there are challenges with maintaining contact with returnees in the longer term, this is still achievable, and a before/after comparison is necessary for the objectives of the evaluation.

However, the COVID-19 pandemic caused significant reduction in the flow of new returnees arriving in the three countries. It was therefore necessary to use another strategy to create the baseline: retrospective enumeration.

FIGURE 49: DECISION TREE FOR IMPACT STUDY DESIGN – HOW WILL YOU MEASURE THE IMPACT? UPDATED STRATEGY FOR THE BASELINE


Can a baseline be directly measured?
NO
Can the baseline be enumerated retrospectively?
YES
Reconstruct the baseline by asking retrospective questions.

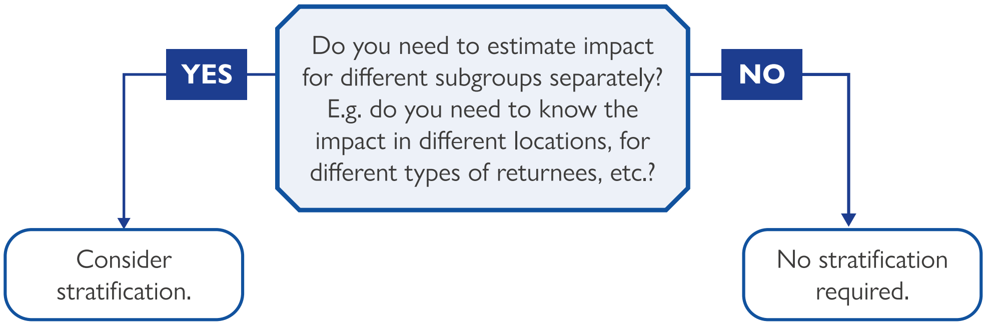
The effect of the COVID-19 pandemic meant that the IMPACT study baseline was not conducted as planned, and retrospective enumeration was used instead.

An endline questionnaire was conducted which also asked respondents to recall what their situation had been at the time of baseline for each of the questions on the RSS.

IMPACT team response:

As has been mentioned in other modules, retrospective enumeration is an approach which is still being explored at the time this course is being developed. There are risks in terms of the reliability of the recalled information and the potential for bias.

However, it offers the option to create a baseline where this would otherwise have been impossible, so may be an appropriate choice if used with awareness of its limitations.


Do you need to estimate impact for different subgroups separately? E.g. do you need to know the impact in different locations, for different types of returnees, etc.?
YES
NO
Consider stratification.
No stratification required.

Question

Look again at Figure 45 (the relevant portion is repeated above for reference). Do you need to estimate impact for different subgroups separately? Which option would you choose?

- □ Yes
- □ No

Answer

Do you need to estimate impact for different subgroups separately?

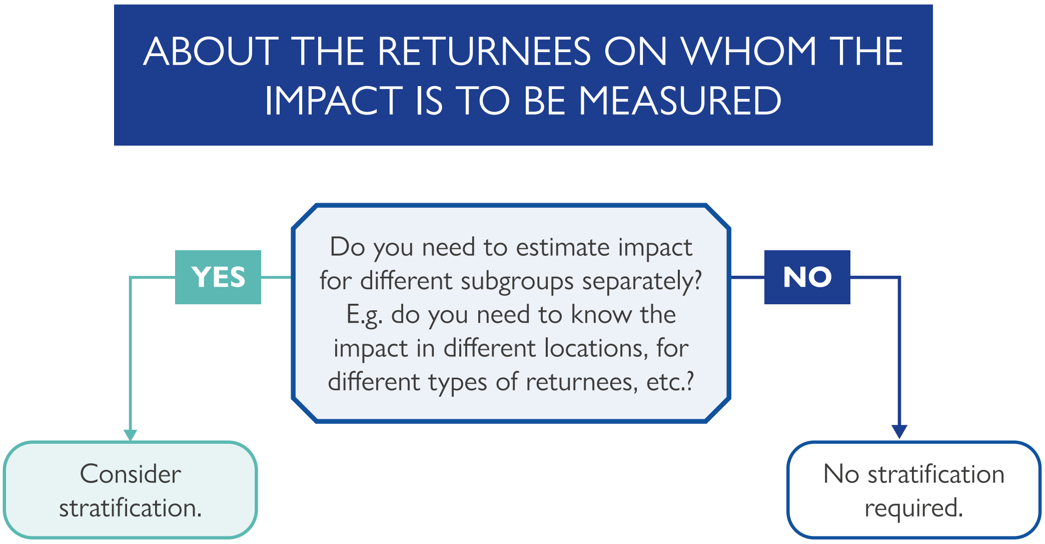
☑ **Yes**

□ **No**

This will be necessary if the IMPACT study is to accomplish its aim of categorizing and comparing different types of returnees over time. Separate results are also needed for each country included in the impact evaluation. This requires stratification.

However, this option adds complexity and expense to the sampling and data-collection process, so it may be sensible to choose not to use multiple strata – or to minimize the number used – in order to prioritize the more vital objectives of the study.

FIGURE 50: DECISION TREE FOR IMPACT STUDY DESIGN – ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED (PART 1)



ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED

Do you need to estimate impact for different subgroups separately? E.g. do you need to know the impact in different locations, for different types of returnees, etc.?

Consider stratification.

No stratification required.

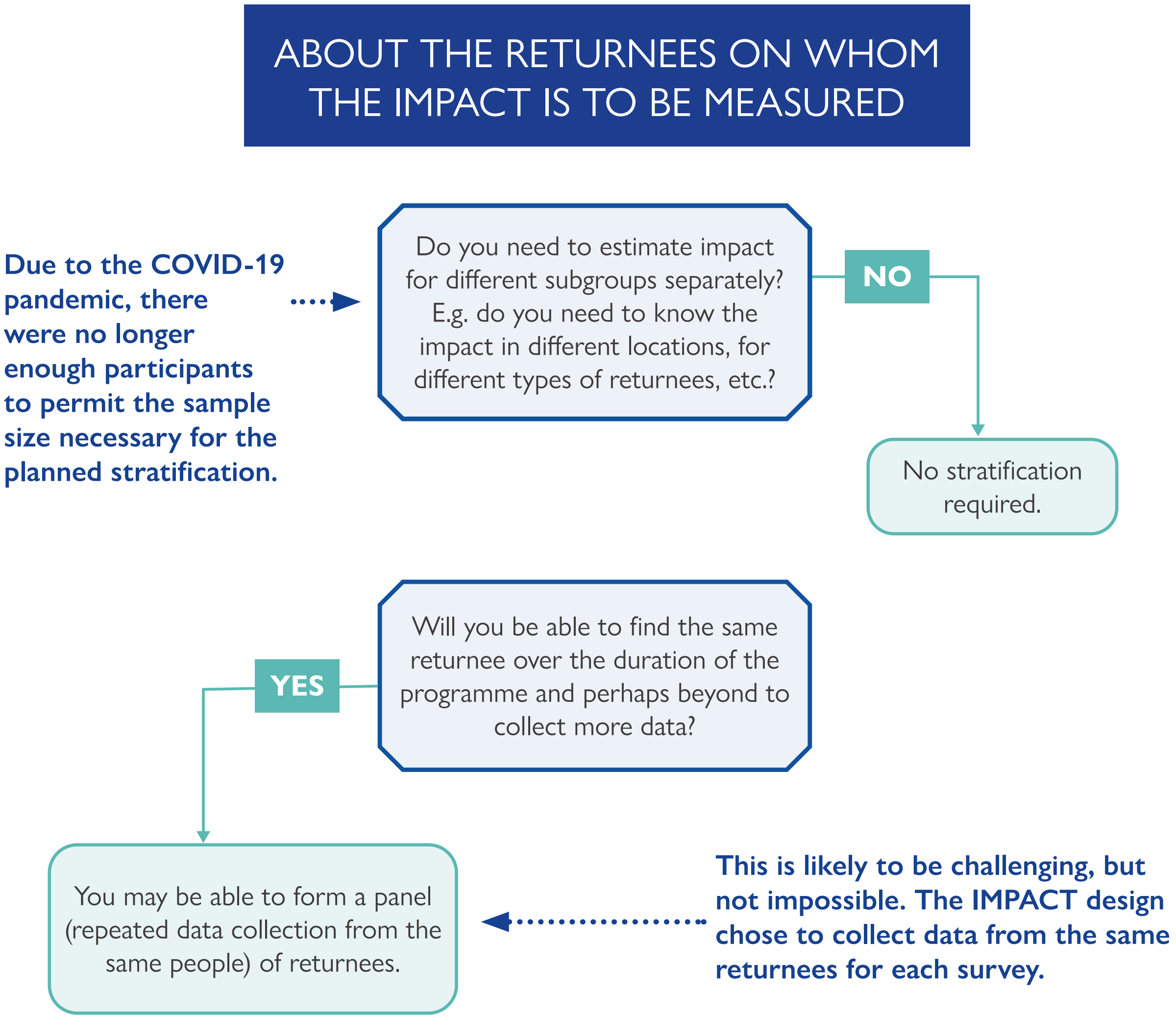
IMPACT team response:

The data collection for the IMPACT study is stratified by country.

The study also aims to be able to compare different types of returnees over time – for example: support levels provided since return, or length of time before receiving assistance. To do this, data needs to be separated according to these types. The design included stratification by some of the above factors, despite the expense involved.

However, this didn't end up happening! The programme roll-out was heavily disrupted by the COVID-19 pandemic, which prevented many people from participating. This means that there were no longer enough participants to permit the sample size necessary for this stratification. Rapid changes had to be made for the entire programme and evaluation:

FIGURE 51: DECISION TREE FOR IMPACT STUDY DESIGN – ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED (PART 2)



ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED

Due to the COVID-19 pandemic, there were no longer enough participants to permit the sample size necessary for the planned stratification.

Do you need to estimate impact for different subgroups separately? E.g. do you need to know the impact in different locations, for different types of returnees, etc.?

NO

No stratification required.

Will you be able to find the same returnee over the duration of the programme and perhaps beyond to collect more data?

YES

You may be able to form a panel (repeated data collection from the same people) of returnees.

This is likely to be challenging, but not impossible. The IMPACT design chose to collect data from the same returnees for each survey.

Question

Why might the IMPACT study use a panel data-collection approach in its design, despite this being a difficult option to implement in the context? Choose the option you think is the most important factor in making this decision.

- □ It is even more challenging to use a cross-sectional approach as it is difficult to motivate people to participate in the survey.
- □ A panel study can produce more helpful data, as the actual change over time can be measured for observational units (in this case individual returnees), rather than the average change of a sample.
- □ A panel study is helpful if the study will use a matched design.

The answer is on the next page.

Answer

Why might the IMPACT study use a panel data-collection approach in its design, despite this being a difficult option to implement in the context?

□ **It is even more challenging to use a cross-sectional approach as it is difficult to motivate people to participate in the survey.**
This is not generally true. While motivating participation can be a challenge in some circumstances, panel studies are typically the more challenging option.

