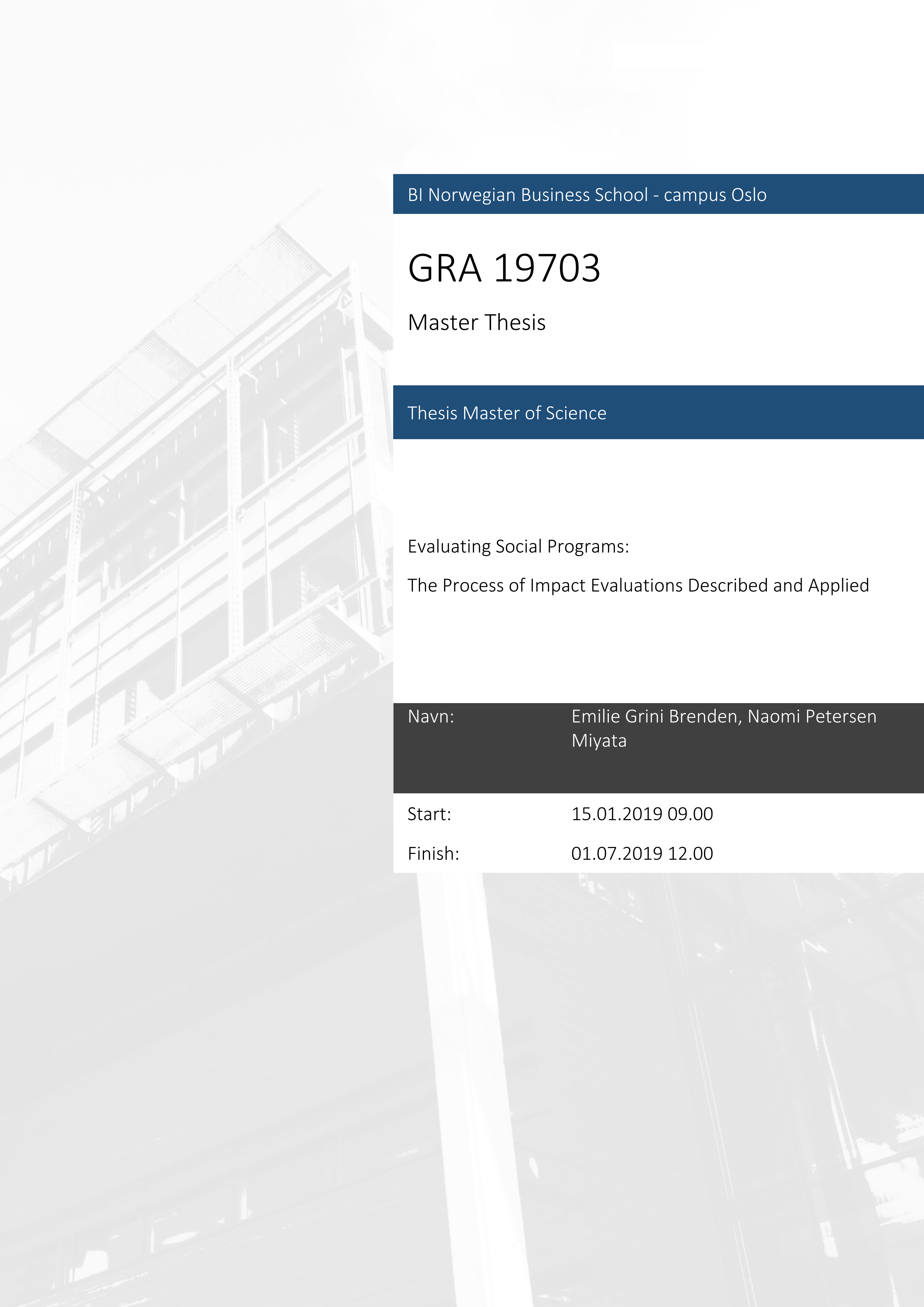

BI Norwegian Business School - campus Oslo

GRA 19703

Master Thesis

Thesis Master of Science

Evaluating Social Programs:

The Process of Impact Evaluations Described and Applied

Navn:	Emilie Grini Brenden, Naomi Petersen Miyata
Start:	15.01.2019 09.00
Finish:	01.07.2019 12.00

Evaluating Social Programs:

The Process of Impact Evaluations Described and Applied

Program:

Master of Science in Business - Major in Economics

Supervisor:

Jørgen Juel Andersen

This thesis is a part of the MSc program at BI Norwegian Business School. The school takes no responsibility for the methods used, results found, or conclusions drawn.

Abstract

There is an increasing demand for quality evaluations of aid projects to find out what works for development and what does not. The effect of specific aid projects should be measured through impact evaluations using robust methodologies. We believe that evidence from proper impact evaluations can help move the world towards better policy-making and poverty reduction. This thesis provides a framework for the process of conducting an impact evaluation from beginning to end. Our proposal is that the best way to design an aid project is to (1) include the steps of careful result-based monitoring in the impact evaluation process, and (2) make the impact evaluation a randomized controlled experiment if the circumstances allow for it. Randomization should ensure that the measured effects can be attributed to the project in question. To demonstrate the practical application of our framework, the framework is applied to an agricultural development project by Norwegian Church Aid aimed at smallholder farmers in Malawi.

Acknowledgement

Firstly, we would like to express our sincere gratitude towards our supervisor, Associate Professor Jørgen Juel Andersen for the continuous support of our work on this thesis, for always being positive, for the valuable inputs, and for giving us insight into his extensive knowledge of economic development. His counsel always had us inspired and his passion about the field is transmitting.

Our sincerest thanks also go to Norwegian Church Aid, Jakob Fagerland, and Johannes Ensby in particular, for trusting us with insights into their ongoing project and unpublished information. Without their cooperation, we would never have been able to write a thesis of this sort. It has been inspiring to witness firsthand the passion some people have for poverty eradication.

Last, but not least, we thank our families for their never-ending support, for their words of encouragement when days were dark, and for proofreading.

Table of contents

1. Introduction

Every year, huge sums of money are allocated to foreign aid by private institutions, individuals, and governments. In 2018 alone, Norway allocated 3.96 billion USD and the United States 33.0 billion USD to official development assistance (OECD Data). However, the lack of proper evaluations of aid projects makes it difficult to determine whether funds are spent in the best possible way. A Norad review of Norwegian aid projects found that over 65% of the projects reviewed did not build on sound methodological foundations (Chapman, Lloyd, Villanger, & Gleed, 2017). Too little light has been shed on the net impact of aid, and too many recommendations from social programs are based on insufficient evidence (Chapman et al., 2017; Dhaliwal & Tulloch, 2012, p. 2). Motivated by this, the objective of our master thesis is to answer the question:

"*What is the best way to evaluate the impact of a specific social program?"*

The term "social program" is for the purpose of this thesis used to encompass any targeted aid project or program. To measure the isolated effect of a social program and to *learn* from the experience, program managers need to conduct *impact evaluations.* Impact evaluations can also be conducted to measure the effect of new policies. If we are able to measure whether and how programs and policies are successfully achieving their goals, money can be distributed more effectively and ultimately be used to help more people.

In this thesis, we have prepared a framework which we believe can be useful for project managers with limited prior knowledge of impact evaluations of social programs. The framework touches upon the most important aspects of conducting an impact evaluation. However, due to the complexity of evaluating social programs, we by no means wish to imply that the framework is a complete account of the literature surrounding this subject. One can easily get lost in the jungle of detailed frameworks and complex econometric methodologies in existing literature. We believe that one of the most important contributions of this thesis is that it provides a relatively short and simple introduction to the process of an impact evaluation in its entirety.

As part of a complete impact evaluation, it is crucial to properly monitor the process over the course of a program. However, to be able to draw *causal claims* from the results of an evaluation, we argue that a social project should ideally be designed to be performed as a social *experiment* from the very beginning of the process. A properly designed and implemented social experiment should give us an accurate measure of the net impact of a specific initiative.

Over the last two decades, impact evaluations and the learning aspect of evaluating social programs has slowly gotten more attention (See, for example, White & Raitzer (2017) or Savedoff, Levine, and Birdsall (2006)). Quite a few organizations are already focusing on sharing knowledge around the importance of impact evaluations, amongst them Jamal Latif Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA), and the International Initiative for Impact Evaluations (3ie) (White & Raitzer, 2017, p. 150). J-PAL is a research center at the Massachusetts Institute of Technology (MIT). Their goal is to create a link between researchers and policymakers, and they especially argue for the use of randomized controlled trials (RCTs) in social programs (J-PAL, 2017, 2019b). The resources they provide on the subject of impact evaluations are conveyed in a simple and understandable way and has been an important building block for the formation of our thesis. Our framework is based on their online lectures and resources.

The thesis consists of three main parts. In the first part, we argue that it is important to properly evaluate social programs. In the second part, we provide a framework for the process of an impact evaluation. Our proposal is that the ideal way to design an aid project is to (1) include the steps of careful result-based monitoring in the impact evaluation process, and (2) make the impact evaluation a randomized controlled experiment if the circumstances allow for it. Randomization should ensure that the estimated effect is as unbiased as possible and continuous monitoring of results throughout should help us understand where and why any problems arise.

In the third and final part, we provide a practical application of our framework. With information provided by Norwegian Church Aid (NCA), we have been given the opportunity to apply and demonstrate how an impact evaluation could be implemented in practice. NCA is currently working on a program with an overall vision to help smallholder farmers across Sub-Saharan Africa lift themselves out of poverty. This is to be achieved through a concept called “Micro Investing”. As part of their broader program, a project is currently being implemented in Malawi. This specific project will for the remaining part of the thesis be referred to as “Project M”. We illustrate how Project M could have been designed as a social experiment.

The thesis is organized as follows. In Chapter 2 we establish the importance of conducting impact evaluations. In Chapters 3 and 4 we present our framework for the process of impact evaluations: in Chapter 3 we introduce the first steps to result-based monitoring, while we in Chapter 4 move on to the steps that are specific to impact evaluations. We here argue that RCTs produce the most accurate results and explore the benefits and challenges of this method. In Chapter 5 we describe the project in courtesy of NCA. We treat the project as if we were to conduct it as a social experiment, and in Chapter 6 we apply our framework to the project. In Chapter 7 we present some limitations to our work and, finally, in Chapter 8 we give some concluding remarks.

2. The importance of impact evaluations

Historically, the evaluation of social projects, programs, or policies has mainly served the purpose of holding implementers accountable and assure donors that their money has been put to good use. It is common practice when conducting a social project (or program or policy) to perform a *process evaluation*. Process evaluations are the monitoring and evaluation of the implementation of a project. A process evaluation measures how well things are going and provides an early warning if any improvements are needed. Information of this sort can especially be valuable in pilot projects (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, p. 17). This is of course an important part of an evaluation, but it is the bare minimum that should be done in such projects.

In addition to an assessment of implementation, one should also evaluate whether and why a program actually works. Rather than focusing on input versus immediate output, the question in focus should be whether the program induces the *change* we are looking for. For example: rather than asking how much money was spent on a program compared to how many malaria nets were distributed, one should ask how much money was spent compared to how much the program reduced malaria rates. The distribution of nets does not automatically imply reduced rates of infection. This is where *impact evaluations* can provide valuable insight. Impact evaluations can, if conducted correctly, tell us the isolated effect of a particular project and whether the change is induced by the project or if the same changes would have happened regardless of the intervention. The fact that policymakers and stakeholders are requiring more evidence when considering which programs to invest in, is an important trend in moving towards more evidence-based policy-making (OECD, 2012; Stevens, 2011).

It is only through impact evaluations that we can answer questions about cause and effect. Not only do we need to look at whether an outcome is improved after a program is implemented, but we also need some proof that the outcome is improved because of the program and not simply due to some factors that are unaccounted for. That is, we need to distinguish between *correlation* and *causation*. A misunderstanding of these concepts can mislead managers and policymakers in their final decisions. To answer the question about cause and effect, one can introduce elements of an experiment into social programs.

It is common consensus in the pharmaceutical industry that one should conduct controlled experiments before introducing a new drug to the market, but the same standard does not seem to hold in development economics. Funds are allocated without sufficient evidence on what works and what does not. For example, recent studies of the impact of microcredit on poverty alleviation have found that the benefits are modest compared to previous claims (Banerjee, Karlan, & Zinman, 2015). Not only are we at the risk of wasting public resources, but also of harming those in need of help. For example, a 1994 public health campaign in Bangladesh prompted the population to switch water sources because of the discovery of arsenic in groundwater. A recent analysis of the initiative found evidence that unintended consequences of the campaign lead to considerably higher rates of child mortality as well as increased adult mortality (Field, Glennerster, & Hussam, 2011).

To avoid such costly mistakes and improve resource allocations, well designed and well implemented evaluations should be an integrated part of program design from the beginning. It is as important to have research transparency in development economics as it is for medical trials, and this requires reporting more information about the final sample size, manipulations, and data, to improve the quality and credibility of research (Camerer et al., 2016; E. Miguel et al., 2014).

3. Evaluating Social Programs

The implementation of any social project can be described as a process, and in this chapter, we will describe the initial steps of that process. In Section 3.1 we discuss one important thing that needs to be established before even deciding to conduct an impact evaluation: does the problem at hand really need assessment? In Section 3.2 we introduce the Theory of Change model as a means to map the causal chain of a project - from intervention to final outcomes. In Section 3.3 we clarify what to keep in mind when formulating an evaluation question. Finally, in Section 3.4 we define performance indicators and explain how they should be selected to most accurately measure the results of a project.

3.1 Is there a problem in need of assessment?

The first step in the process of designing a social program is to find a problem in need of a solution. This might sound painfully obvious, but it is nonetheless an essential part of the process. Is the problem you are looking to address really a big issue? Once the problem is defined, the next step is to locate the source of this problem. Once the source of the problem is identified, we can begin to look for a solution. In this part of the process, one should look at past and existing proposed solutions to the same problem and examine why they are failing or falling short. The initial assessment should be a systematic approach to identify the nature and scope of a specific social problem, define the target population to be served, and determine the means needed to address the problem (J-PAL, 2017).

