Practical Guidelines for conducting research

Summarising good research practice in line with the DCED Standard

February 2013 (links updated August 2021)

By Mohammad Muaz Jalil

for the Donor Committee for Enterprise Development

www.Enterprise-Development.org

DCED

The Donor Committee for Enterprise Development

Practical Guidelines for conducting research

Contents

1. Introduction to the study 4
   1.1. Background 4
   1.2. Structure of the report 4
   1.3. Rightsizing expectation 5
2. Research Design 5
   2.1. Research as Part of a Results Measurement System 5
   2.2. Design & Methods 6
   2.3. Much ado about causality 7
   2.4. Types of Research Design 8
      2.4.1. Experimental 8
      2.4.2. Quasi Experimental 10
      2.4.3. Non Experimental Design 11
3. Research Methods 12
   3.1. The spectrum of Qualitative and Quantitative method 12
   3.2. Understanding mixed method 13
4. Data collection tools 15
   4.1. Types of data collection tools 15
   4.2. Surveys 16
5. Characteristics of good measurement 18
   5.1. Reliability 18
      5.1.1. Reliability in quantitative methods 19
      5.1.2. Reliability in qualitative methods 19
   5.2. Validity 19
      5.2.1. Types of validity 20
      5.2.2. Threats to validity 21
   5.3. Degrees of Evidence 22
6. Sampling Strategy 23
7. Conclusion 23

Annex I: Outlier Analysis 24

Annex II: Writing a terms of reference for external research 25
Annex III: Case studies 28
Case Study 1 28
Impact assessment of promoting the use of appropriate soil nutrients by palm oil producing farmers in Thailand with T-G PEC 28
Case Study 2 30
Impact assessment of Minipack seed intervention with Katalyst 30
Case Study 3 33
Impact assessment of EACFFPC Training Course on Freight Forwarder Performance in Rwanda with TMEA 33
Resources 37
References 40

1. Introduction to the study

1.1.Background

This report offers practical guidelines for conducting research in line with the DCED Standard for Measuring Results in Private Sector Development (PSD).

The DCED Standard is a practical eight point framework for results measurement. It enables projects to monitor their progress towards their objectives and better measure, manage, and demonstrate results. As more programmes begin to implement the Standard, a growing need has emerged for guidance on how to conduct research in accordance with good practices, presented in an accessible and condensed form for the ready use of practitioners.

For more information on the DCED Standard, visit the website through this link. Newcomers to the Standard may wish to start by reading an introduction to the Standard, while more experienced users can consult the implementation guidelines.

About the author

Mohammad Muaz Jalil is the Director of Monitoring and Result Measurement Group in Katalyst, a multi-donor funded M4P project operating in Bangladesh. He has a post graduate degree in Economics from the University of British Columbia. He has published numerous articles in peer reviewed journals, presented papers in international conferences and has over 5 years of experience in the field of International Development. He has received training on randomized control trial from J-PAL, Massachusetts Institute of Technology (MIT), USA. He was recently invited as a presenter on M&E at the introductory course on M4P organized by DfID for its PSD advisors in London (2012).

Email address : muaz.jalil@kings.cantab.net

1.2.Structure of the report

This report follows the major steps in a research process. It starts by describing the difference between research design and method. Then it looks in to major types of research designs, touching on various experimental and non-experimental designs. In the section on research methods, a particular emphasis is given to mixed research method because of its strong efficacy in M&E systems within PSD programmes.

The report also discusses tools for data collection, from survey to focus group discussions. Since existing literature is quite strong in these areas, this report provides summaries and references to

the relevant literature. Surveys are one of the most important tools for results measurement, and so receive particular attention.

Strong emphasis is placed on two characteristics of good measurement; reliability and validity. This is often overlooked, but a crucial aspect of research designs. Various threats to external validity and internal validity are also discussed, and examples given. The next two sections deal with sampling and data analysis. The annex of the report contains a step by step guide to removing outliers from data, along with advice for writing terms of reference for external research. There are also three case studies of research conducted by existing programmes.

1.3.Rightsizing expectation

The report is a guideline, and not a step by step toolkit for conducting research. Given the diversity of PSD programmes it is impossible to develop a single toolkit to suit everybody. However the report will describe good practice in social research, with specific examples and tips to assist practitioners. The report is by no means exhaustive, but readers are directed towards existing literature for greater details on specific topics.

2. Research Design

2.1.Research as Part of a Results Measurement System

The DCED Standard identifies eight elements for a successful results based measurement system. It starts by requiring programmes to clarify what exactly they are doing, and what outcomes are expected. This is represented in a 'results chain'. The programme should then set indicators to measure each key change expected, and the measure them on a regular basis. A strategy should be developed to measure attribution, systemic change, and programme costs. Results should be reported on a regular basis, and finally the programme should manage its own results system, ensuring that information generated feeds into management decision making.