☑ **A panel study can produce more helpful data, as the actual change over time can be measured for observational units (in this case individual returnees), rather than the average change of a sample.**
This was indeed the primary reason why the panel approach was used.

□ **A panel study is helpful if the study will use a matched design.**
This was not the priority for the IMPACT study, but it is worth considering if there are strong reasons why it is necessary to use matching in your design. Using the same units for each round of data collection means the process of finding matching only has to be performed once, at the start.

IMPACT team response:

The reasoning behind this choice was that panel studies can produce some very helpful information; the ability to track changes in particular individuals allows insights that might not otherwise be possible.

For example, when using a panel study, it is possible to use the baseline state as a variable, such as by looking at whether returnees with high levels of debt at baseline experienced different impacts than those who did not.

The complexity of return and reintegration situations creates significant challenges for impact evaluation design. Often, a good solution is to adopt a combination of methods (see page 174 for more information). Designing an impact evaluation, in practice, requires creative problem-solving and the ability to find a balance between competing priorities.

WHAT NEXT?

What we have at this point is the foundation of an impact evaluation design (from a primarily quantitative perspective).

This is the starting point upon which the specific requirements for the project can be built and which might be suitable to include in calls for proposals to work with external evaluators and specialists.

INTERVIEW TRANSCRIPT: DESIGNING THE IMPACT STUDY

In this interview, Davide Bruscoli, IOM Regional Information Management Officer and technical coordinator for the IMPACT Study, gives some background and explanations of the decisions made when planning the study.

DESIGNING THE "IMPACT" STUDY

Interview with Davide Bruscoli, Regional Information Management Officer, IOM

Hello. My name is Davide Bruscoli. I am an applied economist by background. I have worked for IOM since 2019, and since then I've been developing a research and evaluation portfolio for the Joint Initiative programme. The centrepiece of this portfolio is of course the IMPACT study for which I serve as a technical coordinator on the side of IOM. In this short video, I will be providing some background and additional explanations on the decisions we made in the context of the IMPACT study.

Random allocation of treatment

Following the decision tree, let's start from the reason why random allocation of the treatment was impossible. It was mentioned in the course material that this was mostly due to ethical considerations. The caseload of the Joint Initiative programme is a very vulnerable one. The programme itself was designed to assist the most vulnerable returning migrants. We are talking about returnees who in the vast majority of cases did not manage to reach their intended destination and [who] found themselves stranded during the journey, often in very precarious conditions.

Of course, randomized control trials are the gold standard. We looked hard into the possibility of implementing a pure experimental design but it was almost immediately clear, based also on discussions with the with programme staff, that it wouldn't have been in any way justifiable to deny assistance to some returnees just for the sake of the evaluation.

The importance of ethics in impact evaluations

Ethics is extremely important in the context of impact evaluations, especially when the programme or project being evaluated assists vulnerable individuals. If programmatic decisions on who gets assistance, who doesn't, how assistance is provided, when assistance is provided, if this sort of decisions are taken also based on the need of implementing a certain design for the impact evaluation then it becomes extremely important – if not even mandatory – to subject your study design to thorough ethics review. Very often you have to pay for those reviews so make sure that you have some budget for it. Those reviews also take time so make sure you consider the time that is needed to complete them in the timelines of your impact evaluation.

Non-random allocation of the treatment

Having ruled out the possibility of implementing a pure experimental design, we moved to look at possible comparison groups based on a non-random allocation of the treatment. For example, within the Joint Initiative programme there is a group of beneficiaries who don't get all of the assistance that is available because they are not deemed as vulnerable enough. This is a problematic comparison group because we would end up comparing more vulnerable returnees in the treatment group with systematically less vulnerable returnees in the comparison group. Another thing we looked into is whether we could compare Joint Initiative returnees with the beneficiaries of other reintegration programmes. This as well was deemed as problematic because there is a very defined geographical specialization for those reintegration programmes. Very often, the other programmes being implemented at the same time of the Joint Initiative programme were focusing on returnees coming from different geographical areas than the Joint Initiative. Very often those were returnees who did manage to reach their intended destination, so in a way the beneficiaries of the other programmes were fundamentally different from Joint Initiative returnees.

Stepped-wedge design and naturally occurring delays

The so-called stepped-wedge approach presents an important ethical advantage over pure randomization, and this is the fact that everybody will receive the treatment, everybody will be assisted. What you manipulate as an experimenter is just the timing of assistance provision. We considered this particular design for the IMPACT study but the same ethical concerns that were raised for pure randomization applied as well to the stepped-wedge approach: it was not deemed as justifiable to delay the provision of support on purpose, because of the levels of vulnerability of the caseload that was being assisted. Even though a pure stepped-wedge approach was not possible, we are still making comparisons between different groups of returnees, but we are doing it by exploiting naturally occurring delays that affect the programme. At the time of the endline some beneficiaries are still to receive reintegration assistance and this gives us the opportunity to compare the outcomes of those who received assistance before the endline with those who are still waiting.

Comparisons with non-beneficiaries

At this point, we are left with one option, which is comparing our treatment group with non-beneficiaries. For the IMPACT study, we decided that we would compare the treatment group with non-migrants, so people with no previous experience of migration who could be found in the same communities where our beneficiaries were living. Comparing migrants with non-migrants is indeed a bit problematic. We even call this group with which we compare our treatment group a "calibration" group. We don't call it a "comparison" group, and this is to emphasize that we are very far from an experimental ideal. That the conclusions we can draw from those comparisons are much weaker than what we could do in a randomized control trial.

The value of comparing returning migrants with host community members

Even though not ideal from an experimental point of view, comparing returning migrants in the treatment group with non-migrants in the calibration group can be very useful to understand sustainable integration better. Those comparisons are not done very often, at least in a systematic way, and in our specific case they address one of the limitations of the measurement framework we use: the Reintegration Sustainability Index. This index defines sustainable reintegration as an absolute concept, without any relativity to it. So, to decide whether you are sustainably integrated or not, it doesn't matter whether you are better off or worse off than the people around you. Those comparisons allow us to start filling this gap. They are a reality check on the thresholds we are using and they also allow us to understand the differences between returning migrants and host community members in terms of what among all the indicators we have in the index is most important to predict reintegration.

TRY IT YOURSELF

Now that you have seen the how the decision-making process works to design an impact evaluation, it is your turn to try working through the decision trees yourself. Here is another scenario, based on a real programme.

An initiative is being put in place to support and follow-up with returnees in their communities of origin. Alongside other forms of reintegration assistance, the initiative will pilot a mentoring approach in which local community members (often returning migrants themselves) are trained as mentors. They will meet regularly with returnees to assist them with their reintegration plan, offer advice and encouragement and refer them to other services where appropriate.

The initiative is aimed at facilitating the social and psychosocial reintegration of returnees by helping them strengthen their relations in the community and improve their well-being. It is expected that this approach will improve the reintegration of returnees into their communities. The programme theory and evaluation questions have been defined for this initiative.

In this programme, while the economic dimension is the foundation of the support, social and psychosocial assistance are essential to consolidate reintegration gains.

The intervention has been established as a learning opportunity for the organizations supporting reintegration to determine the extent to which the mentoring intervention contributes to reintegration. An impact evaluation is to be carried out to achieve this aim.

The returnees:

- The evaluation team have access to information about each returnee who arrives back in the country. Returnees are expected to return to their original communities.
- For practical reasons and to inform decision-making, communities were selected into the study area to represent three different settings: rural, peri-urban and urban. There is an interest in seeing if there is a difference in impacts between these settings.
- It is difficult to establish what the situation of each returnee was before migration was attempted. Different options may be considered, but expert opinion suggests that interviews about past conditions of each returnee are unlikely to yield reliable data.
- The evaluation team does not have access to data or addresses of individuals who did not migrate and remained in the communities.

Mentoring approach:

- From a practical point of view, it is not possible to assign a mentor to all the returnees, and some way to decide who will receive mentorship will be needed.
- It is expected that 10 mentors will be recruited and trained and that each mentor can only work with up to 20 returnees during the period of the intervention. All mentors will receive training at the same time. When they finish their training, they will be contracted to work for 12 months. The mentoring support is expected to last for six months for each returnee.
- During the period of the pilot, it is expected that around 1,200 returnees will come back to their country.
- Whether or not they receive the mentoring, all returnees are given some kind of reintegration support from the point of their arrival.

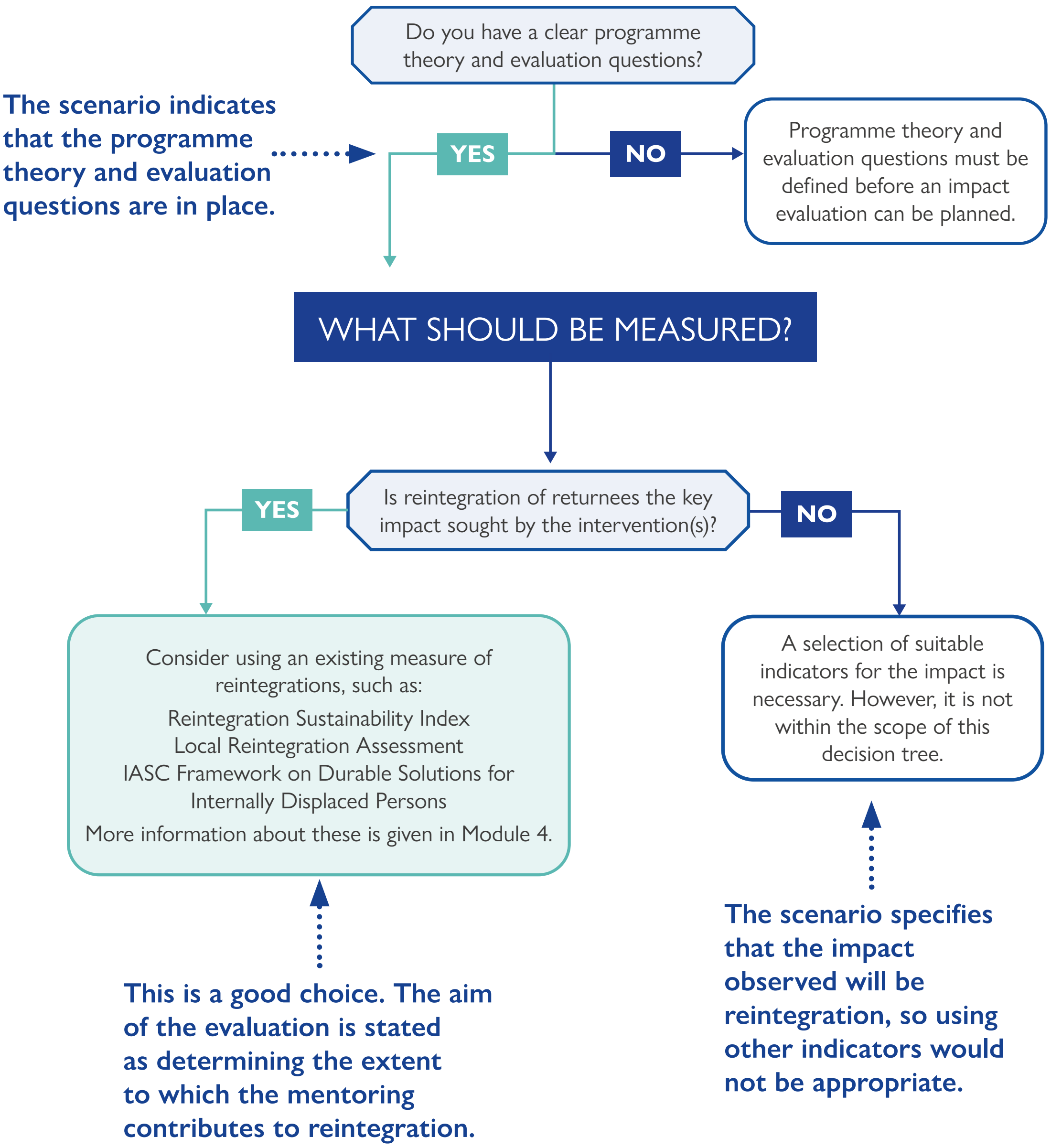
Your task: design the impact evaluation.