3.2 Creating a Theory of Change

Once an overall problem has been identified and a proposed solution is beginning to take form, we need to make sure that any suggested interventions are actually solutions to the problem. In other words, we need to identify the causal path from intervention to outcome. The Theory of Change (ToC) methodology is a tool used to address a problem which is sometimes described as "the missing middle" between a certain change initiative and a desired outcome (Center for Theory of Change, 2019). Concepts such as charity and philanthropy are not new, but for social programs to induce change, it is important to understand how that change is to be achieved. This is what a ToC does; a ToC describes the causal logic from the implementation of a particular project all the way to a desired change.

A ToC is useful for all types of social projects. Such a theory makes program goals explicit and points out the information needed to assure proper program implementation. Hence, a ToC reveals the data one has to collect in the different steps of a program implementation. ToCs also help uncover the indicators required to measure outcomes and make it easier to specify and review an evaluation question. When constructing a ToC, one should start by identifying the anticipated long-term outcomes and then work backwards until arriving at the specific proposed solution, identifying all the conditions that must obtain for the change to take place.

When creating a ToC, the theory can be depicted in a number of different ways – from relatively simple pathway illustrations to detailed and comprehensive models. Approaches to prepare a ToC include logic models, logical frameworks (logframes), and results chains. A basic results chain should map out inputs, activities, outputs, outcomes, and final outcomes (Gertler et al., 2016, p. 34). Figure 1 is a depiction of the different elements of a results chain that illustrates how implementation (inputs, activities, and outputs) leads to results (outcomes, impacts, and final outcomes).

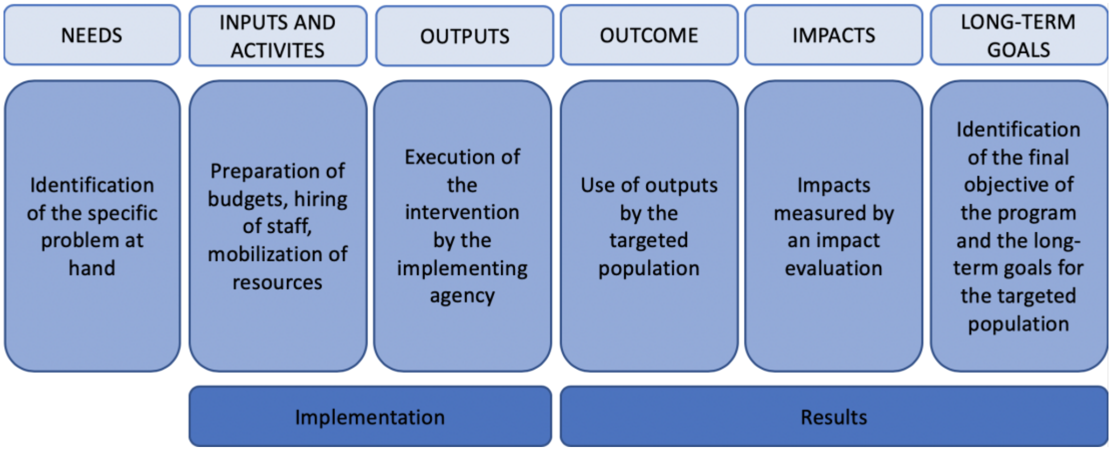
NEEDS
INPUTS AND ACTIVITES
OUTPUTS
OUTCOME
IMPACTS
LONG-TERM GOALS
Identification of the specific problem at hand
Preparation of budgets, hiring of staff, mobilization of resources
Execution of the intervention by the implementing agency
Use of outputs by the targeted population
Impacts measured by an impact evaluation
Identification of the final objective of the program and the long-term goals for the targeted population
Implementation
Results

Figure 1: A simple example of a logical framework. The figure is made with inspiration from J-PAL (2017, p. 5) and Gertler et al. (2016, p. 35).

3.3 The evaluation question

At the heart of any professional evaluation is a well formulated evaluation question. As mentioned in the preceding section, a project's ToC should be used as guidance when specifying such a question. The evaluation question must be tailored to address the problem we are looking to tackle. In the case of an impact evaluation, the evaluation question needs to be formulated as a testable hypothesis (Gertler et al., 2016, p. 36). The purpose of an impact evaluation is to generate credible evidence to prove or dismiss this hypothesis. The basic evaluation question for an impact evaluation is always: What is the impact of the program on the outcome of interest?

The hypothesis following the evaluation question should be clear, testable, and quantifiable.

3.4 Performance indicators and expected effect sizes

After establishing the problem in need of a solution, creating a ToC and formulating an evaluation question, program managers need to specify a set of *performance indicators* to assess results and measure success. These performance indicators will let main stakeholders in the evaluation team know whether implementation of the program was carried out as planned, and whether desired outcomes were achieved.

Outcome indicators are not outcomes per se, but quantitative or qualitative variables that *"allow the verification of changes in the development intervention or show results relative to what was planned"* (OECD DAC, 2002, p. 29). Together, these variables should be a simple and reliable tool to monitor program implementation, evaluate results, and measure achievement. To ensure that the outcome measures are good indicators of program performance, stakeholders from both the research team and the policy team should be included in the process of selecting these performance indicators (Gertler et al., 2016, p. 41).

Ideally, outcome indicators should be affected solely by the intervention. They should be attributable to the project and targeted to the objective population. Outcome indicators should be as clear, direct, and unambiguous as possible. Quantitative indicators should be presented in terms of a specific number or percentage (Zall Kusek & Rist, 2004, p. 69). Qualitative indicators should be applied with caution as they measure perception of progress rather than actual progress and may therefore easily be biased. It might be easiest to define progress in quantitative terms, however, the progress that really matters (e.g., whether people are living better lives), might well be a matter that can best be investigated qualitatively. If needed, indicators can be added or dropped later in the process, but the decision to change indicators should be carefully considered.

Expected effect sizes

Once the performance indicators have been decided upon, targets, or minimum *expected effect sizes*, need to be established. Expected effect sizes are the anticipated values of the outcome indicators (Gertler et al., 2016, p. 41). In other words, they represent the changes expected to occur as a result of a program, such as the quantitative change in test scores as a result of smaller class size or the qualitative reported change in life quality as a result of access to credit. These targets form the basis for the technical elements of an evaluation, including deciding on the required sample size and conducting power calculations (see Subsection 4.3.4). The expected effect sizes should include specific and realistic time frames for achievement. They should be set for the intermediate term, as opposed to the long term, so that they can be compared to the results of the impact evaluation.

To establish a reference for the expected effect sizes, it is crucial to collect *baseline data* on the outcome indicators. Baseline data is the measurement of initial conditions and is used to compare the results of a program to the starting point. As will be discussed further in Section 4.5, the need for baseline data is, in theory, eliminated when a randomized controlled experiment is correctly conducted. Comparison of endline data to baseline data then shifts to comparison of endline data from a treatment group to endline data from a control group. Nonetheless, baseline data is undismissable for the *process evaluation* part of an impact evaluation. Moreover, social experiments are always complicated, and a lot can interfere with the execution of a social experiment, even when well-designed. Collecting baseline data is useful for verifying that the assignment has not been accidentally skewed in the randomization process (see Subsection 4.2.1), and can help shed light on where in the process something went wrong if results are not in alignment with predetermined goals.

It is important that there is clarity on the source of data and how data will be collected for each performance indicator. Program designers need to ask themselves questions such as: “Will the data be obtained from a survey, a review, or perhaps from existing administrative data?", "At what point in time and with what frequency will data be collected?", "Who will be responsible for the collection of data?", and "What will be the economic cost of collecting and analyzing the data?" .

Indicator development is a critical aspect of moving towards more result-based monitoring of social programs. It helps recognize success, answer the questions of cause and effect, and forms the basis for all subsequent data collection, analysis, and reporting (Zall Kusek & Rist, 2004, pp. 65-66). In addition, it can assist managers with budgeting, resource allocation, and staffing (Morra Imas & Rist, 2009). Developing the right indicators is both a time-consuming and resource demanding process, but in evaluating social programs it is a fundamental step.

Up until this point, we have described the process which is always relevant when planning to implement a social program. The next step in the process is to decide upon whether an impact evaluation should be conducted. There are different ways to measure impact, and often we see that evaluations are lacking hard evidence on the cause of observed changes. In Chapter 4 we will go deeper into the material on the subject of impact evaluations, discussing the different impact evaluation methods and some methodological challenges.

4. Impact Evaluations

The main objective of an impact evaluation is to determine the *ceteris paribus-*, or "other things equal" impact of a program, preferably by comparing a group that is affected by the intervention (a treatment group) to a group from the same population that is not affected by the program (a control group) (Chabrier, Hall, & Ben, 2017). In this chapter we will dive into the different aspects of impact evaluations and the methods for creating treatment and control groups.

In Section 4.1 we explore the concept of the counterfactual situation and how to measure something unobservable. In Section 4.2 we briefly describe the most common experimental and non-experimental methods of creating treatment and control groups. In Section 4.3 we give a more detailed account of the experimental design of randomized controlled trials. Section 4.4 presents some of the methodological challenges we are faced with when conducting social experiments. In Section 4.5 we explain the process of collecting data for social programs. In Section 4.6 we give a short note on the interpretation and presentation of the results of an impact evaluation and comment on the generalizability of a social experiment. Finally, in Section 4.7 we explain how these results can be used for a comparative cost-effectiveness analysis between different projects.

4.1 The counterfactual

The essence of an impact evaluation is to create a substitute for the counterfactual, that is, we need to simulate what would have happened in the absence of an intervention (White & Raitzer, 2017, p. 32). To evaluate the exact effect of a project, one would ideally like to observe at time t^2 both the results of individual A having received the intervention at t^1, as well as the results of individual A having not received the intervention at t^1. Otherwise, one could not be sure that any differences in outcome could be attributed to the intervention and not to other, unrelated factors, or so called *confounding variables*. The problem is that both administering and not administering the intervention to the same individual at t^1 obviously cannot be done. If we do administer the intervention at t^1 the results of not administering the treatment will be a counterfactual. Without this access to this counterfactual, we cannot know the actual contribution of a project and are thereby at risk of drawing uninformed and inaccurate conclusions. For example, it might be tempting to announce that a certain program is successful if there is a positive increase in an outcome after it has been launched. However, without an impact evaluation we cannot say whether this increase can be attributed to the program or is simply due to a positive underlying trend. The same logic applies if the outcome is declining – maybe the counterfactual situation would have been as bad or even worse.

Since the counterfactual cannot be observed, we have to mimic or simulate it artificially. Conventionally, project performance has been evaluated on a set of criteria for deciding whether the program was successfully carried out or not (see Section 3.4 about performance indicators). This approach is called a *before-and-after* analysis. In this method, relevant outcomes of a sample (of people) is compared to the same outcomes of the same sample before the intervention was implemented (Gertler et al., 2016, p. 54). Program implementers might not explicitly state that this “before scenario” represents the counterfactual, but in effect that is what they are stating when claiming that a program had a certain impact. Doing so, might however, lead to invalid conclusions given the plausibility of other factors affecting the outcomes simultaneously. Hence, this approach would only tell us what happened, not why it happened.

A second traditional approach is called *with-and-without* comparisons between enrolled and non-enrolled. The idea is to compare the individuals in a group choosing to receive some form of treatment to the individuals in the same group that choose not to receive the treatment (Gertler et al., 2016, p. 48). This event would only work in studies where there are no systematic differences between the ones that chose treatment and the ones that did not. Not differing systematically means that the “control group” for instance is not much richer or poorer than the “treatment group”. However, there is a great chance that there are some underlying factors affecting the enrollment in itself. If that is the case, a with-and-without comparison will not give a valid estimate of the impact.

Luckily there is one method for constructing the counterfactual that is more likely to create a valid estimate of the true impact of a project. That is by intentionally creating a group that compares to the treated group. Such groups go by the names control, comparison, or placebo group. What we want to know is whether a given program induces a difference in outcomes between the treatment and non-treatment groups. In the following section we will describe different methods for creating valid treatment and control groups.

4.2 Impact Evaluation methods

To define a treatment and a control group, we can conduct evaluations that are either experimental, or non-experimental. In experimental evaluations, design is decided beforehand to create an experimental situation. When the design is experimental, it is called a *randomized controlled trial.* As will be discussed in the forthcoming subsection, RCTs are considered the “gold standard” of impact evaluations and we will therefore explain such experimental evaluations thoroughly in Section 4.3.

If it for some reason is not sensible or possible to conduct an RCT, non-experimental designs can give similarly credible measurements of impacts. Non-experimental evaluations can be conducted when a natural condition, geographic locations, or a government policy in effect separates the same population into control and treatment groups. Such situations are referred to as *natural experiments* and evaluations are then designed after the intervention has taken place. Natural experiments are addressed through either *quasi-experimental designs* or *regression-based approaches*. Quasi-experimental designs include difference-in-differences, propensity score matching, and regression discontinuity design. The most common regression-based approach is the instrumental variables method. The idea with non-experimental designs is that they create two groups that are “as good as randomly assigned”. However, these methods should be applied with caution as they require making various assumptions to get the *ceteris paribus* effect.

4.2.1 Randomized controlled trials

A randomized controlled trial (also known as a randomized evaluation, field experiment, social experiment, or experimental design) is an experimental type of impact evaluations which involves randomization of the allocation of units to treatment and control. The aim is to measure the effect of a project by comparing outcomes in a group where the project is implemented with outcomes in another group from the same population not affected by the project. Since participants are assigned at random, they have the same chance of receiving treatment and, hence, we can have confidence in that the difference in outcome can be explained by the program introduced. Because they consistently produce the most accurate results, RCTs are often considered the gold standard of impact evaluations (J-PAL, 2017, p. 9). We will circle back to RCTs in the subsequent section, but first, we will give a brief description of some alternative, non-experimental methods of impact evaluations. For a more thorough explanation of these methods, see, for example, Angrist and Pischke (2008).