Read through the scenario above carefully, then refer back to the trees on page 190. Work your way through the trees and select the options you think are most appropriate for an impact evaluation based on the scenario. You may wish to print out or copy the trees. When you think you have a set of decisions that comprise a suitable impact evaluation design, look through the suggested solution and guidance given on the next page.

Suggested solution

Below is one appropriate set of options (highlighted in green) for an impact evaluation design that would be appropriate in the given scenario. It is not the only option, so if the solution you chose does not match, it may still be a suitable impact evaluation design. Read the guidance given for each option to evaluate the suitability of your design.

FIGURE 52: SUGGESTED SOLUTION DECISION TREE – WHAT SHOULD BE MEASURED?



Do you have a clear programme theory and evaluation questions?

YES

NO

Programme theory and evaluation questions must be defined before an impact evaluation can be planned.

The scenario indicates that the programme theory and evaluation questions are in place.

WHAT SHOULD BE MEASURED?

Is reintegration of returnees the key impact sought by the intervention(s)?

YES

NO

Consider using an existing measure of reintegrations, such as:

Reintegration Sustainability Index
Local Reintegration Assessment
IASC Framework on Durable Solutions for Internally Displaced Persons

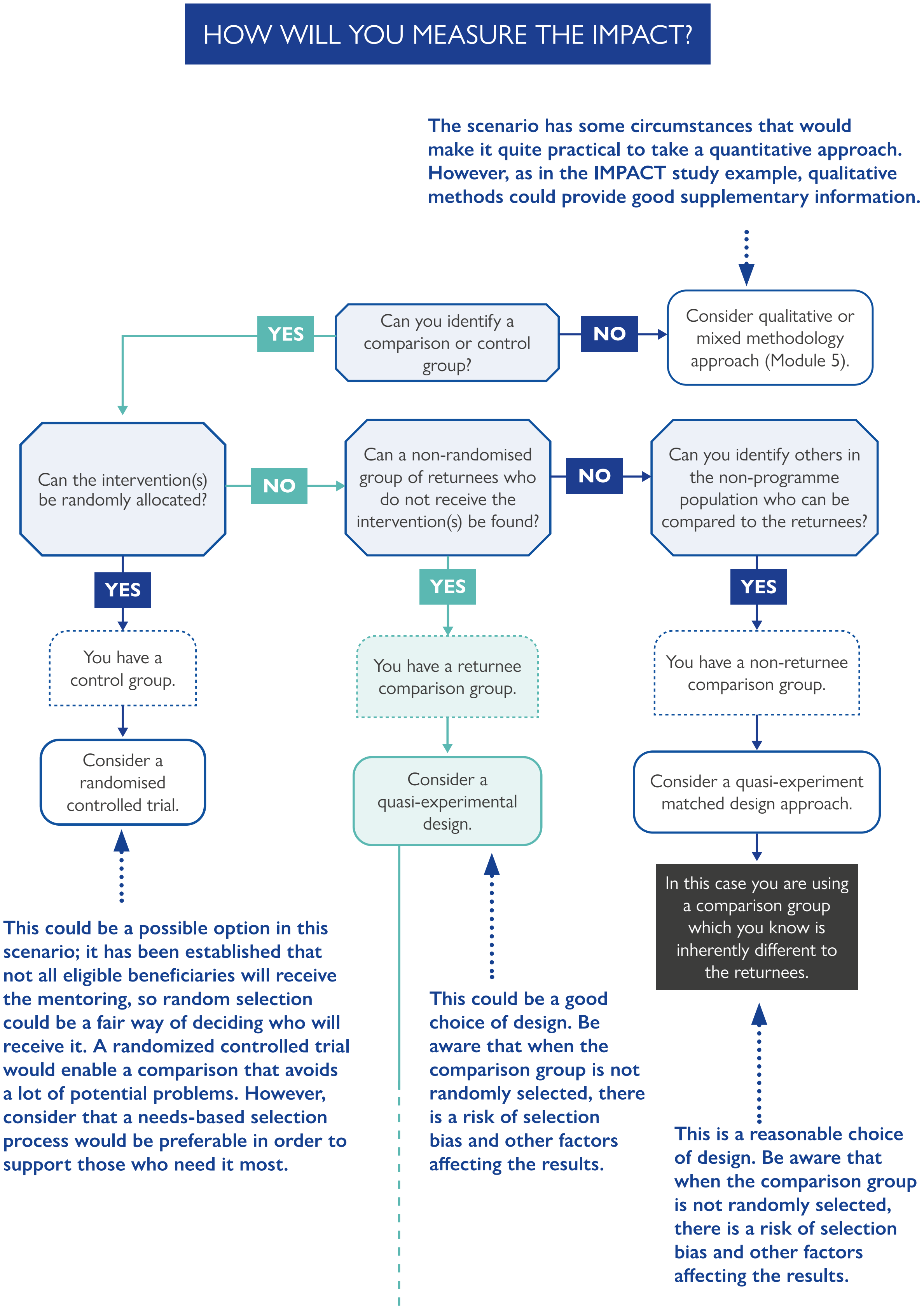
More information about these is given in Module 4.

A selection of suitable indicators for the impact is necessary. However, it is not within the scope of this decision tree.

This is a good choice. The aim of the evaluation is stated as determining the extent to which the mentoring contributes to reintegration.

The scenario specifies that the impact observed will be reintegration, so using other indicators would not be appropriate.

FIGURE 53: SUGGESTED SOLUTION DECISION TREE – HOW WILL YOU MEASURE THE IMPACT?



HOW WILL YOU MEASURE THE IMPACT?

The scenario has some circumstances that would make it quite practical to take a quantitative approach. However, as in the IMPACT study example, qualitative methods could provide good supplementary information.

Can you identify a comparison or control group?

NO → Consider qualitative or mixed methodology approach (Module 5).

YES → Can the intervention(s) be randomly allocated?

YES → You have a control group. → Consider a randomised controlled trial.

This could be a possible option in this scenario; it has been established that not all eligible beneficiaries will receive the mentoring, so random selection could be a fair way of deciding who will receive it. A randomized controlled trial would enable a comparison that avoids a lot of potential problems. However, consider that a needs-based selection process would be preferable in order to support those who need it most.

NO → Can a non-randomised group of returnees who do not receive the intervention(s) be found?

YES → You have a returnee comparison group. → Consider a quasi-experimental design.

This could be a good choice of design. Be aware that when the comparison group is not randomly selected, there is a risk of selection bias and other factors affecting the results.

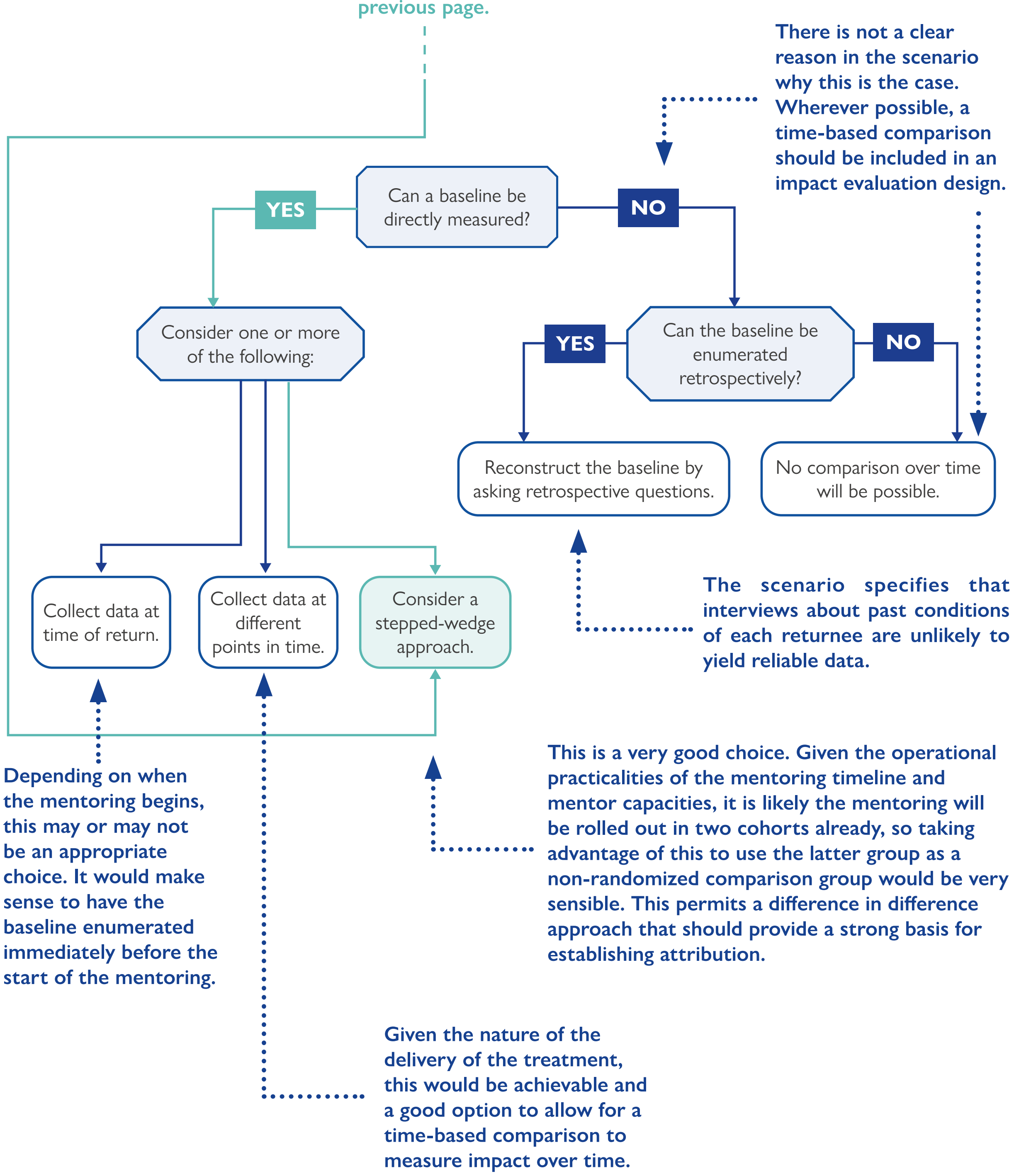
NO → Can you identify others in the non-programme population who can be compared to the returnees?

YES → You have a non-returnee comparison group. → Consider a quasi-experiment matched design approach. → In this case you are using a comparison group which you know is inherently different to the returnees.

This is a reasonable choice of design. Be aware that when the comparison group is not randomly selected, there is a risk of selection bias and other factors affecting the results.

Continued on the next page.

Continued from previous page.


Can a baseline be directly measured?
YES
NO
Consider one or more of the following:
Can the baseline be enumerated retrospectively?
YES
NO
Reconstruct the baseline by asking retrospective questions.
No comparison over time will be possible.
Collect data at time of return.
Collect data at different points in time.
Consider a stepped-wedge approach.

There is not a clear reason in the scenario why this is the case. Wherever possible, a time-based comparison should be included in an impact evaluation design.

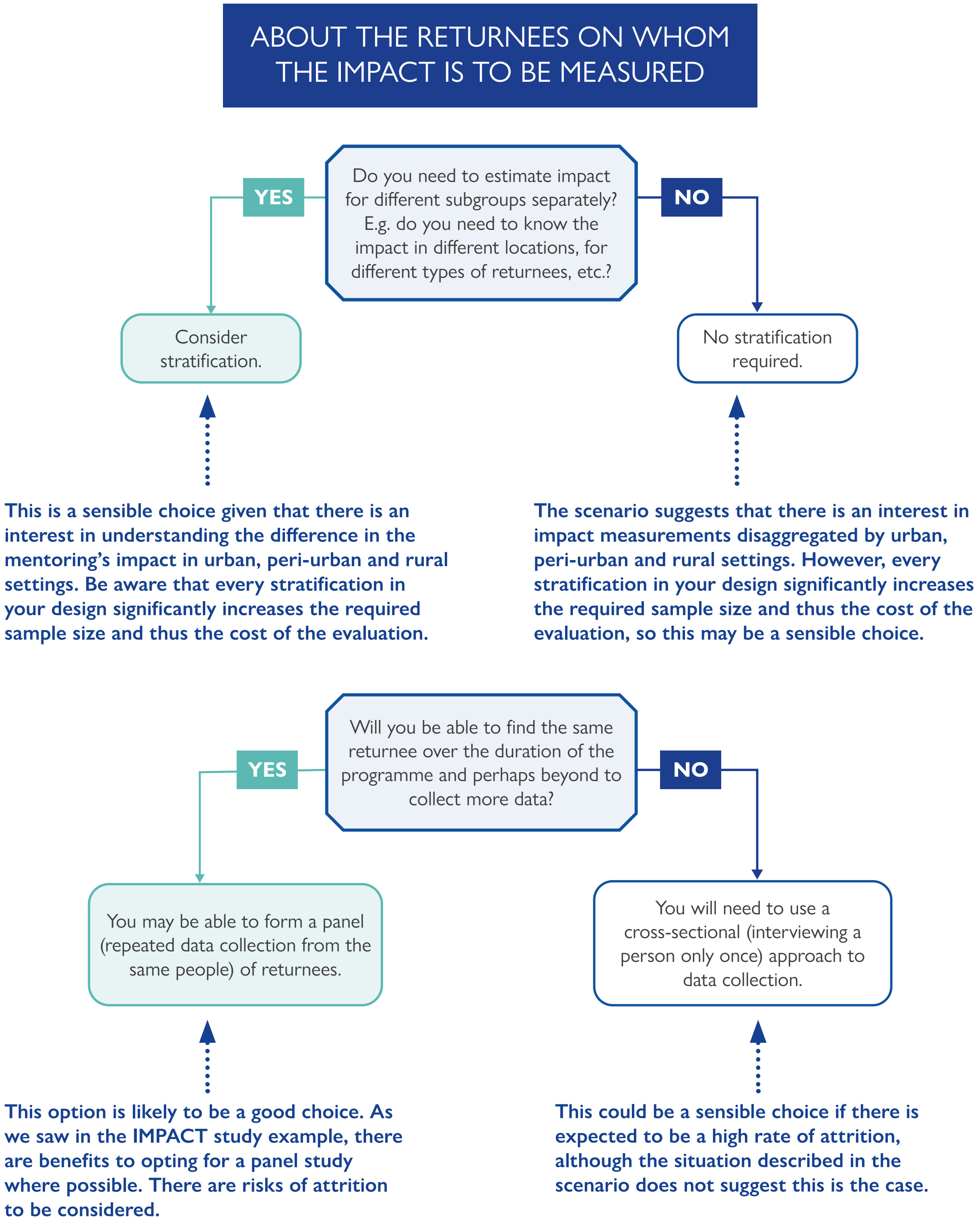
The scenario specifies that interviews about past conditions of each returnee are unlikely to yield reliable data.

Depending on when the mentoring begins, this may or may not be an appropriate choice. It would make sense to have the baseline enumerated immediately before the start of the mentoring.

This is a very good choice. Given the operational practicalities of the mentoring timeline and mentor capacities, it is likely the mentoring will be rolled out in two cohorts already, so taking advantage of this to use the latter group as a non-randomized comparison group would be very sensible. This permits a difference in difference approach that should provide a strong basis for establishing attribution.

Given the nature of the delivery of the treatment, this would be achievable and a good option to allow for a time-based comparison to measure impact over time.

FIGURE 54: SUGGESTED SOLUTION DECISION TREE – ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED



ABOUT THE RETURNEES ON WHOM THE IMPACT IS TO BE MEASURED

Do you need to estimate impact for different subgroups separately? E.g. do you need to know the impact in different locations, for different types of returnees, etc.?

YES

Consider stratification.

This is a sensible choice given that there is an interest in understanding the difference in the mentoring's impact in urban, peri-urban and rural settings. Be aware that every stratification in your design significantly increases the required sample size and thus the cost of the evaluation.

NO

No stratification required.

The scenario suggests that there is an interest in impact measurements disaggregated by urban, peri-urban and rural settings. However, every stratification in your design significantly increases the required sample size and thus the cost of the evaluation, so this may be a sensible choice.

Will you be able to find the same returnee over the duration of the programme and perhaps beyond to collect more data?

YES

You may be able to form a panel (repeated data collection from the same people) of returnees.

This option is likely to be a good choice. As we saw in the IMPACT study example, there are benefits to opting for a panel study where possible. There are risks of attrition to be considered.

NO

You will need to use a cross-sectional (interviewing a person only once) approach to data collection.

This could be a sensible choice if there is expected to be a high rate of attrition, although the situation described in the scenario does not suggest this is the case.

CONCLUSION

You have completed this module.

You may wish to print or save a copy of the decision trees starting on page 190 for future reference. It could be helpful as a checklist or reference tool for thinking about the design of impact evaluations.

MODULE 7: EXTENSIONS – GOING BEYOND THIS COURSE

Khartoum, the Sudan. As part of community reintegration under the EU-IOM Joint Initiative, IOM partnered with a local non-governmental organization called Rural Community Development Organization to rehabilitate a multipurpose community centre. © IOM 2021/ Muse MOHAMMED

MODULE 7: EXTENSIONS – GOING BEYOND THIS COURSE

INTRODUCTION

This module will introduce suggested "extension" topics relevant to the implementation of impact evaluation studies in a reintegration context, provide a basic overview and direct trainees to resources that will enable them to continue learning on these topics autonomously.

OUTCOMES

At the end of this module, trainees will be able to:

- Outline key concepts of sampling theory and practice relevant to impact evaluation.
- Give an overview of good practices for questionnaire design, including quality assurance, with reference to digital data collection.
- List some strategies that may be used to resolve implementation challenges such as longitudinal data collection on the same participants and the high cost of face-to-face panel household- or individual-level surveys, associated with conducting impact evaluations.
- Independently build on existing knowledge of topics relevant to the practical implementation of impact evaluations.

INTRODUCTION

This module is different from the other modules. It provides a starting point for trainees to undertake further independent reading and learn about topics of which it would be helpful to develop an understanding, although they are not covered by this course. These are areas related to implementing impact evaluation for return and reintegration programmes, which would be useful for anyone involved in such evaluations.

The selection of topics covered is not a complete set by any means. However, it is expected that they will be useful when implementing, commissioning or using the results of an impact evaluation.

In the following sections, a brief introduction to several topics will be given, along with an annotated list of recommended resources to learn more. The topics included are:

Resources are provided with a brief description, an indication of whether they are at a basic, intermediate or advanced level and – where relevant – specific page ranges or sections to look at.

SAMPLING THEORY AND PRACTICE

INTRODUCTION

A key component in the design of surveys, including those to assess impact, is the sampling decisions. These decisions also have very important consequences for the cost of the study or evaluation you are undertaking. The principles for making these decisions for impact assessment of return and reintegration programmes are the same as for other studies, although the terminology may vary.

Statisticians with expertise on sampling methods are of great help and should be consulted for complex surveys. However, they will need specific inputs from you as an evaluator or decision maker.

This section contains definitions and concepts that may help you interact with statisticians about sampling.

EXAMPLE: SAMPLING FOR AN IMPACT EVALUATION OF A RETURN AND REINTEGRATION PROGRAMME

In the following example, we will see how decisions about sampling strategies can quickly become quite complex in real life situations and why gaining a strong base of understanding of sampling concepts can therefore be beneficial to anyone involved in planning an impact evaluation.

Think back to the IMPACT study for the Joint Initiative programme, which we have discussed multiple times in previous modules (see page 9). Let's imagine we are planning the sampling strategy for the quantitative study portion of the impact evaluation. We want a representative sample of the returnees in each country, with a sufficient sample size to estimate the key indicators for the impact evaluation.

Suggested solution 1

Using the list of registered returnees of each country in the Migration Management Operational System Application (MiMOSA; the system that IOM uses to register beneficiaries of return and reintegration programmes), we can take a simple random sample of returnees in each country to evaluate the how the initiative supports their reintegration.

What is a simple random sample?

A simple random sample is when a subset (sample) is taken from a population at random. All members of the population must have the same chance of being chosen for the sample. This is an unbiased way to create a sample; there is no risk of any characteristics of the individuals affecting their chance of being selected

You can use a sample size calculator to explore what kind of sample might be needed from a population.

Sample size calculator: *www.qualtrics.com/blog/calculating-sample-size/*

At first glance, this strategy might seem fairly reasonable. It has the advantage that probability sampling (using a randomly selected sample) means the study can make estimates about the population sampled. Random selection is the best way to avoid bias in a sample.