4.2.2 Difference-in-Differences

Sometimes treatment cannot be randomized because the intervention has already taken place, or it might be unethical to withhold treatment from only a certain group of participants. When treatment is not randomized, the *difference-in-differences* approach can be used to control for the possibility of an underlying trend affecting outcomes. The key assumption for difference-in-differences evaluations is thus an assumption of common trends between treated and non-treated (Angrist & Pischke, 2008, p. 171).

The *difference-in-differences* approach combines a before-and-after analysis with the control group approach described above, as a treatment group is compared with a control group both before and after the program is implemented (Gertler et al., 2016, p. 130). This method is commonly used when looking at something that changes at a specific point in time (e.g., a new policy is introduced). Longitudinal data is used, meaning that the chosen unit – say an individual – is observed over time. The simplest form of this method is when there is one group that is affected by the program and another is not, and outcomes of the groups are compared pre-treatment and post-treatment. If the treatment is random, we do not need a difference-in-differences to get unbiased estimates of the effects because one can then simply look at the differences between the treatment and control groups. Having said that, even in those cases it can be valuable to use difference-in-differences to improve the precision of the estimates.

4.2.3 Matching

Matching is a method in which one uses large data sets and statistical techniques to construct a control group with very similar covariate values as the treated group (for instance the same gender, roughly the same income, education etc.) (Gertler et al., 2016, p. 143). For every treated unit, matching attempts to find a non-treated unit that has as similar observable characteristics as possible. The comparison of all these "matches" makes up a list of treatment differences that give us the average treatment effect.

There are different ways of finding people with similar covariates, but the most commonly used method is *propensity score matching* (Rosenbaum & Rubin, 1983). The first step in propensity score matching is to estimate the propensity score, that is, the likelihood or chance that an individual gets allocated to the treatment group. Thereafter, the propensity scores are used to match individuals who had similar scores to get a more convincing control group. Then, the next step is to evaluate the quality of the matching and, assuming that the match has a balance of covariates, the last step is to evaluate the intervention or policy.

4.2.4 Regression Discontinuity Design

A third quasi-experimental approach is *regression discontinuity design.* This method can be used when there is a "threshold" or cut-off point (e.g., the poverty line or a certain test score requirement). When there is a precise enough threshold where people above the threshold are treated and people below are not treated, we can sometimes assume that there are no systematic differences between the people who are just above and just below the threshold. This allows us to use the people just below the threshold as a control group against the treatment group composed of the people just above the threshold. Of course, this requires that there are no systematic differences between those just above and those just below the cut (Gertler et al., 2016, p. 113).

4.2.5 Instrumental variables

The instrumental variables method can be used to produce valid estimates from partial or incomplete random assignment, whether naturally occurring or generated by researchers (Angrist & Pischke, 2014, p. 98).

Instrumental variables methods deal with the problem of endogeneity. Endogeneity arises when one of the independent variables in our model is correlated with the unknown error term (Wooldridge, 2002). When receiving the treatment is correlated with the error term, i.e. unknown factors, it becomes hard to say whether the observed effect is a result of the treatment itself or a result of the unknown factors correlating with receiving the treatment. Sometimes, however, there is a third variable that we know affects who receives the treatment, and we know does not correlate with the unknown factors. In such cases, it is possible to estimate the effect of the treatment by measuring the effect of this variable instead.

Instrumental variables are also relevant in experimental designs, because not everyone offered to take part in a treatment group will choose to participate, and we cannot guarantee that everyone in the control group will not be affected by the treatment. In these cases we can use the randomized assignment to treatment as an instrumental variable (White & Raitzer, 2017, p. 86).

4.3 Experimental studies minimize bias

The key to an accurate impact evaluation is to construct a treatment and a control group with no systematic differences. It was the statistician R.A Fisher that found a way around this problem by introducing random assignment, which eliminates systematic differences between the two groups and thereby solves the problem of selection bias (see Subsection 4.4.1) (Gerber & Green, 2012, p. 6). In a well-executed RCT, the groups should have no systematic differences regarding both observed (e.g., test scores) and unobserved characteristics (e.g., motivation). When there are no systematic differences between groups, because all the differences are due to chance, we can use statistical methods to analyze the likelihood that any difference in outcomes between the treatment and control groups is due to the treatment and not to chance. Hence, when conducting an RCT, we eliminate the need to identify any confounding variables. When the two groups are equal in every way possible, confounding variables which we are not able to envision are controlled for automatically (Kendall, 2003).

As RCTs produce the most valid representations of the counterfactual, they will produce the most accurate results and are therefore considered a benchmark by which results from other evaluation methods should be judged (Angrist & Pischke, 2014).

4.3.1 Whether and when to conduct an impact evaluation

Even though experimental designs produce the most accurate results of an impact evaluation, it cannot be argued that an RCT, or even an impact evaluation at all, should be conducted for all social programs. Impact evaluations are only valuable if the evidence generated will be used in one way or another. Evidence from impact evaluations can back up the decision to continue, upscale, or replicate a project. It can also help managers understand how programs can be adjusted to become more effective. Impact evaluations should be prioritized where there are gaps in the existing body of evidence (White & Raitzer, 2017).

Two common concerns about randomized evaluations are that they are considered unethical and costly. It is true that impact evaluations in general, and perhaps RCTs in particular, are costly to conduct. However, compared to other evaluations, randomized evaluations are not necessarily more expensive because the cost varies with the type of data required. Evaluations that are using data that already exists publicly are less costly than collecting new data that is not already in place (Chabrier et al., 2017). The two main ethical issues concerning impact evaluations relate to the concern regarding research on human subjects, and to the fact that the control group does not receive the intervention. The human subject issue is discussed in Subsection 4.5.2.

In most cases, there is an untreated population or group when a social program is implemented anyway. Impact evaluations and RCTs does not create the untreated population, they simply identify it. It can be argued that random assignment of treatment is actually the fairest way to decide who gets to participate in the program, as treatment is granted by chance. That said, individuals of the control groups may still feel unfairly treated, especially as they are subject to data collection for an intervention without getting any direct benefits (White & Raitzer, 2017, pp. 136,137).

It might be tempting for program managers to “wait and see” whether it would be beneficial to evaluate the impact of a program, as resources are often limited and impacts are generally measured years after an intervention is first implemented. However, evidence from impact evaluations designed ex-ante, is almost always more rigorous than from ex-post designs. This follows mainly from the opportunity for random assignment and collection of baseline data.

As mentioned throughout this paper, program managers should plan for evaluations already in the design-phase of a program. However, when to conduct an evaluation is not uncomplicated because the evaluators needs to balance the timing of the evaluation (White & Raitzer, 2017, p. 41). An ideal time would be in the pilot phase of a project or before it is scaled up (J-PAL, 2017, p. 12). A lot can be learned from performing pilot projects and this information may be valuable in improving the program.

Even though RCTs are ideally designed ex-ante, it may be possible to use elements of random assignment when an intervention is ongoing. According to White and Raitzer (2017, p. 56), one can use random assignment in the rollout of the program, introduce variations into program implementation for adaptive learning purposes, or use an encouragement design. Thereafter, when the pilot phase is over, it is time to consider the effectiveness and whether it should include the remaining part of the population.

4.3.2 Population of interest and unit of randomization

When conducting a social experiment, the aim is to find results that are valid for a larger group of people. However, as already established, the treatment group needs to be statistically identical to the control group for the experiment to produce valid results. Randomization makes this possible, but the randomization has to be limited to a specific population of interest. Project managers first need to establish who is eligible for the treatment in question.

After identifying the eligible population, managers need to decide on the *level* of assignment. It is possible to randomize on the individual level or on higher levels such as whole groups or clusters (Bloom, Bos, & Lee, 1999). Clusters are usually pre-existing groups such as hospitals, schools, clinics, or geographic areas. Cluster randomized trials are often conducted when the intervention is aimed at the whole group. Another reason to randomize on the cluster level is to control for contamination, that is, when individuals or groups are benefitting from a project that they are not supposed to take part in (see Subsection 4.4.6) (Leeuw & Vaessen, 2009). We want to separate those who get the intervention from those who do not, and the easiest way of doing that is by cluster randomizing. A third reason for using the cluster randomization approach is when it is ethically difficult to randomize individuals. However, cluster randomizing results in larger trials and requires a larger sample size, hence, it adds on to the complexity of the design (Puffer, Torgerson, & Watson, 2005). Considering the risk of contamination, impact evaluations in social projects are often designed as cluster randomized trials (White, 2013).

4.3.3 How to approach random assignment

RCT designs not only differ with respect to the level of assignment, but also regarding the approach to random assignment. A common concern about randomized evaluations is that it is unfair to hold certain individuals outside a program that might be lifesaving. It would of course be unethical to exclude or deny people from entering a program in a case where we had massive evidence of its efficacy and enough resources for everyone qualified to participate. Unfortunately, this is often not the case. However, there are variations on randomized assignment and can be designed to address ethical issues simultaneously. There is also an option to include elements of randomization into programs that already exists and examples of other designs than the one conventionally used – by randomly assigning people into treatment and control groups – is the *lottery design, phase-in design, rotation design, encouragement design, different treatment design,* and *two-stage randomization* (J-PAL, 2017, p. 17).

Amongst the different designs to choose between when randomizing, four designs will be explained further: simple randomization, pipeline randomization, raised threshold randomization, and encouragement designs (White, 2013).

Simple randomization

Simple randomization is the assigning process we have referred to so far, where the unit of randomization is drawn at random from the predefined population and assigned to either treatment or control until the predetermined sample size is reached. This method is the easiest, and most basic way of assigning subjects. This approach to random assignment might be subject to the ethical dilemma of unfair treatment. But with limited resources and as long as assignment really is random, this might actually be the fairest possible approach. Still, issues can arise when collecting data from the control group. Data collection is not only costly for the project implementers, first hand data is costly for the people providing the data as well. Why should they take the time to answer a survey if they do not gain anything from it? Furthermore, if the control group is aware that another group of people is getting some kind of beneficial treatment they may be provoked, intentionally or subconsciously, into giving incorrect or imprecise answers.

Pipeline randomization

An alternative approach which might help with the possible issue of data collection from the control group, is pipeline- or phase-in design. In a pipeline randomization, all units of assignment will eventually receive treatment. Compared to simple randomization where assignment to treatment is random, it is the time of entry to the program that is random. This approach is mostly used for cluster randomized controlled trials. To begin with there will be several units functioning as control groups and only a few treatment groups. Then, as treatment is introduced to more units, they make a permanent switch from control to treatment group. This approach might be useful for example when budgetary or logistical constraints prevent the program from reaching the entire population at once. Using pipeline randomization makes sense for a lot of social programs, which are often rolled out in stages targeting one school, village, etc. at the time. For a practical application of the pipeline design see, for example, Attanasio, Meghir, and Santiago (2011).

Raised threshold randomization

Raised threshold randomization is not as widely used as the two approaches described above, but it expands those qualified to participate and randomize within the group. There is often a threshold that decides who is qualified to receive the program, say the poverty line or entry grades.

Encouragement designs

Lastly, we have encouragement designs which randomly assigns participants to the offer of receiving something that makes them more likely to take part in the program, but they choose for themselves whether to receive the treatment (West et al., 2008). In such a case, those encouraged to receive the treatment are compared with those who are not encouraged to receive the treatment.

4.3.4 Sample size and power calculations

Once the eligible population is identified and the unit of randomization is decided upon, the next step in an impact evaluation would be to determine the sample size required to accurately estimate differences in outcomes between treatment and control groups. Random sampling is the process of drawing units from the population of interest. To ensure that findings from the sample analysis is generalizable to the entire eligible population (see Section 4.6), a sampling frame should ideally coincide exactly with the population of interest. A sampling frame should list all units of the population of interest. Note that, as explained above, units could be clusters, hence the list does not necessarily need to contain information about specific individuals (Gertler et al., 2016, p. 263). After a sample has been drawn from the sample frame, it is from this sample that units will be randomly assigned to treatment and control. The distinction between random sampling and random assignment is important to keep in mind.

How many units to draw from the sampling frame for the sample to provide precise estimates of program impact is decided by *power calculations*. Power calculations indicates the smallest sample size required to go through with an impact evaluation. Calculating power is a technical procedure which needs to be done properly and the evaluation team should therefore include a statistical expert from an early stage.

Power, in this context, is the probability of finding a difference between the control and treatment group given that one truly exists. It is the statistical power of an experiment that determines the probability for results to be *statistically significant*. The simplest way of testing statistical significance is through a t-test. A t-test is a statistical hypothesis test where the null hypothesis is that the intervention has no impact. Statistical significance is tested using a pre-specified significance level. The significance level is the probability of rejecting the null hypothesis when it is in fact true. Statisticians often use a 5% or 10% significance level. For instance, a 5%

significance level indicates that you would incorrectly reject the null hypothesis that the intervention has no effect on average 5% of the times (Stock & Watson, 2014).

Larger sample sizes produce more accurate estimates of differences between treatment and control groups (Gertler et al., 2016, p. 267). However, when determining the size of the sample there might be limitations such as time, logistics, and money.

4.4 Methodological challenges

Because of their complexity, RCTs are often not implemented exactly as planned. When conducting social experiments (either through experimental or non-experimental designs) there are some methodological challenges that one should always be aware of. In this section we will touch upon a few of the challenges that one should be aware of before implementing an RCT.

4.4.1 Selection bias

If individuals have the opportunity to choose whether they participate in a program or not, it is said that they self-select into treatment. With self-selection, there is most likely underlying factors that affect whether individuals choose to participate in a program. This is a source for a problem commonly referred to as "selection bias". *"Selection bias will occur when the reasons for which an individual participates in a program are correlated with outcomes, even in absence of the program"* (Gertler et al., 2016, p. 59). Selection bias can also arise because of program placement. If a program is targeted at the poor, outcomes should not be compared to the non-poor, but a set of others with similar income and characteristics (White & Raitzer, 2017, p. 36). The problem of selection bias arises when participants in the treatment and control groups are chosen in a non-random way so that they differ from each other in some characteristics that will affect the outcome. Hence, experimental and non-experimental designs should solve the problem of selection bias as these methods should ensure that assignment is random or as good as random.

4.4.2 Heterogeneous treatment effects

So far, we have (implicitly) assumed that if treatment is given, it will affect all units of the population in the same way. However, responses to treatment might differ systematically across different groups of recipients. If they do, we have what is commonly referred to as *heterogeneous treatment effects* (Gertler et al., 2016, p. 159). If impact evaluations are to capture these heterogeneous effects, they have to contain subgroup analyses. To be able to conduct a subgroup analysis it is essential to have enough data on the subgroups of interest. Data should be collected through stratified sampling (see Subsection 4.5.3) to make sure that the sample consists of a sufficient number of representatives from each subgroup. Heterogeneous treatment effects are therefore important to consider at a relatively early stage of the evaluation process and should be implemented in the ToC (see Section 3.2).

4.4.3 Attrition bias

Attrition bias can arise if parts of the sample for some reason disappear over time so that the researchers are not able to collect endline data on the whole sample (Gertler et al., 2016, p. 169). Attrition can be a problem because the data collection sample might not accurately represent the population of interest. We might be left with only a subgroup of the original sample and, as a consequence, we might no longer have balanced treatment and control groups. If the groups are not balanced, we can no longer find the “other things equal” or *ceteris paribus* effect of the program. Imagine, for example, an intervention taking place in some poor village. Imagine further that there are heterogeneous treatment effects and that the intervention significantly increases the average income of only part of the villagers. The increased income might cause villagers to move to more urban areas. Data collectors are then left to collect data only on the individuals who remain in the village. In other words, they are left to collect data on the villagers who did not react as strongly to the intervention. An evaluation of the results will then underestimate the effect of the intervention.

4.4.4 The Hawthorne effect

Unintended behavioral responses might occur if you know that you are part of an experiment. This effect is known as the *Hawthorne effect* (Roethlisberger & Dickson, 1939). The Hawthorne effect arises when individuals behave differently simply because of the fact that they are being observed.

4.4.5 Imperfect compliance

In any social experiment, there may be a discrepancy between assigned treatment status and actual treatment status. This is commonly referred to as *imperfect compliance.* Imperfect compliance happens when some subjects assigned to the treatment do *not* receive treatment (non-compliance), and/or when some subjects assigned to the control groups somehow *do* receive treatment. Individuals of the latter case are referred to as *crossovers*.

The evaluation team has to make a choice when analyzing the data in a sample with non-complying participants. One option is to do it "per protocol" where only data from participants who were compliant with the treatment is analyzed. A second option is to analyze the sample "as treated". Data from participants is then examined for the group that they conform to regardless of which group they were randomized to (Sainani, 2010, p. 212). A final option is called "intention-to-treat" were all of the participants data is kept in the group they were originally assigned to by the randomization process. When data is analyzed using the intention-to-treat option, the results will expose the *average treatment effect* but not the *average treatment effect on the treated* sometimes referred to as the *treatment-on-the-treated* effect (Gertler et al., 2016, p. 91; White & Raitzer, 2017, p. 156). The best way to manage the potential bias that occurs from non-compliance is to use intention-to-treat analysis (Sainani, 2010, p. 212). Any impact found analyzing the intention-to-treat will be smaller than the treatment-on-the-treated effect. To find the treatment-on-the-treated effect, one can use an instrumental variable method (Gertler et al., 2016, p. 91).

4.4.6 Spillover effects

Somewhat surprisingly, little light has been shed on the unintended side-effects of aid. A report commissioned by Norad's Evaluation Department concludes that one of three Norad evaluations did not mention unintentional effects, even when specified in the terms of reference (Wiig & Holm-Hansen, 2014). Unintended effects create a problem for analyzing experiments because it complicates the statistical analysis.

If part of the control group is affected by the intervention in one way or another, they have been exposed to *contamination* or *contagion*. In addition to the issue of crossovers mentioned above, contamination commonly happens through *spillover effects*. There are three common types of spillover effects: externalities, social interactions, and general equilibrium effects (Gertler et al., 2016, p. 163).

Externalities

Externalities are effects that go from treated subjects to untreated subjects (Gertler et al., 2016, p. 163). An example of a positive externality is the Kenyan deworming program analyzed by Miguel and Kremer (2004). In a school health project, deworming drugs and health education messages about avoiding worm infections were given to randomly assigned primary schools. The deworming drug interfered with disease transmission, which benefited the children in nearby preschools since they were less likely to have worms although their schools had not been given treatment through the deworming drug.

Social interactions

Spillovers might result from social or economic interactions between the treated and non-treated (Gertler et al., 2016, p. 163). An example of spillovers that happened through social interaction is the famous Perry Preschool Project. Imagine that two next-door neighbors each have one kid of the same age and one of the kids is randomly assigned to attend preschool while the other kid is randomly assigned to not attend preschool. Suppose that the kid attending preschool starts sharing newfound knowledge with the other kid. This child is then affected by the treatment indirectly though his neighbor, despite the fact that he did not get the treatment of attending preschool (Neidell & Waldfogel, 2010).

General equilibrium effects

Evaluations of social programs tend to look only at partial equilibrium effects, if even that, and evaluations of the macroeconomic consequences of the program are often missing (Duflo, 2004). A *partial equilibrium analysis* is based on the analysis of a particular sector, say prices in the rice market in isolation. This approach deals with each market independently without considering the effects of changes in one market on other markets. A *general equilibrium analysis*, however, recognizes interdependencies among different economic units. Interdependence in the economy makes partial equilibrium analysis overly simple because supply and demand in one market depend on prices determined in other markets. A general equilibrium analysis broadens the perspective as it is taking into account the interactions and interdependencies within various parts of the economy.

One way to reduce the chance of spillover effects is by changing the units of randomization to a higher level, by randomizing at the level of for example schools or villages rather than randomizing individuals. If there is still a chance for spillovers to affect the evaluation, it might be necessary to collect data on an additional control group. General equilibrium effects are, however, harder to affect or prevent. They should nonetheless be studied and taken into consideration when evaluating results.

4.5 Data collection

4.5.1 When to collect data

After identifying an eligible population, deciding on the level of assignment and how to approach the random assignment, conducting power calculations and choosing the sample size, implementers of an RCT should randomly draw the sample for analysis from the population and assign units to treatment and control. When treatment and control groups are identified, *baseline data* on the performance indicators (see Section 3.4) should be collected (Morra Imas & Rist, 2009, p. 119). The baseline data should then be used to check for balance between treatment and control groups. As already mentioned, if the groups are unbalanced, the impact evaluation will not produce the accurate effect of the intervention and program implementers will have to go back and examine the implementation of the random sampling and assignment.

When treatment and control groups are balanced, everything should essentially be in place to roll out the project. While the project is ongoing, *midline data* should be collected along the way to provide managers with a continuous flow of performance information and feedback. In addition to data on the performance indicators, data should be collected about program activities and outputs, and preferably also outside influences (Gertler et al., 2016, p. 293).

At one point, the evaluation period for the project must end. It is at this predetermined point that *endline data* should be collected. The mean outcomes of the endline data from treatment and control groups are compared, and the impact of the project (so far) is determined. This comparison will give the impact of the program in its simplest form. However, to be able to state that this impact is statistically significant, there is a need to conduct a statistical hypothesis test (J-PAL, 2017, p. 21). Analyzing the results of an impact evaluation will be further touched upon in the forthcoming section.

Follow-up data can be insightful when programs are expected to have long-run effects. The timeline for follow-up data differs from project to project and could vary from one year up to five years after the intervention took place. Data collection is a costly procedure but is arguably the most important part of an evaluation process and resources should be allocated with this in mind.

4.5.2 Types of data and the ethics of data collection

Data can be collected through either *primary* or *secondary* data. Primary data is gathered through for example surveys, interviews, or direct observations by the researchers conducting the evaluation, whereas secondary data is obtained from existing sources such as administrative data, public statistics, or previous studies (J-PAL, 2019a).

In most social programs there is a need for data on specific objectives, and thus primary data often has to be collected. When collecting data on people, researchers need to be considerate and remember that their subjects are humans not objects. It is common practice to follow a set of ethical principles for the protection of human subjects of research. These principles are presented in a document from 1979 titled "The Belmont Report". The three core principles are *respect for persons*, *beneficence*, and *justice* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 2014). In accordance with the first principle, "respect for persons", all research subjects must give informed consent to participation in research. This means that when researchers want to obtain data from individuals, they should provide these individuals with adequate information regarding the study and give them adequate opportunity to consent or decline to participate. It also entails that extra safeguards are in place to make sure vulnerable people like children or homeless people are not tricked or coerced when deciding to participate. The second principle, "beneficence", implies, at a minimum, that the potential benefits of the research – what is being gained by the society – must outweigh the potential harms. The third principle, "justice", implies, among other things, that the participants in the research should, ideally, also be potential beneficiaries of the research.

4.5.3 Sample size for data collection

In most cases it is too time-consuming and resource demanding to collect data on the whole sample. As a solution, one can take a stratified sample of the treatment and control groups and collect data from that subset of units (Gertler et al., 2016, p. 264).

When forming a stratified sample, it is important that individuals are randomly selected again, even though they have already been randomly sampled from the entire population and randomly assigned to treatment and control. To safeguard against selection bias, the selection must also be blind (see Subsection 4.4.1).

As with the sample size for treatment and control groups (see Subsection 4.3.4), power calculations are used to decide upon the size of the sample for data collection. To get precise estimates, both a large enough sample from the population of interest and a large enough subsample for data collection is required. With too few observations, estimates will be inaccurate, and the sample average will not represent the true average of the population.

4.6 The results of an impact evaluation

When endline data on treatment and control groups has been collected, it is finally time for the research team to evaluate results. One of the benefits of an RCT is that impact can be measured without advanced statistical techniques. The simplest method to estimate an impact is to compare average outcomes of the treatment group to average outcomes of the control group (J-PAL, 2017, p. 21). However, in all serious evaluations one must make sure that results are statistically significant. As explained in Subsection 4.3.4, it is the statistical power of an experiment that determines the probability for results to be statistically significant. It should be noted that power calculation is not an easy task, as it requires accurate estimation of standard errors. If the sample size in a trial meets the requirements of a valid power calculation, the analysis is likely to show statistically significant effects given the existence of real effects. As explained in Subsection 4.3.4, the simplest way of testing significance is through a t-test. A t-test is a statistical hypothesis test where the null hypothesis is that the intervention has no impact. If the difference in outcome between treatment and control is not significant, the conclusion of the experiment is that the observed difference in averages may be due to sampling error. The null hypothesis is then accepted (Stock & Watson, 2014).

It is common to perform slightly more complicated analyses. To get a more precise estimate, one can for example run a regression including a dummy variable – a variable that is either 0 or 1 – indicating treatment status. For even more precision, one can include control variables for various characteristics measured at baseline (Duflo, Glennerster, & Kremer, 2007, pp. 31,35).

Internal validity

The result of an honest impact evaluation conducted through a randomized controlled experiment should have *internal validity*. If results are internally valid, it means that one can trust the conclusions drawn about the specific sample. Internal validity implies that causal inference can be made, that is, the difference observed in the outcome variable between treatment and control are solely caused by the program. The condition for internal validity is that the control groups is a valid estimate of the counterfactual (Gertler et al., 2016, p. 73).

External validity

Randomization should not only ensure that the groups assigned to treatment and control are statistically identical to each other, but also to the entire eligible population they were drawn from. Hence, results from the evaluation should be generalizable to the whole population of interest. This is referred to as *external validity* (Gertler et al., 2016, p. 73). However, this validity rests on the two assumptions that the estimator of the causal effect is unbiased and consistent, and that the standard errors used for power calculations and significance tests are appropriately computed. For various reasons these assumptions might not be met and, if so, validity is threatened (Stock & Watson, 2014, pp. 362,363).

External validity is a term that can also be used in a broader sense. External validity for the population of interest does not entail generalizability to other populations or other settings. External validity does, in contrast to internal validity, not deal with the quality of causal claims. Instead, it looks at whether the findings in our research are uniquely applicable to those participants looked at or if they also concern other groups. Hence, a study that has a high degree of external validity means that the findings can be applied to almost anyone, meaning that the findings are generalizable (Bracht & Glass, 1968). Even if a social experiment is perfectly designed and executed, the findings may not be generalizable because the results lack external validity.

4.7 Cost-effectiveness analysis

When results have been analyzed and impacts determined, it can be beneficial to perform a cost-effectiveness analysis of the project. Cost-effectiveness analyses measure the ratio of the costs of a program to the effects it has on an outcome (Yalouris, 2014). Cost-effectiveness analyses can especially be useful if the project is a pilot with plans for later implementation at scale.

Cost-effectiveness analyses can also help managers and policymakers make systematic choices between different programs. Such comparative cost-effectiveness analyses compare the cost-effectiveness ratio from one program to a similar ratio for several other programs. Cost-effectiveness analysis is a transparent way to synthesize information across different evaluations. Such an analysis can be useful when multiple program options have been rigorously evaluated with positive impacts, but there is uncertainty about which program has the highest impact at the lowest cost. Also, it can be useful in making the case for a non-obvious program, such as providing deworming drugs to improve student attendance, which has turned out to be extremely cost-effective (Edward Miguel & Kremer, 2004).

There are however, several challenges in doing cost-effectiveness analyses. A program might prove to have a significant impact but not be cost-effective. Costs are also very hard to gather from the implementing organizations and the process is not easy because there is no standard procedure for documenting costs. Adding to the complexity is the many assumptions required to complete the analysis (e.g., multiple outcomes, spillovers, inflation rates).

5. Project M

In Chapter 3 we explored the initial steps of a result-based evaluation of any social program, project, or policy. In Chapter 4, we delved into the specifics of impact evaluations. In this chapter, we will introduce the ongoing program by Norwegian Church Aid. In section 5.1 we explain the process of our collaboration with NCA. In Section 5.2, we describe a pilot project of the Micro Investing program launched in Tanzania. In Section 5.3, we briefly describe the economy in Malawi and take a dive into the agricultural sector to get an idea of their current situation. In the last two Sections, 5.4 and 5.5, we look at the specifications of the project. This includes the distribution of kits, who are being included to receive the program, and the phase-out of Project M.