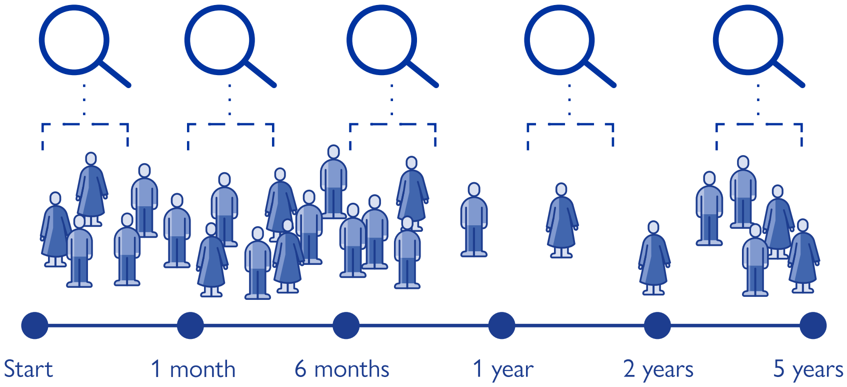
However, returnees do not arrive all at once; they will be continuously joining the programme and added to MiMOSA over the next three years. This means we can't draw the sample from the whole population using this method. We need a new approach that allows for how returnees arrive.

Suggested solution 2

Build the sample of returnees over time as returnees arrive, as illustrated in Figure 55.

This would mean taking a random sample from new returnees in the MiMOSA database in multiple rounds over the duration of the programme (e.g. monthly, quarterly or biannually – the decision for which interval is most appropriate is based on many factors, including the size of the flows, enumeration capacity, costs, etc.).

FIGURE 55: MULTIPLE ROUNDS OF SAMPLING OVER DURATION OF PROGRAMME



This method is more complicated and challenging to implement than the previous suggestion, but it allows you to capture a more diverse group of returnees, as returnees arriving at different times are being included, and it is a good solution to the problem of continuous intake of returnees.

However, there are some further issues to consider. For example, we are sampling from MiMOSA, which includes all the registered returnees. Some returnees are part of the same household and some colleagues of yours suggest that it would not be helpful to have more than one returnee from the same household in the sample since they share the same resources and having one of them is enough.

Another colleague reminds you that you should only interview beneficiaries that are old enough to give proper consent to participating in data collection.

Suggested solution 3

Establish a set of criteria for inclusion in the random sample.

For example, only returnees who are above 18 years of age and the "principal applicant" for their household will be sampled. While this solves some problems, moving away from the method of taking a purely random sample from the whole population could cause problems.

Can you think what problem(s) might result from applying selection criteria to the sampling? Take some time to note down your thoughts, then look at our suggestions on the next page.

Here are our suggestions:

- This means that the definition of the population represented by our sample has changed. It is no longer all returnees, but only those returnees that meet the criteria we are imposing to qualify for the sample. We need to ask, does this introduce biases that reduce the usefulness of our sample?
- There is a compromise to be considered here; this solution offers a resolution to some specific issues in exchange for new potential problems. An understanding of the concepts and theory of sampling is essential here to allow you to make informed decisions.

There are also other things to consider. Once returnees have been selected for the sample, the intention is that the team of interviewers will contact the returnees and run the questionnaire with them. However, contacting and getting responses is not simple. It is common for returnees to change phone number or be otherwise uncontactable. There will also be returnees who refuse to be interviewed.

Question

What problems could this lead to? Select all the answers that apply. The answers are given below.

□ Reduced sample size.

□ Risk of bias.

□ Contamination of results.

□ Ethical concerns.

Answer

What problems could this lead to? Select all the answers that apply.

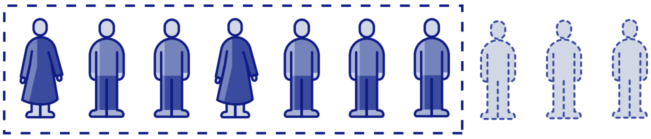
☑ **Reduced sample size.**
Yes, we can lose significant numbers from our intended sample from those who are sampled but from whom we cannot collect data. This is an issue when there is a minimum sample size to meet the information needs that guided the sampling design.

☑ **Risk of bias.**
There is indeed a risk of bias. There may be characteristics of returnees that make them less likely to participate or be consistently contactable. The decreased likelihood of returnees with these characteristics being included in the study introduces selection bias into the evaluation.

□ **Contamination of results.**
Although an important concern, contamination is not a risk associated with this scenario.

□ **Ethical concerns.**
There should not be ethical problems resulting from this situation, although it is worth reinforcing that returnees can only be included in the study if they have given their informed consent; if they do refuse to participate, this should be respected, regardless of the consequences for the sampling.

Regarding the potential for bias, this is an example of how the theoretical ideal (such as a purely random sample) is sometimes simply not reflected in reality. The trick is to understand what can be modified without affecting the ability of the sample to produce estimates that meet required standards. This is why it is helpful to work with statisticians, and it helps to be familiar with sampling concepts and terminology to engage with them.



In terms of the impact to sample size, one option is to start with a larger sample than is needed to account for loss through non-contact or refusal. This is also an advisable strategy to counteract attrition in a longitudinal study, or to account for issues when using matching to create a comparison group.

The planned sample size itself is a matter of negotiation. We need to consider the reality of the field, the available resources and the statistical criteria.

Determining sample size is a challenge that may benefit from including a statistician in the process. They will want to know about the margin of error for the estimates and about the expected variability of the key indicator – being familiar with these and other principles of sampling will be very useful in this process.

As you have seen in this scenario, it is clear why there is a good amount of flexibility and problem-solving needed in applying sampling theory to real-life return and reintegration scenarios. Anyone involved in planning an impact evaluation would benefit from having a good understanding of sampling, as this will inform decisions about the sampling design and be helpful when involving people from other fields, such as statisticians.

The information and resources presented in this section will support you to develop a good level of familiarity with relevant sampling concepts.

CONSIDERATIONS FOR A SAMPLING DESIGN

What do you need to provide as inputs for a sampling design? The following are relevant considerations:

- What are you measuring? For example, Reintegration Sustainability Index score.
- What changes do you expect to see? What is the programme aiming to achieve? This is also called “determining minimum observable effects”.
- What is the target population? For example, returnees in a specific location from which you want to take a sample of people or households.
- Can you assume that the population is large enough to ignore the need for “finite population adjustment” (see following section) in calculating the sample size?
- Are there groupings in the way the target population is distributed in the field?
- Are they “clustered” by settlements or towns or regions in such a way that you can take advantage of this to make fieldwork easier?
- And, if so, do you need to adjust for the potential effect of clustering of responses for sample size calculations?
- Is the target population grouped into types (e.g. male and female returnees, returning from different countries, etc.)?
- Do you have comparison or control groups, before and after observations, etc.?

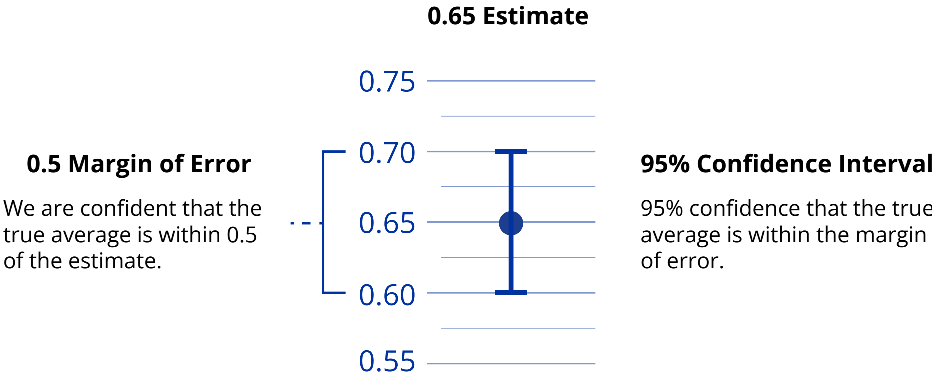
Most of the above has been covered to an extent in Modules 3 and 5. The next section contains some guidance and resources to expand on what you have already learned and discover how this applies to sampling.

Margin of error and confidence interval

When you conduct a survey using a sample from a population, the results provide you with estimates about the whole population. Estimates, by definition, have some room for error (Figure 56).

- **The margin of error for an estimate is a range within which you can be "confident" (see confidence level) that the "true value" lies.**
  For example, imagine you have an estimate of the average reintegration index score for a population, based on a random sample. The estimate is 65 per cent with a margin of error of 5 per cent. This means that you can be confident that, if you were to actually calculate the reintegration score for the entire population, the average would be somewhere within 5 per cent of 65 per cent – so between 60 per cent and 70 per cent.
- **The confidence level is how confident you are that your estimate is within the margin of error of the true value.**
  It is given as a percentage, which measures how much of the time you would get the same result if you were to repeatedly take a new sample of the population. A 95 per cent confidence level means that if you took 100 samples of the population, you would get a result within the margin of error of your estimate 95 out of 100 times.

FIGURE 56: MARGIN OF ERROR AND CONFIDENCE INTERVAL

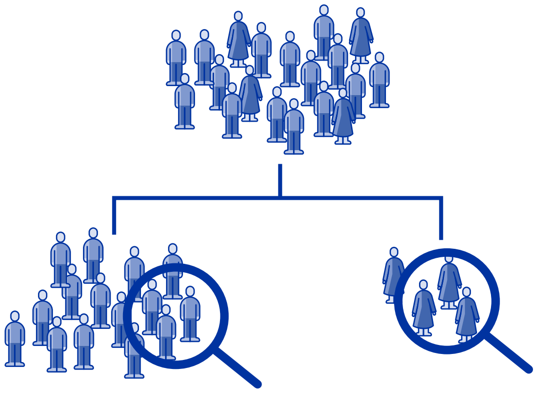


So if you make an estimate about your population based on a sample, you could not say, for example, "the population has an average reintegration score of 0.65", but you could say you are "95 per cent sure that the true average reintegration score for the population is between 0.6 and 0.7".

Reducing the margin of error or increasing the confidence level requires an increased sample size.

Stratification

As mentioned in Module 3 (page 66), it is sometimes useful to divide the population into subgroups, or strata, when sampling. This might be for the sake of reporting separate results for specific subgroups, or to ensure representation of subgroups that would otherwise be missed or underrepresented in a random sample of the whole population.

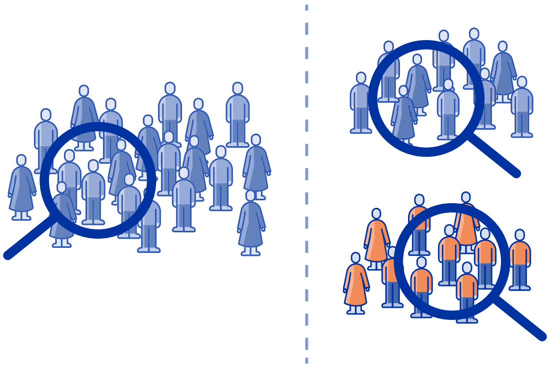


For example, in many return and reintegration programmes, most returnees tend to be men. In a random sample of returnees, women may not be represented. This can be solved by stratifying the target population into men and women, thus taking a separate sample for each gender. This might mean that women are then overrepresented, and analysis of the results would account for this using weights.