5.1 Our collaboration with Norwegian Church Aid

The program at hand was developed by independent consultant, Jakob Fagerland, in collaboration with Norwegian Church Aid. We first heard about the program through solutions architect, Johannes Ensby, whom we reached out to after a tip from a guest lecturer at BI Norwegian Business School. Mr. Ensby inspired us to write our thesis about development projects, and has in effect been a mentor for our work regarding Project M.

Our first correspondence with NCA was through Mr. Ensby via email. After approximately two months of email correspondence, we set up a meeting with both Mr. Ensby and Mr. Fagerland at the NCA main office in Oslo. They briefed us about the program and discussed why they believed it would work. We then pitched our preliminary thesis report and explained our motivation for writing a thesis about impact evaluations. Agreeing that a collaboration could be mutually beneficial, we decided to keep in touch throughout the process of writing our thesis. Mr. Ensby has provided us with what information he could along the way. Collaborating with NCA has been a unique opportunity to get insight into the process of designing a social project. Insights from this collaboration has been as close to firsthand knowledge about a specific social program as we could have hoped to get.

The description of Project M in this chapter is mainly a recollection of unpublished concept descriptions of the project, henceforth referred to as “Concept Description, 2018” and “Concept Description, 2019”. In Chapter 6 we present some numbers and findings from an unpublished baseline study, henceforth referred to as “Baseline Report, 2019”. The Concept Description and Baseline Report are available upon request.

5.2 The pilot project

The program has been developed through trial and error in a pilot project which was launched in Tanzania in 2015. In this project, participants received access to different agricultural inputs and to know-how about farming and irrigation. The program developers found improved solutions for irrigation using equipment already available to most smallholder farmers. Further, they figured out which crops are most valuable for smallholder farmers. The focus was on crops which mature quickly and can help maintain a steady cash flow. In order for the farmers to secure a prompt payback, so that some of the profit can be reinvested, they must select crops with a short growth season. Growing maize, for instance, takes time and there is a risk of either hunger or oversupply – which leads to low prices – depending on the timing in the growth cycle. They also packed important inputs such as plastic pipes for irrigation and fertilizer into smaller, more affordable quantities and made them available to the farmers (Fagerland, 2018). Lessons from the pilot project have been used to form the concept of “Micro Investing”.

A major constraint for smallholder farmers is both access to agriculture inputs and input affordability. Especially poor farmers are not able to save enough money to help themselves increase agricultural productivity, as their income is barely enough to maintain their present level of productivity. Farmers are therefore offered to invest in affordable “Micro Investment Kits” which should increase their profitability and thereby help them lift themselves out of poverty. Rather than providing handouts or credit, the kits consist of agricultural inputs such as improved seeds, fertilizer, and irrigation, combined with knowledge support. The kits are presented as an investment idea to the smallholder farmers. The farmer is approached as a decision maker, rather than a recipient of aid and decides for themselves whether to enter the program or not. The kit was developed to increase the farmers' productivity, which in turn hopefully will increase their income. If they decide to participate, then the farmers will be encouraged to save parts of their new income and invest in more kits. This way, the program is set to facilitate economic growth.

NCA's long term vision is to introduce this concept in as many countries in Sub-Saharan Africa as possible and potentially lift thousands, if not millions, of smallholder farmers out of extreme poverty. An improved version of the project has recently been set in motion in Malawi, aiming to perfect the method of Micro Investing, improving data collection, and assuring better evaluation of the results. Before going into the specifics of Project M, we take a closer look at the economy in Malawi and give an overview of the agricultural sector in the country.

5.3 The economy and agricultural sector of Malawi

Malawi is one of the world's poorest countries. Approximately 70 percent of its 19 million inhabitants live in extreme poverty, that is, on an income below 1.90 USD a day (Roser & Ortiz-Ospina, 2017). The economy is highly dependent on aid assistance from the World Bank and individual donor nations (CIA, 2019). In 2018, Malawi was one of the 10 largest recipient countries of Norwegian development aid (Norad, 2019).

Of the 19 million people living in Malawi, 80 percent live in rural areas (UNESCO, 2019). Most of the rural population work in the agricultural sector (approximately 85 percent), and the agricultural sector contributes to about a third of the country's GDP (Munthali & Murayama, 2013; Team & Region, 2018). The majority of farmers are smallholder farmers. Smallholder farmers are commonly defined as those farmers operating a family farm with less than two hectares of land, although the definition varies across literature (Lowder, Skoet, & Raney, 2016). The fact that so many smallholder farmers live in extreme poverty is curious, as smallholder farmers account for about 80 percent of total food production in Sub-Saharan Africa and Asia (AGRA, 2017).

Smallholder farmers mainly grow food crops to themselves and their family. This is known as “subsistence farming”. They are heavily dependent on rainfed production and they cultivate various crops such as maize, sorghum, millet, sweet potatoes, groundnuts, cassava, soybeans, pigeon peas and vegetables, as well as certain cash crops (i.e., what goes beyond subsistence farming) including tobacco, paprika, and cotton (AFDB, 2006). Tobacco production is the country’s largest and most important cash crop and accounts for about half of the country’s exports. There has long been a trend of falling prices due to a global decline in demand for tobacco. The reduced demand for tobacco is a huge challenge for smallholder farmers who highly depend on it as their major export, and there is a great need to help them transition towards alternatives that adds on value. In addition, the agricultural productivity among smallholder farmers is characterized by low performance. On the continent in general, Africa has had little improvement in yields over the years and are still using traditional production methods.

5.4 The distribution of Micro Investment Kits

NCA are cooperating with local NGOs who are responsible for the packaging of inputs into smaller quantities. Thereafter these Micro Investment Kits are being distributed by so-called agronomists who are paid and trained at the project’s expense. In order to build local knowledge, the agronomists will have regular meetings every three weeks with the smallholder farmers. Having a fixed schedule for when the meetings are held by the agronomists makes it easier for the farmers to plan. However, on a day-to-day basis, it is selected lead farmers who will follow up the implementation of the kits. Their function will be to provide training, communicate information, and organize farmers for later meetings with the agronomists. The lead farmers are frontrunners in smallholder farming who have already invested in Micro Investment Kits. The agronomists are to gradually hand over tasks and knowledge to lead farmers who can help the small farmers gain market access. Planning a gradual exit from the beginning is going to prepare the lead farmers for the day when the intervention is phased out.

5.5 Area of implementation and phase-out

Malawi is divided into 28 districts distributed over three regions: the central, northern and southern region. The central region is divided into nine districts and these nine districts are further subdivided in Extension Planning Areas (EPA). Dowa, which is located in the central region of Malawi, is the first district where the program is initiated. Only parts of Dowa will be included to receive the program at first. The EPAs within Dowa district are Bowe, Mndolera, Mponela, Mvera, and Nachisaka where Mndolera is the first EPA to receive the program. Dowa district is chosen purposively and the EPAs from which each village are selected is selected randomly. NCA is present in the capital of Malawi, Lilongwe, and since Dowa district is bordered by Lilongwe it permits close follow up when evaluating the rollout of the project.

Because of limited resources, the partnering NGOs are present in one area at a time. Before the NGOs move on to another EPA, the kits should have reached a critical mass, and smallholder farmers in the EPA in question should continue to grow without support. Hence, the presence of NGOs in each EPA is temporary and as quickly as possible replaced by “exit partners” such as lead farmers and agripreneurs. When the presence of NGOs is phased out, their competence can be reused in another area. The criteria for when the phase-out can begin is that 50 percent of the targeted group is onboard. These 50 percent consist of early adopters and early majority, where the criteria are that the early adopters need to earn more than 20 dollars a day and the early majority needs to earn more than 10 dollars a day and continue to grow (Concept Description, 2019).

6. An impact evaluation of Project M

In the preceding chapter, we introduced the concept of Micro Investing and NCA's ongoing project in Malawi (Project M). This chapter will follow the theory provided in chapters 3 and 4 and where it is feasible and appropriate, we will exemplify and apply it on the case of Project M, given the information available. Through our suggested impact evaluation design for this specific case, we hope to give valuable insight into what project managers should keep in mind when designing, implementing, and evaluating social programs.

Sections 6.1 through 6.4 follows Sections 3.1 through 3.4 chronologically. First, in Section 6.1, we establish whether the problem that NCA is looking into is really in need of assessment. Then, in Section 6.2, we create a proposed Theory of Change for Project M. In Section 6.3 we formulate an evaluation question for Project M and three hypotheses which need to be true to be able to establish that the project has the intended effect. Section 6.4 mirrors the theory in Section 3.4, and we touch upon the selection of performance indicators for Project M.

Sections 6.5 through 6.8 are based on the theory in Chapter 4. However, not all aspects of the theory are relevant for the case at hand and those parts will thus not be applied. In Section 6.5 we explore suggestions for Project M methodology and design, define the treatment of Project M and suggest what the eligible population, approach to random assignment, unit of randomization, and sample size should be. In line with the theory in Section 4.4, Section 6.6 presents some methodological challenges that our proposed design might be subject to. In Section 6.7 we look at some of the baseline data actually collected by the implementers of Project M. Finally, Section 6.8 gives a short note on our thoughts around the external validity and generalizability of Project M.

6.1 A needs assessment for Project M

In Section 3.1 we explained the importance of assessing whether a problem we are trying to solve is actually a real problem. Now, let us have a look at the problem that NCA is trying to solve. In a concept description of the project, NCA states that: *"By helping smallholder farmers lift themselves out of poverty and grow, we address poverty and food security at the same time"* (Concept Description, 2018).

In the following subsections, we will first look into whether the problem NCA is trying to address is really in need of a solution and then what the source of the problem is. Then, we will look at what NCA's proposed solution to the problem is, whether this proposed solution is a new idea, and whether this idea might actually be a good solution.

6.1.1 The problem at hand

More than 700 million people worldwide live in extreme poverty today (Roser & Ortiz-Ospina, 2017). Extreme poverty is for the most part concentrated in South Asia and Sub-Saharan Africa (Mellor, 2017, p. 209). In addition, poverty is concentrated in rural areas with over 78 percent of the poor located here. An estimated 63 percent of the world's extreme poor are working in agriculture – most of them smallholder farmers (Olinto, Beegle, Sobrado, & Uematsu, 2013, p. a5).

World poverty and world hunger are without a doubt two of the world's most pressing problems. In fact, no poverty and zero hunger are number one and two respectively of the 17 Global Goals for Sustainable Development (United Nations). It is evident that the problems NCA are looking to solve, poverty and food security, are indeed in need of a solution. For clarity, we will define the main problem as "poverty and hunger among smallholder farmers in Sub-Saharan Africa". Next, we need to identify the source of this problem.

6.1.2 The source of the problem

In the decades following 1960, the developing world and East Asia in particular witnessed an impressive growth in agricultural productivity. This period is often referred to as “the Green Revolution”. Although the population more than doubled in the 50 years following 1960, the proportion of people suffering from hunger fell by half in the same period (Wik, Pingali, & Brocai, 2008). The production of cereal crops tripled even though land area cultivated only increased 30 percent (Pingali, 2012), implying an increased productivity. This remarkable increase in production could be attributed to a combination of the high rates of investment in crop research, infrastructure, market development, and policy support taking place during the period from the 1960s to the 1980s (Pingali, 2012).

However, not all regions followed the same trend. Sub-Saharan Africa stands out as the only region in which per capita agricultural output did not see a sustained increase during the Green Revolution. Latin America and South Asia had a small increase, while East Asia and the Pacific had an increased agricultural production per capita by almost 80 percent (Wik et al., 2008, p. 4).

In summary, it seems like the main source of the problem stated above is the low productivity amongst smallholder farmers in Sub-Saharan Africa. Now that we have identified the problem and the source of the problem, we can begin to look for a solution.

6.1.3 The proposed solution

NCA is a non-governmental charity organization. Their funding comes from public grants and collected funds and they work in a number of poor countries to improve the lives of civilians in need (Innsamlingkontrollen, 2019). NCA’s proposed solution is to help smallholder farmers lift themselves out of poverty and create market-based growth over time. By cooperating with local NGOs, their aim is to make sure that inputs and know-how are affordable and available to small farmers. Their reasoning is that *“modern agricultural inputs don’t reach the small farmers in affordable* *quantities, neither does their produce reach urban and international buyers with purchasing power that could fuel development"*. This opens up for three different, but related, questions: first, is it true that modern agricultural inputs do not reach the small farmers? Second, does agricultural development lead to economic growth? And third, does foreign aid fuel agricultural development?

6.1.4 Existing solutions

We need to establish whether it is true that smallholder farmers in Malawi does not have access to modern know-how and affordable inputs. The idea to provide smallholder farmers with basic, important inputs to boost agricultural productivity is not new. Input subsidies have long been a controversial topic in agricultural development. In a systematic review, Hemming et al. (2018), found that agricultural input subsidies are associated with higher agricultural yields and increased income. With rapid learning about the use and benefits, the subsidies are only needed for a short period of time and can thereafter be phased out (Chirwa, 2013). However, the impact has become increasingly questioned because the input subsidies are not distributed to the poorest households, instead it tends to benefit the wealthier farmers (Ricker-Gilbert, Jayne, & Shively, 2013).

The Malawi "starter pack program" launched in 1998/99 were to provide almost all rural smallholder farmers with a free pack of inputs to increase agricultural productivity. About 2.86 million starter packs were distributed among the Malawian population (Chibwana, Fisher, Jumbe, Masters, & Shively, 2010; Duflo, Kremer, & Robinson, 2004). Later on, however, the program was substituted with the Targeted Input Program and limited to only a number of households (Harrigan, 2008).

Even though input-kits already exist, there seems to be some form of market failure, as they are not distributed properly and are often not available to small farmers. According to the baseline study conducted in Mndolera EPA in the district of Dowa, large input supply companies are located in the main cities providing a range of fertilizer, pesticides, and seeds (Baseline Report, 2019). Part of the problem is that most farmers are not linked to a well-functioning market which makes them unaware of the range of crops and inputs that exist. In other words, there is incomplete market information, and constraints are put on the possibilities for smallholder farmers to diversify away from primary staple crops and become more efficient.

Findings from the baseline report shows that there is room for improvement when it comes to seeds, fertilizer, and especially pesticides (Baseline Report, 2019). Around 40 percent of the farmers reported that they only have access to traditional seeds or no access at all, while the remaining respondents have access to improved seeds or both. Furthermore, 61 percent of the respondents apply inorganic (improved) fertilizer, while 39 percent reported that they solely apply organic (traditional) fertilizer or no fertilizer at all. As much as 32 percent of the respondents reported that they do not use any pesticides, while 6 percent use traditional means to combat pests. The remaining respondents apply commercial pesticides or both.

Smallholder farmers in Malawi seem to have little or no access to extension services which could help promote crops and innovations in their farms. The responses differed between the sections surveyed, but in general, only 17 percent of the target population reportedly had access to such extension services.

One of the main ideas behind the Micro Investing program is to facilitate the meetings of small farmers with the providers of inputs and know-how in the market. In sum, there are two interventions: the actual logistics of getting the kits out to the farmers, and the kits in themselves along with information about how to use them. Even though the Micro Investment Kit is not an entirely new idea, there is still a need to solve the coordination problem around inputs and information.

6.1.5 Agricultural growth, foreign aid, and economic growth

As NCA’s goal is to create market-based growth and not just a one-time improvement of the livelihoods of a small group of people, we need to make sure that foreign agricultural aid can have a positive impact on economic growth. The effect of foreign aid on agricultural growth and on economic growth in general is a disputed topic. We can divide the different viewpoints into three main schools of thought: 1. Foreign aid does not have a positive impact on economic growth. It should be noted that some advocates belonging to this group strongly believe that foreign aid actually has a *negative* impact on economic growth (See, for example, Mosley (1986) or Doucouliagos and Paldam (2015)) 2. Foreign aid can have a positive impact on economic growth, but agriculture is in itself a symptom of poverty and aid and resources should thus be directed elsewhere to help the country in question move from an agrarian- towards an industrialized economy 3. Agricultural foreign aid can have a significant positive impact on agricultural growth and agricultural growth drives overall economic growth.

Through a brief review of existing literature, we will explore whether the vast number of farmers in Sub-Saharan countries is simply a symptom of poverty or if growth in the agriculture sector can drive economic growth. Furthermore, we will review whether agricultural growth can be sparked by foreign aid to agriculture.

Agricultural growth and poverty reduction

Starting in Great Britain in the late 18th century before gradually spreading across the rest of Europe and America in the 19th century, the Industrial Revolution set off a period of unpreceded economic expansion in the western world. During the Industrial Revolution, societies moved from being mainly agrarian to highly industrialized. This involved the introduction of new technology, modernization, mechanization and mass production, urbanization, and improved infrastructure. Ever since the Industrial Revolution and its immense effect on economic growth, the world has been waiting for today’s developing countries to take the same path. However, over the last decade, a growing body of literature that questions and refutes the notion that growth is best achieved through structural change and industrialization has emerged.

Ravallion and Datt (1996) used data for poverty measures in India spanning 40 years and found that rural growth reduced poverty in both rural and urban areas whereas urban growth had no impact on rural poverty. Moreover, shifts from rural to urban population had no significant impact on poverty. In another study, they found that higher agricultural growth rates was associated with poverty reduction (Ravallion & Datt, 2002).

Following the same line, Thirtle, Lin, and Piesse (2003) documented that the sector of growth does matter and that agricultural productivity growth had a substantial impact on poverty reduction in Asia and Africa.

Christiaensen et al. (2011) concluded that growth in agriculture is especially beneficial for the poorest of the poor, while non-agricultural growth is more powerful in reducing poverty among the poor just above the threshold for extreme poverty. Data in this study included Sub-Saharan Africa.

A World Bank report from 2008, documented that growth in agricultural production reduces poverty at least twice as effectively as growth generated by other sectors and stated that agriculture can be the lead sector driving overall economic growth in the agrarian countries (Mondiale, 2008, p. 6 & 29). Furthermore, the report points to a “productivity revolution in smallholder farming” as the key to economic growth in agricultural-based countries (p. 1).

Throughout his book “Agricultural Development and Economic Transformation”, John W. Mellor argues strongly that the way to increase agricultural production is to increase productivity of the non-poor, small commercial farmer (2017). Increased agricultural production will in turn promote economic growth and reduce rural poverty. Mellor argues that “the bulk of agriculture in every low- and middle- income country has the potential for a lengthy period of rapid “catch-up” growth” (p. 64).

Foreign aid and economic growth

Juselius, Møller, and Tarp (2014) conducted a multivariate time series analysis and concluded that aid has had a positive long-run impact on the macroeconomy in the majority of countries in Sub-Saharan Africa. In 27 of the 36 Sub-Saharan countries studied, aid has had a significantly positive effect on investment, GDP or both and only in two countries was the effect of aid significantly negative on investment or GDP.

Alabi (2014) found in his econometric analysis of 47 Sub-Saharan countries that the impact of foreign agricultural aid on agricultural productivity is positive and significant at the 10% significance level.

Some scholars also stress the importance of institutional support and that aid should be directed at countries with favorable policy environments. Burnside and Dollar (2000) found that foreign aid has a positive impact on growth of per capita GDP in developing countries where fiscal, monetary, and trade policies are directed at growth but has little effect in countries where policies are poor. However, the quality of policy has only a small impact on the allocation of aid and in the period studied (1970-1990) donors were not favoring good policy environments, implying that aid might have had a greater overall impact on growth in the developing world had it been systematically conditioned on good policy. Intuitively, if aid allocations are made systematically towards good policy environments, foreign aid could in turn begin to affect policy-making.

Mellor (2017) states that *"Aid to agriculture must be in the context of a clear national plan for rapid growth with a national level of support which is substantial and growing relative to the total to provide a high level of long-term development."*. He even goes as far as saying that without national support, foreign aid should not be provided. Following this bold statement, he notes that the government of many African countries are likely to struggle to make a commitment of the sort. His belief is that failure by governments to prioritize agriculture is the reason why poverty reduction has slowed in much of Sub-Saharan Africa and a few low- and middle-income countries in Asia.

Furthermore, it is not only the amount of aid that matters, but different types of aid can have different impacts. Albi (2014) found that while bilateral aid can have a bigger influence on agricultural productivity, multilateral aid can have a greater impact on agricultural contribution to GDP growth.

According to Mellor (2017, p. 209), *"foreign aid now gives special attention to Africa where the agriculture sector-specific allocations increased by 2.3 times from 2002 to*

2010. However, the proportion [of aid] allocated to agriculture held steady at between 6.5 and 7.8 percent of the total. That […] hardly shows a high priority to agriculture or, implicitly, to poverty reduction”. This implies that if foreign agricultural aid has a positive impact on growth, there is indeed room for more social programs aimed at agricultural growth in Sub-Saharan Africa.

6.1.6 Need of assessment conclusion

We have established that many smallholder farmers in Malawi do not have access to the modern agricultural inputs and information needed to increase productivity. Increased productivity in smallholder farming could lead to agricultural growth, and agricultural growth might both reduce poverty and drive overall economic growth. Furthermore, foreign aid can have a positive impact on agricultural growth, but the nature, origin, and purpose of the aid will influence the effect it has. Given the potential for poverty reduction from agricultural growth, it could be beneficial to allocate a larger part of aid to agriculture. However, foreign aid alone is not sufficient to tackle the problem, and respective governments need to make serious commitments to agriculture if the sector is to grow and poverty is to be reduced. It is important that foreign agricultural aid is aimed at supporting change in agricultural policy-making to create synergies between the aid and the domestic government expenditures.

With our need of assessment questions answered, we can move forward to the next step in the evaluation process. In the following section, we will develop a framework which should be the foundation for any impact evaluation and begin to answer the question of *how* an intervention will lead to the desired change.

6.2 A Theory of Change model for Project M

Following the preceding discussion, the question that needs to be answered is: “How can we increase the income of smallholder farmers through improved productivity in farming?” This is the problem that NCA is aiming to solve through the concept of Micro Investing.

In Section 3.2, we established the importance of creating a framework encompassing the causal logic of a social program. Figure 2 maps out a causal chain starting with the overarching goal of increasing the income of smallholder farmers and working backwards mapping out the assumptions that must be in place, the primary outcomes, and lastly, the interventions that is the program itself. Hence, Figure 2 forms a simple ToC for the NCA-project. This ToC framework will act as a foundation for our forthcoming analysis of the case in question. For starters it is the base for the formulation of our evaluation question and accompanying hypotheses in the subsequent section.

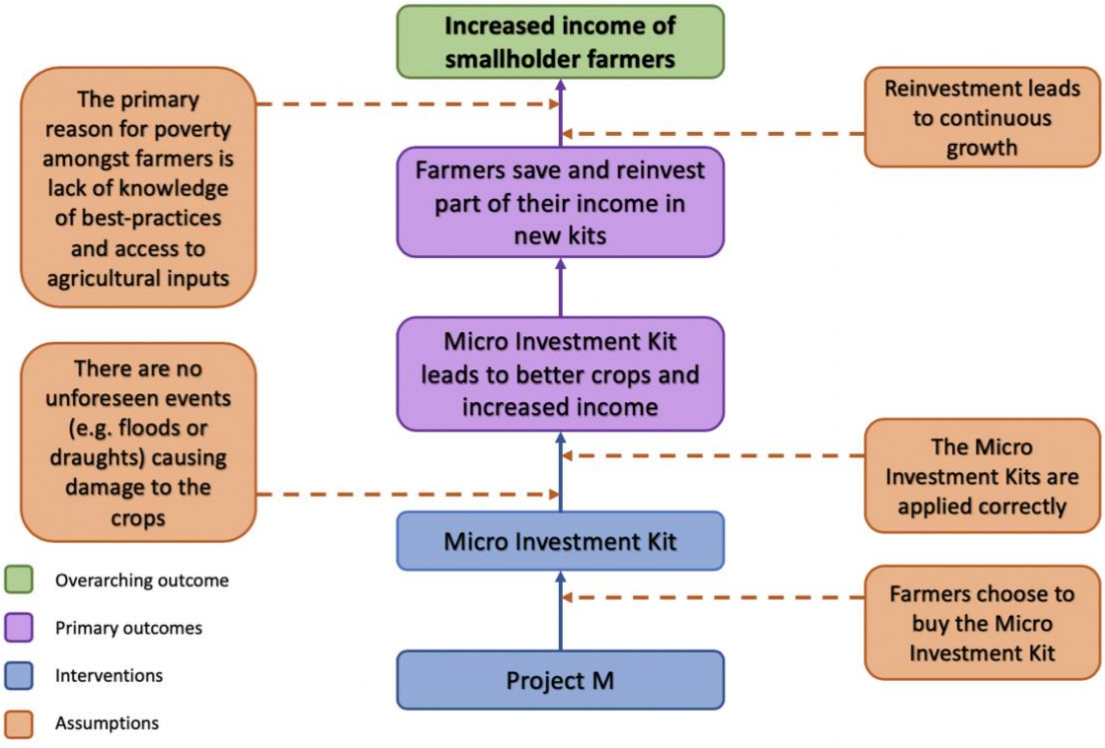
Increased income of smallholder farmers
The primary reason for poverty amongst farmers is lack of knowledge of best-practices and access to agricultural inputs
Reinvestment leads to continuous growth
Farmers save and reinvest part of their income in new kits
Micro Investment Kit leads to better crops and increased income
There are no unforeseen events (e.g. floods or draughts) causing damage to the crops
The Micro Investment Kits are applied correctly
Micro Investment Kit
Farmers choose to buy the Micro Investment Kit
Project M
Overarching outcome
Primary outcomes
Interventions
Assumptions

Figure 2: A proposed simple Theory of Change model for Project M.

6.3 Evaluation question and hypotheses for Project M

From the proposed ToC model developed in Section 6.2, we see that the basic evaluation question in the case of the NCA-project is: “What is the impact of Project M on the income of smallholder farmers?” Using the ToC model as a guideline, we have formulated three clear, testable, and quantifiable hypotheses:

> *Hypothesis 1:* The presence of Project M increases the probability that a farmer buys a kit aimed at improving crops.

There is an increasing supply and availability of better seeds, fertilizer, pesticides, and irrigation systems in Malawi and other Sub-Saharan countries. Hence, it is not given that NCA's presence will increase the probability that a given farmer buys the kit. However, as explained in Subsection 6.1.4, there is a coordination problem between the existing inputs and smallholder farmers. There is thus a chance that the presence of Project M increases the probability that a farmer buys an investment kit.

> *Hypothesis 2:* The kit provided by Project M increases the return of the farmer's crops and her/his income.

Following the discussion in Subsection 6.1.2, it seems plausible that the income of small farmers will rise if they have access to modern agricultural inputs and technology. If both Hypothesis 1 and 2 is true, the project can be said to improve the livelihoods of all smallholder farmers buying the kit.

> *Hypothesis 3:* The kit provided by Project M increases the farmers' willingness to continuously save and reinvest part of their income in new kits or more vegetable beds. The reinvestment leads to further growth.

Findings in the baseline report indicate that farmers are willing to reinvest into agriculture if they have the means for it (Baseline Report, 2019). Around 70 percent of the respondents had set a relatively high priority of reinvesting into agricultural inputs. If Hypothesis 1, Hypothesis 2, and Hypothesis 3 are all true, Project M can be said to lift smallholder farmers out of extreme poverty. Following the reasoning of Mellor (2017), this could in turn accelerate economic growth and reduce overall poverty.

6.4 Performance indicators for Project M

To be able to determine the impact of Project M on the income of smallholder farmers, we need to decide upon some measurable performance indicators. As explained in Section 3.4, these indicators should be specified with the help of the projects ToC and should be as clear and unambiguous as possible. The outcome indicators should be selected with the help of the main stakeholders in the evaluation team from both the policy- and research branch of the team.