Stratification increases the overall sample size requirements for a study. If there is a need to report separate results for the strata, for example, then whatever sample size would have been needed from the whole population without stratification will be needed for each subpopulation, or the precision of the estimates will be reduced.

Comparisons

Simple random samples are used to make estimates about a population.



- When the aim is to compare a treatment and comparison group, you will need to take a sample for each group, and the calculation will be different from a single sample; in this case it will be for a two-sample comparison.
- Many impact evaluations make before/after comparisons.
    - The choice of whether to use a longitudinal or repeated cross-sectional study will affect the sample size required.
    - Longitudinal studies, for example, have less random variation, as the same units are used each time. This means that the basic sample size can be smaller than for repeated cross-sectional surveys – although remember that other adjustments may also be needed, such as increasing the sample size to allow for attrition.

Defining minimum observable effects

In statistics, this is about defining the smallest change (impact) that would be practically important and ensuring you survey enough people to allow you to observe this change in the resulting data if it has occurred. In impact evaluation, this should be thought of as a policy decision – what is the minimum change that the programme would consider as evidence that the intervention is having an effect, or that would be relevant to policy and decision makers?

For example, in a study that includes measuring the Reintegration Sustainability Index of returnees, it might be decided that the study needs to be able to detect changes to the average RSI of the sampled population of at least 5 per cent – this might be based on a combination of policy, budget and the expected impacts of the intervention being evaluated.

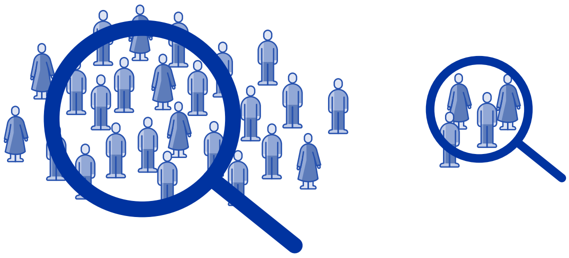
This means it is necessary to have a sample size that can provide a precise enough estimate to detect this change with the required confidence. If the sample size could only provide an estimate with a margin of 10 per cent then this would not be enough to be able to detect a 5 per cent change with any certainty.

The sample sizes needed for smaller margins of error and higher confidence levels can become very large, so a balance must be found between available resources and information needs. Once you have made this decision, you have a good part of the problem of sample size calculation solved and standard statistical theory can be used in combination with decisions about the sampling scheme.

Finite population adjustment to sample size

The standard calculations that are used to work out sample sizes assume there is an infinite population to sample from. When working with a relatively small population, the standard calculations might give a larger sample than is needed – or even a sample size that is larger than the available population

Imagine that a project has only 300 beneficiaries. What do you do if the recommended sample size is 350? In those cases, the calculation needs to be adjusted using a finite population correction factor.



Compensating for potential clustering of respondents

Sample size calculations generally assume that all observations are independent – i.e. a measurement taken from one sampled unit is not related to a measurement taken from another.

However, if you interview households who have aspects in common (e.g. they live in the same community, are members of the same self-help group, etc.) then the answers they give are likely to be more correlated than households with nothing in common. For example, returnees in the same community have the same community leadership, the same access to medical and education services and so on.

We can adjust the sample size calculation for this correlation between observations, known as the "design effect".

RESOURCES FOR FURTHER EXPLORATION ABOUT SAMPLING

Sampling Guide for Displacement Situations & Practical Examples

Joint IDP Profiling Service
Level: Basic

The Joint IDP Profiling Service (JIPS) provides an essential toolkit for profiling IDPs (that contains useful information about sampling, including this guide).

- The introduction on pages 6–25 is a good starting point, but the whole document is worth exploring.
- The terminology list on page 7 is very good and relevant.

PDF: *www.jips.org/uploads/2020/05/JIPS-SamplingGuideForDisplacementSituations-June2020.pdf*

Designing Household Survey Samples: Practical Guidelines

United Nations, Department of Economic and Social Affairs, Statistics Division
Level: Advanced

The handbook's purpose is to include in one publication sample survey-design issues for convenient referral by practising national statisticians, researchers and analysts.

- Methodologically sound techniques grounded in statistical theory are presented, implying the use of probability sampling at each stage of the sample selection process.
- The handbook could be used as reference material for multiple topics.

PDF: *https://unstats.un.org/unsd/demographic/sources/surveys/Series_F98en.pdf*

Household Sample Surveys in Developing and Transition Countries | Chapter 6: Estimating components of design effects for use in sample design

UN Statistics Division (UNSD)
Level: Advanced

This chapter is part of UNSD's guidance for Survey Design and Implementation.

- It provides a detailed description of design effects and how their different components can be estimated.
- This is a very advanced text for those who have a desire to go deeper into the different elements of a design effect and how these might vary from survey to survey.

PDF: *https://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf*

International Household Survey Network

Level: Intermediate to Advanced

IHSN is a useful and rich source of information about sampling and surveys. The section of their website that provides guidelines covers topics that are important for designing and conducting surveys.

Web page: *www.ihsn.org*

Sampling Decision Assistant Tool

Developed for UNHCR by Statistics for Sustainable Development
Level: Basic to intermediate

This is a tool designed to help with the design of sampling schemes to provide estimates of the values of characteristics of a population.

- It guides the user through a sequence of decisions that eventually build a sampling scheme.
- It contains a series of "additional resources" – instructional materials including videos for each step in the process that provide an explanation about statistical issues that are of importance when designing a sampling scheme.
- The videos are also collected together here.

Web page: *https://stats4sd.org/resources/63*

Note on the use of size calculators:

There are plenty of sample size calculations available on the web. They are all very similar. Below is an example:

Web page: *www.calculator.net/sample-size-calculator.html*

The calculator uses the population size and asks for a confidence level and margin of error. The calculator will provide a sample size that would allow you to report results with the confidence level and margin of error you select.

When using sample size calculators, be aware that most of them assume a simple random sampling design. In the work of IOM with returnees, simple random samples are rarely used. The calculations from those calculators need to be adjusted to account for the complex sampling designs used in practice.

QUESTIONNAIRE AND QUESTION DESIGN FUNDAMENTALS

Questionnaire design is a complex process. It is time-consuming and requires considerable expertise.

It is very likely that well-tested questions (or modules of questions) already exist for that the data you require to fulfil your information needs. For example, IOM's Monitoring and Evaluation Tools for Return and Reintegration Programmes can be downloaded here:

🔗 Web page: *https://returnandreintegration.iom.int/en/resources/guideline/monitoring-and-evaluation-tools-return-and-reintegration-programmes*

An important recommendation for questionnaire design is: "Do not reinvent the wheel". Using questionnaires that are used by others also has the advantage that your data will be comparable with previous data collections or data collected in other locations. This is a big plus as we move from scattered and ephemeral data to working with large and integrated data sets.

The following references provide an overview of questionnaire design issues:

Household Sample Surveys in Developing and Transition Countries | Chapter 3: An overview of questionnaire design for household surveys in developing countries

United Nations Statistics Division (UNSD)
Level: Intermediate

Provides an overview of the questionnaire design process for household surveys in developing countries (pages 47–52). This chapter is part of their guidance for Survey Design and Implementation.

🔗 PDF: *https://unstats.un.org/unsd/hhsurveys/pdf/Chapter_3.pdf*

Capturing what matters: Essential Guide for Designing Household Surveys: LSMS Guidebooks Second Edition December 2021

World Bank Group

Level: Intermediate

This is an updated go-to publication for practitioners designing multitopic household surveys, typically for estimating poverty rates among other themes. There is a module particularly focused on migration (5.3.1 Migration-page 42), that can provide inspiration and guidance for designing surveys for measuring reintegration.

🔗 Web page: *www.worldbank.org/en/programs/lsms/publication/CapturingWhatMattersEssentialGuidelinesforDesigningHouseholdSurveys*

Statistical Guides: Guidelines for Planning

Statistical Services Centre, University of Reading

Level: Basic

A booklet on planning surveys for research. The guide covers planning, when a survey is appropriate to use, setting up a survey, designing questionnaires, sampling principles, quality assurance and survey data analysis.

🔗 Web page: *https://stats4sd.org/resources/404*

DIGITAL DATA COLLECTION

Conducting survey data collection for impact evaluation via digital means, sometimes referred to as computer-assisted personal interviewing (CAPI) and often done using smartphones or tablets, has many advantages:

- Removes the need for separate data entry after the survey.
- Can enhance the quality of the data being collected, by restricting the information enumerators can enter; this is sometimes called data validation.
- Reduces the time needed for data cleaning (also called data verification).
- If the digital tool for the survey is also linked to previous data on the same subject, this can be used to look in real time at changes in responses between surveys and query those (e.g. "you said No when we spoke to you last; what reasons are there for why your response changed to Yes now", or "last time we spoke your household was eight members, has this changed?").
- Allows for data monitoring, which is useful for monitoring enumerators and improving the quality of their interviews (e.g. by checking the duration, repetitiveness of responses).

Digital data collection also has disadvantages and contexts in which it may not be the most useful format:

- The use of digital collection does require longer and more intensive preparation and pretesting time than paper questionnaires to get the best quality tool. However, this is arguably offset by the increased quality and efficiency of using digital data collection.
- At least some electricity and data connectivity are required to ensure that the data can be uploaded as soon as possible and avoid data loss.
- Surveys which include a lot of qualitative information and discussions with people rather than a series of questions may not be best captured using this method.

Digital Data Collection – Opportunities and Challenges

Statistical Services Centre, University of Reading
Level: Intermediate

A collection of videos giving an introduction to how digital collection works and sharing perspectives on good practice for writing digital questionnaires and conducting interviews.

Web page: *https://stats4sd.org/collections/10*

A Comparison of CAPI and PAPI through a Randomized Field Experiment

Bet Caeyers (University of Oxford), Neil Chalmers (EDI), Joachim De Weerdt (EDI)
Level: Advanced

For a more in-depth look at the digital data collection versus the use of paper forms. A report of an interesting randomized survey experiment comparing pen-and-paper (PAPI) to computer-assisted personal interviewing (CAPI).

PDF: *https://documents1.worldbank.org/curated/en/467401588063959793/pdf/A-Comparison-of-CAPI-and-PAPI-through-a-Randomized-Field-Experiment.pdf*

TOOLS FOR DIGITAL DATA COLLECTION

Although there are many options available for digital data collection, at the time of preparing this course in 2022, a very popular open-source digital survey tool used widely in the development and humanitarian community, including IOM, is Open Data Kit (ODK). There are various platforms which support ODK tools and data collection, some open-source and some with licensed versions.

Many humanitarian and development actors use ODK Collect on Android mobile devices served by KoboToolbox. KoboToolbox provides access to two ODK servers: one for humanitarian organizations and one for researchers, aid workers and everyone else. IOM has its own KoboToolbox ODK server, which has the advantage that IOM can maintain its own data on its own servers, reducing the risk of unauthorized data access.

To create the questionnaire, there are ODK form builders available (KoboToolbox has one), but the most powerful way to create a survey form with all the data validation and automated metadata importing is to use XLSForm format in Excel for developing the questionnaire and then uploading to an ODK server (KoboToolbox or other ODK Central server), that then makes that form available to enumerators on their Android devices.

ODK Open data kit

Web page: *https://opendatakit.org*

KoboToolbox

Web page: *www.kobotoolbox.org*

> Designing a simple data-collection form in ODK is easy. Designing a complex data-collection form takes time, experience and lots of testing.
>
> Do not underestimate this process when you plan your data collection.

Alternatives to ODK include:

- Qualtrics
  www.qualtrics.com/uk/core-xm/survey-software
- Enketo (ODK-based)
  https://enketo.org
- Survey 123
  https://survey123.arcgis.com
- Redcap
  www.project-redcap.org
- Ona (ODK-based)
  https://ona.io/home/products/ona-data/features

Introduction to ODK

Statistics for Sustainable Development
Level: Basic to Intermediate

A collection of videos presenting the fundamentals of Open Data Kit (ODK) and mobile data collection, including guidance on writing XLSforms, using aggregators such as KoboToolbox and carrying out data collection using a mobile device.

A Spanish version is also available here.

Web page: *https://stats4sd.org/collections/29*

XLSForm.org

Level: Intermediate to Advanced

An online reference resource for designing an ODK form using XLSForm. This would be useful as a reference and to expand on the XLSForm basics presented in the "Introduction to ODK" series.

Web page: *https://xlsform.org/en/*

QUALITY ASSURANCE IN DIGITAL DATA COLLECTION

One of the benefits of digital data collection is an increase in the quality of the data collected and reduction of the data cleaning required later. This is because:

- **The tool can be designed to control the data being entered.**
  E.g. age of interviewee can only range from 16 to 100 years old, number of children going to school cannot be more than the number of school-age children in the household, etc.
- **You can use additional validation or triangulation.**
  E.g. "Your household size is eight if we add up all the people you mentioned earlier, is this correct?"

Capturing what matters: Essential Guide for Designing Household Surveys: LSMS Guidebooks Second Edition December 2021

LSMS and World Bank Group
Level: Basic

Chapter 7 (p.47–50)

Gives some practical guidance on implementing digital data collection and lightly covers quality assurance.

PDF: https://documents1.worldbank.org/curated/en/381751639456530686/pdf/Capturing-What-Matters-Essential-Guidelines-for-Designing-Household-Surveys.pdf#page=55

Introduction to ODK: Part 7 – Using KoboToolBox – Data Monitoring and Downloading

Statistics for Sustainable Development
Level: Intermediate

This video from the "Introduction to ODK" series mentioned previously introduces basic data monitoring and quality assurance through using KoboToolbox.

Web page: *https://stats4sd.org/resources/515*

Cousera – designing CAPI for data quality assurance

John Hopkins University
Level: Intermediate

Part of a free online course developed by John Hopkins University in the context of household surveys for programme evaluation in low and middle-income countries (LMICs). This video gives information about how quality assurance can be designed into digital data-collection questionnaires and implementation.

Video: *www.coursera.org/lecture/household-surveys-for-program-evaluation/designing-capi-for-data-quality-assurance-bFfEe*

Data Quality Checks

Mike Gibson, Poverty Action Lab
Level: Basic

This guide covers three kinds of checks for data quality assurance.

The use of back-checks and spot-checks, for example, where a field supervisor returns or calls an interviewee, refresher training or interview shadowing can all be used to improve and assure the quality of digital data collection.

Web page: *www.povertyactionlab.org/resource/data-quality-checks*

IMPORTANCE OF UNIQUE IDENTIFIERS

The use of a unique identifier for a programme or survey participant is vital to our ability to combine different data (e.g. activity participation, previous survey round data) on the same participant for impact evaluation.

For example, in IOM's global, web-based movement database, MiMOSA (Migrant Management Operational System Application), each migrant has a unique identifier, which uniquely identifies the beneficiary across multiple services, programmes and activities.

The use of unique identifier is also useful for anonymization of data.

Unique identifiers can have meaning, e.g. the code tells you the community in which the household is found, the ID is a person's national ID card number, etc., or be randomly generated, e.g. an ID created within an ODK form.

An identifier may not always refer to an individual person; depending on the data collection, it may be a household, for example.

> While it may seem obvious that unique identifiers are needed, in practice many data sets lack them, and this causes serious problems at the analysis stage.
>
> Make sure your data has appropriate unique identifiers.

Unique Identifier (UID): A Crucial Aspect of Survey Design

Humans of Data
Level: Basic

A short but useful summary of the importance of unique identifiers, the four basic ways to build a UID and the benefits of defining these early in the programme.

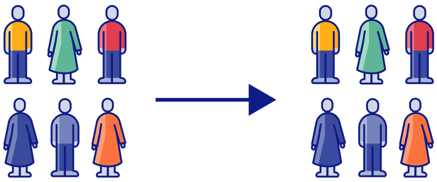
Blog post: *https://humansofdata.atlan.com/2017/08/unique-identifier-uid-survey-design/*

LONGITUDINAL/PANEL SURVEYS

LONGITUDINAL SURVEYS

We have introduced longitudinal surveys in previous modules (see page 71).

A longitudinal study involves measuring change over time by conducting multiple rounds of data collection with the same respondents, rather than using a new random sample each time. This has multiple advantages in terms of the useful information that can be gained, but can be challenging to implement successfully in return and reintegration contexts. This section aims to provide sources and examples to learn more about potential solutions to these challenges.



Challenges

The main challenge with longitudinal surveys is being able to repeatedly interview the same person or household multiple times. Your ability to do this may be reduced because of:

- Inability to contact returnees (e.g. returnees don't have own phone, change the number, etc.)
- The mobile nature of the survey population (e.g. returnees moving between communities, remigrating, etc.)
- Reluctance to be interviewed several times, especially if the questionnaire is long
- Survey participants dropping out of the overall programme (which could also be due to one of the reasons above).

Solutions

Below are some strategies for reducing or compensating for attrition in longitudinal studies:

- Other programmes and expert knowledge can be used to estimate attrition rates for longitudinal surveys and used to increase the sample size prior to the first round of survey – see page 72.
- Social media contacts (such as a Facebook profile) can be used rather than phone numbers as these tend to be more persistent.
- Later rounds of panel surveys can plan for the challenge of locating the same households by having a presurvey process to contact, request consent and book an appointment to interview the respondent. This stage could include multiple options to identify the respondent including their contact information, GPS coordinates from previous round of the survey and community leaders in their last known location.
- A "rolling-in rolling-out" design, where a proportion of the panel sample is retired and a new cohort recruited to maintain the overall sample size as the survey rounds progress. This tries to maintain the advantages of a panel observation while minimizing the downside of respondent fatigue. Normally, this approach is only taken with an unusually large number of taxing observations; e.g. quarterly observations over a number of years.

Using a Mobile App When Surveying Highly Mobile Populations: Panel Attrition, Consent, and Interviewer Effects in a Survey of Refugees

Jannes Jacobsen, Simon Kühne
Level: Advanced

An article investigating the use of mobile phone applications as a tool to reduce panel attrition among refugees. Useful to understand the strengths and weaknesses of this approach and the factors that make a difference to its successful implementation.

Article: *https://journals.sagepub.com/doi/full/10.1177/0894439320985250*

Reducing attrition in phone surveys

Berk Özler, P. Facundo Cuevas, World Bank
Level: Intermediate

This blog entry describes approaches taken to resolve issues with attrition in an evaluation of a refugee aid programme. This is a useful example of how small measures can make significant differences.

Blog post: *https://blogs.worldbank.org/impactevaluations/reducing-attrition-phone-surveys*

Paying respondents for participation in surveys

Sometimes, respondents are compensated for their time to encourage survey completion and future participation. This is a widely debated issue in terms of ethical considerations and the influence these may have on data quality.

Organizations usually provide some guidance to researchers under their institutional research ethics guidelines. However, the decision of what, if any, compensation should be paid to survey participants is highly contextual and relates to the length and timing of the survey and the survey participant's role or participation in the overall programme, if any.

While several organizations in the development and humanitarian sector (including IOM) do not encourage the practice of paying respondents for participation in surveys, this is an issue that arises frequently either as a way to encourage participation or because in the context where the work is done, previous data-collection operations have paid respondents.

Payment brings about issues of ethics and biases that need to be considered when deciding on whether or not paying respondents and how.

The following resources discuss these issues in more detail.

Compensation and Reimbursement of Research Participants

University of Toronto
Level: Basic

An example of an institution's guidelines on compensating respondents. Gives a good overview of the ethical concerns involved in this practice.

Web page: *https://research.utoronto.ca/compensation-reimbursement-research-participants*

If you pay your survey respondents, you might get a different answer

World Bank Blog, Marcus Goldstone
Level: Basic

This blog post discusses the potential for compensation of respondents to affect or bias responses. It highlights the fact that very little research has been done on this issue outside of developed country context.

He points to a 2017 paper, *Can incentives improve survey data quality in developing countries?: results from a field experiment in India* and while this paper is not in the public domain, Marcus Goldstone summarizes some of the findings, suggesting that it is only in the domain of asset and income that the compensating respondents report lower consumption.

🔗 Blog post: *https://blogs.worldbank.org/impactevaluations/if-you-pay-your-survey-respondents-you-just-might-get-different-answer*

🔗 Paper: *https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12333*

Retrospective enumeration

Retrospective enumeration has sometimes been used to recreate baseline data. For example, in the IMPACT study, the number of new returnees was much smaller than anticipated due to the COVID-19 pandemic, so returnees that arrived earlier were recruited for the evaluation and baseline data had to be retrospectively gathered.

IMPACT study:

🔗 PDF: *https://eastandhornofafrica.iom.int/sites/g/files/tmzbdl701/files/documents/iom_methodological_report_final_20102020.pdf#page=119*

This approach comes with some risk – while recent research has highlighted some cost-effectiveness advantages of retrospective data, there is also some uncertainty about the reliability of recall data.

While more research is needed, it is reasonable to think that the possible biases of retrospective data are specific to the context in which they are used. A detailed assessment of this, as well as a cost and benefit evaluation, should always be conducted if you plan on using retrospective data.

Measuring Once Twice

Jaspers, E., Lubbers, M., Graaf, N. D. D.
Level: Intermediate

This paper evaluates retrospective accounts compared with data gathered at the time to study the usefulness of recall data.

The evaluation was based on asking people about their attitudes in the past to euthanasia, homosexuality and the presence of migrants and thus is focused specifically on recall of attitudes.