Looking at our ToC model from Section 6.2, we see that we at least need indicators to answer this list of questions:

1. How many of the farmers who have access to the Micro Investment Kits choose to purchase a kit?
2. Do the farmers who have purchased kits actually make use of the kits?
3. How much does the productivity of the farmers who have purchased one kit increase?
4. What percentage of farmers reinvest in more kits?
5. How much more does the productivity of farmers increase if they buy more than one kit?

Question 1 will tell us something about how far we are from full-compliance (see Subsection 4.4.5) and will help us test Hypothesis 1 from the preceding section. Question 2 is important for assessing whether the project is implemented and running as intended. Question 3 will help us test Hypothesis 2. Questions 4 and 5 will help us test Hypothesis 3 as well as review whether the project is successfully implemented.

Answering these questions is an important step in the process of evaluating the impact of Project M. A traditional result-based monitoring would end after answering and evaluating such questions. However, to get the *ceteris paribus* impact of Project M on the income of smallholder farmers, we have to construct a representation of the counterfactual and compare it to the income of smallholder farmers who have purchased kits. That is, we need to introduce the elements of an experiment into the project design.

6.5 Project M as a field experiment

6.5.1 The Project M methodology and design

If RCTs, or field experiments, are the gold standard of impact evaluations, it follows that we should ask ourselves if the project of NCA can sensibly be designed as such an experiment. It is our belief that the project could very well have benefitted from being conducted as a field experiment. The project taking place in Malawi is already underway, we nonetheless hope that a description of how the project could have been designed as a field experiment, will provide valuable insight into how to conduct an impact evaluation in practice.

6.5.2 The Project M treatment

The treatment in the case at hand is the application of an investment kit provided by NCA and their partners, which consists of valuable inputs and know-how. As previously explained, the investment kits are affordable by the standard of the small farmers. There are no handouts, and the purchase of a kit is entirely optional. Hence, the farmers in the treatment group will *self-select* to receiving treatment. The treatment group is the entire sample which are offered to purchase the investment kit. To clarify, they do not themselves select whether they are part of the treatment group (whether they are offered to purchase the kit), but the farmers already in the treatment group select whether they want to receive treatment (whether they actually purchase the kit).

Because of the self-selection, we can be certain that there will be imperfect compliance (see Subsection 4.4.5). As stated above, we should tackle this either by analyzing the intention-to-treat where we would get the average treatment effect, or by using an instrumental variables method to get the average treatment effect on the treated.

6.5.3 The Project M population

The broad population of interest is all smallholder farmers in Sub-Saharan Africa. However, as already established, this particular project is taking place in Malawi. Drawing a sample from Malawi as a whole will not be possible for logistical reasons. As described in Section 5.5, Malawi is divided into three regions, which are made up of 28 districts in total. The districts are divided into several extension planning areas (EPAs), and these EPAs consists of sections.

As the project has restricted resources, the project managers have decided to start the rollout of the project in the district of Dowa located in the Central Region. Dowa district consists of nine EPAs, and in the first round, the project is being implemented in Mndolera EPA. Mndolera EPA consists of nine sections and 234 villages. Our suggestion is that for the purpose of an impact evaluation, the eligible population of this project should ideally be all sections in the Central Region of Malawi. This, of course, depends on whether the required sample size does not surpass the number of sections.

6.5.4 The Project M unit of randomization

Because of logistical, economic, and ethical concerns, Project M cannot be efficiently implemented at the individual or household level. If the unit of randomization for Project M was households, the project implementers would have to go door-to-door and offer the investment kit only to the households randomly selected for treatment. This would not only be incredibly impractical and costly but would most likely generate tensions when households from the control group see that their neighbors are offered something that they are not. In addition, a lower level of assignment increases the risk of spillover effects.

When a higher level of assignment is chosen, there are fewer units to assign to treatment and control, even though the total number of individuals receiving treatment might be larger. As explained in Subsection 4.3.4, for interventions implemented at higher levels, an even larger sample is needed to be able to detect a program’s true impact (Gertler et al., 2016, p. 198). Together, these two facts make higher levels of assignment subject to risks of damaging internal validity. The goal should therefore be to find the smallest unit randomization that is operationally feasible.

An imaginable level of intervention in the case at hand is villages. However, at the village level, there will most likely still be high risks of spillover effects between the nearby villages. Moreover, this will require the implementing NGOs and agripreneurs to move around a lot. The next level of already existing geographical areas are the sections mentioned in the preceding subsection. We believe that the smallest operationally feasible unit of intervention is these sections and, hence, our suggestion is that the level of assignment should be sections.

6.5.5 The Project M approach to random assignment

Our suggestion for how NCA should approach random assignment is through pipeline randomization. As explained in Subsection 4.3.3, in pipeline (or phase-in) randomization, all units in the population will eventually receive treatment. The order in which they receive this treatment is random, and at the end of a predetermined period, data is collected on all units and some of the groups make a permanent switch from control to treatment. In the end, there are no control groups left.

To the best of our knowledge, NCA is planning to roll out the Micro Investing program in at least the whole Central Region of Malawi.With these ambitions, simple randomization would not serve its purpose. At the same time, the program has limited resources, and cannot possibly roll out the program to the whole region at once. In an annotation from NCA, the NGOs present in each section will have a planned phase-out (see Section 5.6), and because of this early exit, they will have renewed capacity to move to another section. Hence, a phase-in design seems like the natural choice for Project M’s approach to random assignment. Although a pipeline design might act as an incentive for the control group to contribute to data collection, this approach is not safeguarded against the risks of bias.

Even though the Micro Investing program is scheduled to roll out in the whole Central Region, it is not necessarily of added value for the evaluation team to collect endline data on every single section in the region. As long as there is external validity to the population of interest (see Section 6.8), the project managers should be able to trust that the intervention has a positive effect on the remaining sections of the Central Region as well. This logic is resting on the assumption that the impact evaluation reveals a significant positive effect after the concept has been implemented in enough sections to satisfy the required sample size. This argumentation lends support for the project implementers to use simple randomization.

6.5.6 The Project M sample size

Because of limited access to data, calculating the minimum sample size for Project M is beyond the scope of this analysis. As explained in Subsection 4.3.4, the sample size should be calculated through power calculations and prepared by professionals. The sample size for the case at hand will be the number of sections in the Central Region of Malawi which will later be randomly assigned to either treatment or control. We emphasize that the sample size must be sufficiently large in order for the experiment to be sensitive enough to detect outcome differences between the treatment and control groups (J-PAL, 2017). In addition, the issue of imperfect compliance, which as stated above is a highly relevant matter for the project at hand, needs to be taken into account when deciding on the sample size (Duflo et al., 2007, p. 33).

6.6 The Project M methodological challenges

Heterogeneous treatment effects

As explained in Subsection 4.4.2, heterogeneous treatment effects occur if responses to treatment differ systematically across different subgroups of the sample. One common demographic to analyze is gender. Malawi is ranked as number 148 on the UNDP Gender Inequality Index (2018). Norway, in comparison, is ranked as number 5. Research by the Food and Agriculture Organization of the United Nations have found that women farmers are 20 to 30 percent less productive than men, and importantly that women do not have equal benefits such as training, information, and knowledge (Diouf, 2011, p. 4). Furthermore, a study of the agricultural gender gap in Malawi, found that input subsidies significantly increased modern maize cultivation by female household heads (Fisher & Kandiwa, 2014).

Other demographics that might be interesting to single out are age, years of education, and household income. It could very well be that differences in these factors will affect the impact of the investment kits. For example, the literacy rate among the population in Malawi that is 15 years or older is only 62.14 percent (UNESCO, 2019). Being illiterate will most likely hamper with the benefit of a kit, as some of the information may be in written form.

A critical prerequisite for farmers to be able to grow is access to land. In Malawi, farmers can own- or rent land and land tenure is further divided into customary land, public land, and private land (FAO, 2019). It could be interesting to assess whether land ownership interferes with the intervention of the kit. Finally, it is possible that existing access to extension services could affect the average treatment effect of the investment kit as less farmers would purchase the kits.

To tackle the risk of heterogeneous treatment effects, we hence suggest that the project managers should prepare for subgroup analyses of gender, age, years of education, literacy status, household income, land ownership, and access to extension services. Data should therefore be collected on all these factors.

Spillovers

Since we have established that the project should be designed as a cluster randomized trial, a great part of the risk of spillovers should already be eliminated.

Imagine that treatment was assigned on the individual level. If you as a farmer notice that your neighbor’s crops are growing faster than before, chances are that you will actively try to figure out the reason for your neighbor’s newfound success to get your hands on the same inputs. Maybe the NGO providing the kits does not have the heart to turn you down, and just like that, you have crossed over from control to treatment.

Or maybe you and your neighbor are good friends and when you ask her, she willingly shares her new knowledge about best practices and some of her new seeds and fertilizer. All of a sudden you are contaminated through social interaction. Or maybe you make no notice of your neighbor's lush farmland and continue business as usual. When the time comes to harvest your crops, they are ruined because some of your neighbor's new pesticide which is too strong for your seeds have blown over on your farmland. You have become victim to negative externalities. These are all entirely hypothetical scenarios, of course, but they illustrate the huge risk of contamination when randomizing a social project on the individual level.

By randomizing on section level, the risk of contagion is minimized but not eliminated. An important factor when considering contamination is the geographical distance between control and treatment groups. To avoid contamination there should be a geographic separation between the treatment and control areas while at the same time they should be close enough to be comparable (White & Raitzer, 2017, p. 37).

General equilibrium effects

When implementing social programs, for example in introducing modern technology to farmers, general equilibrium effects may arise if the project is affecting a wider area than the local economy, such as input and output prices or wages. This is supported in a paper written by Svensson and Yanagizawa-Drott where they look at the impact of providing farmers access to price information. This paper emphasizes the importance of measuring general equilibrium effects in the case of policy interventions. They conclude that linking farmers to the market by giving them access to price information may *"help poorly functioning markets work better, improve farmers' bargaining positions, and thereby increase the incomes of the poor"* (2012, p. 25). Also, Schuh (2000, p. 234) states that the focus should shift towards the general equilibrium effects of providing improved technology into the agricultural sector.

Although general equilibrium effects are difficult to measure, we recommend that project managers should consider the possibility of such effects affecting the smallholder farmer sector when analyzing the results of the project.

6.7 Project M data collection

As described in Subsection 4.5.1, after assigning units to treatment and control, the researchers should collect baseline data on the performance indicators. Data should also be collected on any subgroups of interest. While the project is underway, continuous data collection will provide managers with critical information about project implementation.

The nature of Project M suggests that data collectors will have to collect substantial amounts of primary data. Implementers of Project M have already executed a baseline data collection. The baseline study was conducted in the district of Dowa and involved the collection of firsthand data through personal interviews recorded electronically with survey software (Baseline Report, 2019). Before the survey was implemented, a pilot survey was carried out to evaluate the quality of the questionnaire and make appropriate adjustments. The survey was designed to capture demographic characteristics such as gender, age, level of education, land ownership, and access to extension services. In addition, the questions covered the farmers involvement in the horticultural value chain.

As the farmers self-select into treatment, the project implementers are faced with more of a challenge when it comes to the collection of data on the distribution and application of Micro Investment Kits. They will have to collect data on both the purchase rate and the reinvestment rate, and, to keep tabs on any subgroups, the characteristics of the buyers. Collection of some secondary data is probably also useful. The researchers should investigate the availability of existing data on information such as literacy rates.

Baseline and midline data collection on all performance indicators and other factors such as outside influences are important, however, the protagonist of the impact evaluation narrative, is the endline data collection on the outcome variable of interest. As we see from our evaluation question (see Section 6.3), we need to collect data on the income of smallholder farmers from the treatment and control groups. An example of concrete data collection for income of smallholder farms is average net income per household per day from farming derived from price and volume of horticultural crops.

When data on smallholder income is collected from a large enough sample of households from the treatment and control groups (see Subsection 4.5.3), we can compare the mean values of this outcome variable. Note that since the level of assignment is clusters not individuals and as there is imperfect compliance and most likely also heterogeneous treatment effects, we will need a larger sample size to get an internally valid and statistically significant result.

6.8 A note on the external validity of Project M

As explained in Section 4.6, internal validity does not imply external validity. To be able to judge whether the results from Project M, if conducted as a social experiment, would be generalizable to a broader setting, we would have to combine the rigorous evidence from our impact evaluation with lessons learned from the monitoring (process evaluation) and knowledge about local contexts. Even though we can never say with 100 percent certainty that a project will work if transferred to other locations and other contexts, it is important to make the most of the impact evaluations we go through with. Impact evaluations are expensive, time-consuming, and require effort and devotion. Results from one impact evaluation should therefore be utilized in other contexts if and where they are relevant. The combination of theory, knowledge about local context, descriptive evidence, *and* the results of rigorous impact evaluations can help us answer whether results from one program are likely to replicate in another and whether we need to conduct new evaluations. This globally informed, locally grounded way of thinking is the essence of evidence-based policy-making (J-PAL, 2019b).

As a start we should look at the Central Region of Malawi compared to the two remaining regions: Northern and Southern. As mentioned in Section 5.3, 80 percent of the total population in Malawi is rural and approximately 85 percent of rural households are smallholder farmers. If treatment was randomized from the Central Region on the section level, and if the results from the impact evaluation showed a significant positive impact, we believe, based on what knowledge we have, that the program could be successfully scaled up to the national level.

External validity across countries is difficult to claim, as contexts and climate tend to differ dramatically. Moreover, we have to take into account any cultural and institutional differences as they strongly affect behavioral differences and implementation. Implementing a project in a different country without any further analysis also means that there are potential interaction effects with other, existing treatments. Some Sub-Saharan countries may be somewhat similar when it comes to climate and sectoral composition. It is not unthinkable that the program, at least with certain modifications, could be successfully implemented in a few countries across Sub-Saharan Africa. Outside Sub-Saharan Africa, the external validity would probably notably diminish or vanish all together.