🔗 PDF: *https://pure.rug.nl/ws/portalfiles/portal/2638676/JaspersE-Measuring-2009.pdf*

MARTIN SCHMITT: LONGITUDINAL STUDIES IN RETURN AND REINTEGRATION CONTEXTS

In this interview, Martin Schmitt, Regional M&E Officer at the IOM Regional Office in San José, Costa Rica, talks about attrition and difficulties contacting returnees in longitudinal studies and how these challenges can be overcome.

LONGITUDINAL STUDIES IN RETURN AND REINTEGRATION PROGRAMMES

Interview with Martin Schmitt, IOM Regional M&E Officer, San José, Costa Rica

In my former position as head of the research team at IOM in Germany, I coordinated a large-scale quantitative study with the returnees from Germany to 12 countries, asking about their return motivations, their first step to reintegration and also their potential remigration aspirations. And as we couldn't really rely on an experimental design and a study, but we still wanted to isolate somehow external factors and isolate a bit the effects of the programme itself and the reintegration, we decided to build a longitudinal study. And actually we did this, so the first one was 8 to 12 months after the return and then our second survey was held two to three years after their return. So, we could get much better insights into their journey and also see maybe who is dropping out, or who is still there to answer our questions.

Staying in contact with returnees

A very important issue, especially in longitudinal studies, is the challenge to stay in contact with persons – especially migrants are very mobile, by definition are very mobile populations. And, by this, they might – after the first round, they might change the phone number; this happens very often, not only directly after return but also then when they are in the country. For example, their group of friends are changing the phone number so they are all changing their phone company. And also, they might be changing the city where they are living, and it's hard to get in contact with them.

One way to reduce the issue of losing contact details is to is to collect more contact details. So not only a single source like the phone number, and the phone number is changing very often, so try to additionally collect some other data; for example, of messenger services of social media. Because these are usually things that do not change that often, and then you might be able to contact them when you are not getting a hold of them via phone.

It might also be an option to collect the phone number of a good friend or of the neighbour, and, by this, you might be able to reach this person even afterwards.

It definitely helped to make use of the expertise of the colleagues in the field. Some of them really knew better how to get in contact with those persons, and also colleagues in the field at times already had registered different phone numbers that we hadn't. When we could not reach them, we asked them, "Hey, could you please give us the phone number?" And then we connected this with our survey data, and then we had at times four different phone numbers for one person. And we just tried what worked.

But this, we did for another study where we just used, for example, Facebook Adwords to get in touch with persons that were assisted by IOM at any time, and then we just filtered afterwards. So, for this study right now, we are talking about, in the first round we had about 1,200 persons in the sample, and in the second round we still had over 900 so after two to three years. So this was pretty impressive, even for us, that we get so many persons in the sample!

Attrition in longitudinal surveys

Another point is the attrition, and attrition, I think, is something that is common to all panel or longitudinal studies, so you always will have attrition, and attrition only becomes an issue when it's systematic. So, when you have certain groups that are systematically dropping out of your sample.

And this can happen, especially in the field of return and reintegration, for example, that all the people that are leaving the country in the meantime, you probably won't have them in the sample afterwards. So, then when you're talking about their migration aspirations after their return, you won't get the whole picture as those that already left the country won't give you an opinion on that.

What might be a very good way to reduce attrition in longitudinal studies, or also in impact evaluations, where you have an ex-ante and ex-post element is to, where possible, combine programmatic aspects with your data collection. And, by this, you can also reduce a self-selection bias, for example. When persons were picking up their integration assistance, we handed out a survey and asked them to complete it, and by this we could reduce self-selection significantly, as also persons from remote places were coming to the IOM office to fill in the survey. And this can also be done when you are, for example, providing computers in the office so they can complete the online survey; you provide assistance in completing, especially for illiterate persons. Basically, it's all about reducing the barriers to participate in your survey.

PHONE SURVEYS

Phone surveys may be an attractive alternative to face-to-face interviews in challenging locations, for example those with low accessibility or security concerns. They are also cheaper to implement as they require no transportation costs.

Challenges

There are contexts in which these alternatives may be challenging or even not appropriate. For example:

- This option may not be suitable for long surveys (more than 20 minutes) and those with large qualitative elements requiring discussions.
- These types of surveys are especially challenging if there is no prior relationship or trust built between the enumerator and participant in earlier interactions.
- Similarly, surveys which ask sensitive questions (e.g. gender-based violence) may not be appropriate to give over the phone, although the counterargument to this is that privacy may be easier to achieve.
- Some contexts also present their own practical concerns, such as certain locations having unreliable phone connections, returnees changing their phone number or multiple neighbours sharing a phone.

Ethical aspects relating to prior informed consent and privacy should be specifically considered for phone surveys:

- When conducting surveys over the phone, it is not possible for the respondent to sign a form indicating their consent to participate. You may need to consider making an audio recording or obtaining consent ahead of time.
- In terms of privacy, consider that the respondent may be using a shared phone and will need to find somewhere private to take the call.

There are some risks of bias to consider when using phone or Internet-based surveys.

- Using a programme beneficiary telephone contact list may result in interviews for only the wealthier participants who have their own phone.
- Similarly, an Internet-based survey may only be completed by literate programme beneficiaries only.

There is also a concern about data quality; subtle body language and non-verbal cues are not observed. These can be useful for the interviewer to know whether the question is being understood and answered appropriately.

Best practices for conducting phone surveys

Abdul Latif Jameel Poverty Action Lab (J-PAL)
Level: Intermediate

This blog and webinar, written in the context of COVID-19, provides numerous resources and links to other references on designing and conducting phone surveys, including ethical aspects and reducing selection and participation bias.

Blog post: *www.povertyactionlab.org/blog/3-20-20/best-practices-conducting-phone-surveys*

Phone surveys in developing countries need an abundance of caution

Subha Mani, Bidisha Barooah, International Initiative for Impact Evaluation
Level: Basic

This 3ie blog provides some more details and cautionary tales on the use of mobile phone surveys in developing countries.

Blog post: *www.3ieimpact.org/blogs/phone-surveys-developing-countries-need-abundance-caution*

Reducing Bias in Phone Survey Samples

Alemayehu Ambel, Kevin McGee, Asmelash Tsegay, World Bank
Level: Advanced

This policy research working paper from the World Bank looks at the challenge of reducing bias in phone survey samples in four African countries.

The paper concludes that successfully contacted respondents in the four countries were biased towards wealthier households, resulting in an upward bias in estimates of well-being. If the sample was drawn from existing face-to-face representative surveys, application of survey weight adjustment can be done to remove most of the bias.

PDF: *https://openknowledge.worldbank.org/bitstream/handle/10986/35637/Reducing-Bias-in-Phone-Survey-Samples-Effectiveness-of-Reweighting-Techniques-Using-Face-to-Face-Surveys-as-Frames-in-Four-African-Countries.pdf;sequence=1*

Mobile Phone Surveys for Understanding COVID-19 Impacts: Part I Sampling and Mode

Kristen Himelein, Stephanie Eckman, Charles Lau, David McKenzie, World Bank
Level: Intermediate

This World Bank blog on phone surveys to understand the impacts of COVID-19 has some useful ideas for creating the sampling frame (from which to select respondents) and ideas for how to reduce the bias in selection of participations. They also describe different types of phone surveys including interactive voice response (IVR) and short message service (SMS) surveys as alternatives to interviewer-administered surveys.

Blog post: *https://blogs.worldbank.org/impactevaluations/mobile-phone-surveys-understanding-COVID-19-impacts-part-i-sampling-and-mode?CID=WBW_AL_BlogNotification_EN_EXT*

USE OF SECONDARY DATA

Secondary data can be a valuable source of information, especially to look back into what has happened in the past. It is crucial to be aware of what data are already available in the early planning stages of an impact evaluation.

The use of secondary data can complement primary data collection and can provide an alternative to collecting certain types of data; for example, information on something like health centre locations may already be available, or if there is another programme conducting a survey in the same time period then it may be possible to reuse some of the data.

Sources of secondary data which may be useful include:

- National-level representative surveys such as the World Bank Living Standards Measurement Study and national statistical bureaux
- National census data
- Data from other programmes operating in the return and reintegration locations.

Why You Should Consider Secondary Data Analysis for Your Next Study

Alchemer
Level: Basic

This page provides a concise review of the advantages and disadvantages of using secondary data.

Be aware that although this is a useful resource, it is not specifically focused on return and reintegration contexts.

Web page: *www.alchemer.com/resources/blog/secondary-data-analysis/*

GENERAL RESOURCES FOR IMPACT EVALUATION AND SURVEY DESIGN

The following resources are more general and would be helpful references or as further reading to deepen the understanding of Impact Evaluation gained in this course.

Better Evaluation

This site has become the "go-to" knowledge platform with information on more than 450 evaluation approaches, tasks, methods and processes and over 4,000 resources.

Web page: *www.betterevaluation.org*

International Initiative for Impact Evaluation (3ie)

3ie is an international initiative developing evidence and how to effectively transform the lives of poor in low- and middle-income countries. Established in 2008, they support the production synthesis and uptake of impact evaluation evidence in international development. 3ie work across the spectrum of actors from governments through to non-governmental organizations. Their site has a number of resources, including:

- How-to videos – where 3ie experts explain how to apply the theoretical concepts and evaluation designs. (Scroll down to the "3ie How-To videos" heading to find the series.)
- Video lecture series – this resource includes useful primer videos on what is an impact evaluation through RCTs, quasi experimental methods among others.

Web page: *www.3ieimpact.org*

International Household Survey Network – Guidelines

This section of the International Household Survey Network's website provides guidelines and best practices on all stages of survey implementation including:

- Designing survey programmes
- Creating survey budgets
- Implementing surveys
- Integrating surveys
- Archiving and dissemination of microdata.

Web page: *www.ihsn.org/guidelines*

World Bank Living Standards Measurement Study (LSMS)

This World Bank programme has set the standard for household surveys particularly for measuring poverty and well-being. They publish a series of guidebooks, none covering the topic of reintegration, but much of the advice and recommendations would be useful in applying best practice for sampling, questionnaire design and data quality assurance.

Web page: *www.worldbank.org/en/programmes/lsms*

Series of guidebooks: *www.worldbank.org/en/programmes/lsms/lsms-guidebooks*

United Nations Statistics Division (UNSD) Household Sample Surveys in Developing and Transition Countries

An in-depth resource on survey designs. It consists of five sections:

- Section A. Survey Design and Implementation
- Section B: Sample Design
- Section C: Non-Sampling Errors
- Section D: Survey Costs
- Section E: Analysis of Survey Data

🔗 Web page: https://unstats.un.org/unsd/hhsurveys/

CONCLUSION

You have completed this module.

You should now have the resources to scontinue learning autonomously about the suggested “extension” topics outlined in this chapter.

IOM
UN MIGRATION

17 route des Morillons, P.O. Box 17, 1211 Geneva 19, Switzerland
Tel.: +41 22 717 9111 • Fax: +41 22 798 6150
E-mail: hq@iom.int • Website: www.iom.int