7. Limitations and discussion

This thesis has some limitations. Although an attempt was made to present all the main aspects regarding impact evaluations using state of the art methodology, the framework should be seen in light of the time constraints placed upon us. It should be emphasized that this thesis is by no means a complete recollection of the existing literature on impact evaluations. The emergence of experimental designs in the evaluation of social projects is relatively recent, yet the methodology behind experimental studies has been a subject of research for the better part of the previous century. The assumptions and statistical laws behind randomized controlled experiments are complex enough to be a field of study in their own right. Moreover, RCTs are not the only way to produce rigorous results. Each of the methods described in Chapter 4 could be used to estimate the *ceteris paribus* impact of social programs. To go into details about this methodology would, however, be quite technical and is beyond the scope of this thesis.

The applied parts of the thesis also face certain limitations. The timeframe of the thesis versus NCA's "Project M" did not coincide in ideal way. The rollout of the project took place too early for the thesis to contribute to the project design, but too late for us to be able to analyze endline data from the project. The people of NCA have been immensely helpful and easy to correspond with. However, information available to date is limited, which made a complete and detailed analysis of Project M impossible. Because NCA is still working on completing the report regarding the pilot project it was difficult to say anything substantial about the outcome of the specific project.

8. Concluding remarks

This thesis provides a framework for conducting impact evaluations of specific social programs through result-based monitoring and experimental designs. Monitoring should provide a continuous flow of information, which is necessary for the evaluation team to assess whether the implementation of a project is going as planned, while experimental designs can generate rigorous evidence of any impacts on final outcomes.

Through an application of the framework provided, we have illustrated how the ongoing project of Norwegian Church Aid in Malawi could have been executed as a field experiment. An impact evaluation of this experiment should give internally valid results. We explain how this project ideally should have been conducted with the idea of implementing an impact evaluation into the design from the very beginning. Furthermore, the framework provided should be useful for anyone on the outset of designing a social project as it exemplifies the process through a combination of theory and practical suggestions.

Proper impact evaluations are complex and costly and it should be noted that supporters of randomized controlled trials are facing criticism that the superiority of this methodology is overrated (See, for example, Deaton and Cartwright (2018) or Worrall (2007)). It is nonetheless our belief that making proper impact evaluations the standard practice in aid projects whenever feasible and economically viable would be a crucial step towards poverty eradication. Result-based monitoring together with evaluation of endline data should make it much more plausible for the evaluation team to be heard when attempting to affect policy-making. Changes in policy-making are what ultimately will lead to improving the lives of the poor on a national and international level.

References

AFDB. (2006). *Smallholder crop production and marketing project*. Retrieved from Food and Agriculture Organization of the United Nations: https://www.afdb.org/fileadmin/uploads/afdb/Documents/Project-and-Operations/Malawi_-_Smallholder_Crop_Production_and_Marketing_Project_-_Appraisal_Report.pdf

AGRA. (2017). Africa Agriculture Status Report: The Business of Smallholder Agriculture in Sub-Saharan Africa (Issue 5). In *Nairobi, Kenya: Alliance for a Green Revolution in Africa (AGRA).* (pp. 141).

Alabi, R. A. (2014). *Impact of agricultural foreign aid on agricultural growth in Sub-Saharan Africa: A dynamic specification* (Vol. 6): Intl Food Policy Res Inst.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*: Princeton university press.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*: Princeton University Press.

Attanasio, O. P., Meghir, C., & Santiago, A. (2011). Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA. *The Review of Economic Studies, 79*(1), 37-66. doi:10.1093/restud/rdr015

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics, 7*(1), 1-21.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445-469.

Bracht, G. H., & Glass, G. V. (1968). The External Validity of Experiments. *American Educational Research Journal, 5*(4), 437-474. doi:10.3102/00028312005004437

Burnside, C., & Dollar, D. (2000). Aid, policies, and growth. *American economic review, 90*(4), 847-868.

Camerer, C., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., . . . Wu, H. (2016). *Evaluating replicability of laboratory experiments in economics* (Vol. 351).

Center for Theory of Change. (2019). What is theory of change? Retrieved from https://www.theoryofchange.org/what-is-theory-of-change/

Chabrier, J., Hall, T., & Ben, S. (2017). Implementing randomized evaluations in government. In *Lessons from the J-PAL state and local innovation initiative*: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Chapman, N., Lloyd, R., Villanger, E., & Gleed, G. (2017). *The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation*. Retrieved from Norad: https://norad.no/en/toolspublications/publications/2017/the-quality-of-reviews-and-decentralised-evaluations-in-norwegian-development-cooperation/

Chibwana, C., Fisher, M., Jumbe, C., Masters, W. A., & Shively, G. (2010). Measuring the Impacts of Malawi's farm input subsidy program. *Available at SSRN 1860867*.

Chirwa, E. W. (2013). *Agricultural input subsidies : the recent Malawi experience*. In A. Dorward (Ed.).

Christiaensen, L., Demery, L., & Kuhl, J. (2011). The (evolving) role of agriculture in poverty reduction—An empirical perspective. *Journal of development economics*, *96*(2), 239-254.

CIA. (2019, 03.04/19). The world Factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/mi.html

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Dhaliwal, I., & Tulloch, C. (2012). From research to policy: using evidence from impact evaluations to inform development policy. *Journal of Development Effectiveness*, *4*(4), 515-536. doi:10.1080/19439342.2012.716857

Diouf, J. (2011). Women - Key to Food Security.

Doucouliagos, H., & Paldam, M. (2015). 20 Finally a breakthrough? The recent rise in the size of the estimates of aid effectiveness. *Handbook on the economics of foreign aid*, 325.

Duflo, E. (2004). The medium run effects of educational expansion: Evidence from a large school construction program in Indonesia. *Journal of Development Economics*, *74*(1), 163-197.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, *4*, 3895-3962.

Duflo, E., Kremer, M., & Robinson, J. (2004). Understanding technology adoption: Fertilizer in Western Kenya, preliminary results from field experiments.

Fagerland, J. (2018). Ending Hunger - A New Approach. *TEDxOslo*. Retrieved from https://www.youtube.com/watch?v=KHyNc0GzRr4&list=PLNeZMRGp1v-2utAk5lamWrOpva13Jb3Wy

FAO. (2019). Gender and Land Rights Database. *Malawi: Prevailing systems of land tenure*. Retrieved from http://www.fao.org/gender-landrights-database/country-profiles/countries-list/land-tenure-and-related-institutions/en/?country_iso3=MWI

Field, E., Glennerster, R., & Hussam, R. (2011). Throwing the baby out with the drinking water: Unintended consequences of arsenic mitigation efforts in Bangladesh. *Work. Pap., Dep. Econ, Harvard Univ., Cambridge, MA*.

Fisher, M., & Kandiwa, V. (2014). Can agricultural input subsidies reduce the gender gap in modern maize adoption? Evidence from Malawi. *Food Policy, 45*, 101-111.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*: WW Norton.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice, Second Edition*. Washington, DC: Inter-American Development Bank and World Bank.

Harrigan, J. (2008). Food insecurity, poverty and the Malawian Starter Pack: Fresh start or false start? *Food Policy, 33*(3), 237-249.

Hemming, D. J., Chirwa, E. W., Dorward, A., Ruffhead, H. J., Hill, R., Osborn, J., . . . Coffey, C. (2018). Agricultural input subsidies for improving productivity, farm income, consumer welfare and wider growth in low-and lower-middle-income countries. *Campbell Systematic Reviews, 14*.

Innsamlingkontrollen. (2019). Kirkens Nødhjelp/Norwegian Church Aid.

J-PAL. (2017). Introduction to Evaluations. from The Abdul Latif Jameel Poverty Action Lab https://www.povertyactionlab.org/sites/default/files/resources/Introduction-to-Evaluations.pdf

J-PAL. (2019a). Measurement and Data Collection. Retrieved from https://www.povertyactionlab.org/research-resources/measurement-and-data-collection

J-PAL. (2019b). Teaching Resources. Retrieved from https://www.povertyactionlab.org/research-resources/teaching

Juselius, K., Møller, N. F., & Tarp, F. (2014). The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis. *Oxford Bulletin of Economics and Statistics, 76*(2), 153-184.

Kendall, J. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal, 20*(2), 164-168.

Leeuw, F. L., & Vaessen, J. (2009). *Impact evaluations and development: NONIE guidance on impact evaluation*: Network of networks on impact evaluation.

Lowder, S. K., Skoet, J., & Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development, 87*, 16-29.

Mellor, J. W. (2017). *Agricultural development and economic transformation: promoting growth with poverty reduction*: Springer.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Van der Laan, M. (2014). Social science. Promoting transparency in social science research. *Science (New York, N.Y.), 343*(6166), 30-31. doi:10.1126/science.1245317

Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, 72*(1), 159-217.

Mondiale, B. (2008). World development report 2008: Agriculture for development.

Morra Imas, L. G., & Rist, R. (2009). *The road to results: Designing and conducting effective development evaluations*: The World Bank.

Mosley, P. (1986). Aid-effectiveness: The Micro-Macro Paradox. *Ids Bulletin, 17*(2), 22-27.

Munthali, K., & Murayama, Y. (2013). Interdependences between smallholder farming and environmental management in rural Malawi: A case of agriculture-Induced environmental degradation in Malingunde Extension Planning Area (EPA). *Land, 2*(2), 158-175.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (2014). *The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research*. (0002-7979). The Journal of the American College of Dentists

Neidell, M., & Waldfogel, J. (2010). Cognitive and noncognitive peer effects in early education. *The Review of Economics and Statistics, 92*(3), 562-576.

Norad. (2019). Norwegian development aid in 2018. In. Norad.

OECD. (2012). *Aid Effectiveness 2011*.

OECD DAC. (2002). OECD Glossary of Key Terms in Evaluation and Results-Based Management. In: OECD Development Assistance Committee Paris.

OECD Data. (2018). Net ODA. Retrieved from https://data.oecd.org/oda/net-oda.htm

Olinto, P., Beegle, K., Sobrado, C., & Uematsu, H. (2013). The state of the poor: Where are the poor, where is extreme poverty harder to end, and what is the current profile of the world's poor. *Economic Premise, 125*(2), 1-8.

Pingali, P. L. (2012). Green revolution: impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences, 109*(31), 12302-12308.

Puffer, S., Torgerson, D. J., & Watson, J. (2005). Cluster randomized controlled trials. *Journal of evaluation in clinical practice, 11*(5), 479-483.

Ravallion, M., & Datt, G. (1996). How important to India's poor is the sectoral composition of economic growth? *The World Bank Economic Review, 10*(1), 1-25.

Ravallion, M., & Datt, G. (2002). Why has economic growth been more pro-poor in some states of India than others? *Journal of development economics, 68*(2), 381-400.

Ricker-Gilbert, J., Jayne, T., & Shively, G. (2013). Addressing the "wicked problem" of input subsidy programs in Africa. *Applied Economic Perspectives and Policy, 35*(2), 322-340.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker : an account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago*. Cambridge, Mass: Harvard University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Roser, M., & Ortiz-Ospina, E. (2017). Global Extreme Poverty. Retrieved from https://ourworldindata.org/extreme-poverty

Sainani, K. L. (2010). Making sense of intention-to-treat. In: Elsevier.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). *When will we ever learn?: Improving lives through impact evaluation*: Center for Global Development.

Schuh, G. E. (2000). The household:: the neglected link in research and programs for poverty alleviation. *Food Policy, 25*(3), 233-241.

Stevens, A. (2011). Telling policy stories: an ethnographic study of the use of evidence in policy-making in the UK. *Journal of Social Policy, 40*(2), 237-255.

Stock, J. H., & Watson, M. W. (2014). *Introduction to Econometrics, Update, Global Edtion* (Vol. 13080). NOIDA: NOIDA: Pearson Education Limited.

Svensson, J., & Yanagizawa-Drott, D. (2012). *Estimating Impact in Partial vs. General Equilibrium: A Cautionary Tale from a Natural Experiment in Uganda*. Retrieved from

Team, M. C., & Region, A. (2018). Systematic Country Diagnostic: Breaking the Cycle of Low Growth and Slow Poverty Reduction.

Thirtle, C., Lin, L., & Piesse, J. (2003). The impact of research-led agricultural productivity growth on poverty reduction in Africa, Asia and Latin America. *World Development, 31*(12), 1959-1975.

UNDP. (2018). Table 5: Gender Inequality Index. *Human Development Reports*. Retrieved from http://hdr.undp.org/en/composite/GII

UNESCO. (2019). Malawi. Retrieved from http://uis.unesco.org/en/country/MW

United Nations. (2019). The Sustainable Development Agenda. Retrieved from https://www.un.org/sustainabledevelopment/development-agenda/

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., . . . Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American journal of public health, 98*(8), 1359-1366. doi:10.2105/AJPH.2007.124446

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness, 5*(1), 30-49. doi:10.1080/19439342.2013.764652

White, H., & Raitzer, D. A. (2017). *Impact evaluation of development interventions: A practical guide*: Asian Development Bank.

Wiig, H., & Holm-Hansen, J. (2014). *Unintended Effects in Evaluations of Norwegian Aid*. Retrieved from Norwegian Institute of Urban and Regional Research (NIBR): https://evalueringsportalen.no/evaluering/unintended-effects-in-evaluations-of-norwegian-aid-a-desk-study/Unintended%20Effects%20in%20Evaluations%20of%20Norwegian%20Aid.pdf/@@inline

Wik, M., Pingali, P., & Brocai, S. (2008). Global agricultural performance: past trends and future prospects.

Wooldridge, J. M. (2002). Introductory econometrics, a modern approach, 2003. *New York: South-Western College Publishing*.

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science, 58*(3), 451-488.

Yalouris, A. (2014). Cost-Effectiveness Analysis. Retrieved from https://www.povertyactionlab.org/sites/default/files/6.%20Cost-Effectivness%20Analysis%202014.03.04%20MININFRA.pdf

Zall Kusek, J., & Rist, R. (2004). *Ten steps to a results-based monitoring and evaluation system: a handbook for development practitioners*: The World Bank.