This guide will focus on the crucial third step; measuring changes in indicators. Programmes typically spend a lot of time and money on this step. **However, research is only useful as part of a broader results measurement system**. High quality research will not show impact by itself. It needs to be supported by a well-developed results chain, relevant indicators, and a solid attribution strategy. Information from the research then needs to be reported clearly and used to inform programme management and decision making.

Consequently, the starting point of your research should be to ensure that you have a good results management system, including a clear results chain. This will show exactly what type of changes are expected, and so help frame the research question. There should also be indicators that measure the changes shown in the results chain. The research will normally be designed to measure these indicators directly. Without a solid results chain and indicators, your research may not show you anything relevant to the project.

For more on implementing the DCED Standard, visit the overall guidelines here, or go straight to these guides to each specific element:

1) Articulating the Results Chain
2) Defining Indicators of Change
3) Measuring Changes in Indicators
4) Estimating Attributable Changes (now part of 3)
5) Capturing Wider Change in the System or Market (now 4)
6) Tracking Programme Costs (now 5)
7) Reporting Results (now 6)
8) Managing the System for Results Measurement (now 7)

The first step in the process, provided one has identified key indicators and has a result chain, is to develop the overall research design. Unfortunately many texts confuse research designs with research methods, the mode of collecting data. The following section briefly delineates the two concepts.

2.2.Design & Methods

The terms 'research design' and 'research methods' are often used interchangeably; however they are distinct concepts. 'Research design' refers to the logical structure of the inquiry. It articulates what data is required, from whom, and how it is going to answer the research question. Fundamentally research design affects the extent to which causal claims can be made about the impact of the intervention. Research design thus `deals with a logical problem and not a logistical problem' (Yin, 2009, p. 27). For instance a programme might choose to do quasi experimental design to estimate the attributable impact of an intervention. How to do the research or what info to collect becomes a choice of methods.

Research methods, by contrast, specify the mode of data collection. This includes whether qualitative or quantitative data is required, or a mix of the two. In theory at least there is nothing intrinsic about any research design that requires a particular research method, though in practice more experimental designs tend to use quantitative methods.

These guidelines first explore research designs, explaining how different designs can address the issue of causality in the intervention. It then examines research methods, including data collect techniques, surveys, and sampling.

Before we look in to various research designs, we will first digress a bit and try to clarify the term causality because in the heart of result measurement is the concept of causality. In a result chain the boxes are connected via a causal link and it is very important to understand just what we mean by the term 'causality.'

2.3.Much ado about causality

Causality is a fundamental part of result measurement, as we want to see what impact a particular intervention has on the target population. In other words, is there a causal link between the activity that we undertake, and the result we see? This is the link captured in the results chain, which tries to build a causal chain between the activity and the outcomes or impact.

We might seek to demonstrate this causal link by measuring the variable that we wish to affect before the intervention, and then comparing it to afterwards. For example, in an intervention to reduce poverty, a researcher could measure poverty levels in the target population before the intervention, and then compare it to afterwards. If poverty decreases, we might think that our intervention was successful.

However, a decrease in poverty levels may not have been caused by your intervention. Poverty is affected by many things; global economic forces, local businesses, other private and public programmes– even the weather. A real challenge for programmes is to show that observed improvements were due to their work, rather than other factors. This is often known as the 'problem of attribution'; can you attribute observed improvement to your activities? In the diagram to the right, the change with the intervention is shown by the top, black sloped line. The change without the intervention is shown by the dotted sloped line. The impact of the intervention is the difference between these two lines, which is the change attributable to the intervention.

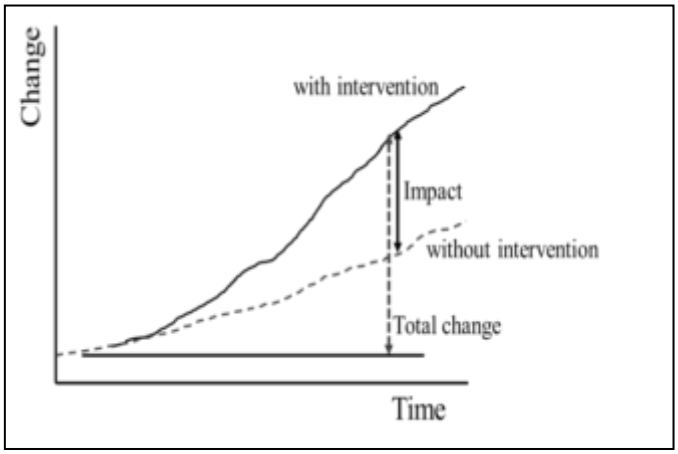


Figure 1: Causality and attribution

In order to demonstrate that the programme caused the observed benefits, it is often necessary to construct a 'counterfactual'. This shows what the world would have looked like if your programme had not been there. The difference between this 'counterfactual' situation and the real one shows how much difference the programme made, and thus how much of the improvement was attributable to the programme. *The objective of different research designs is almost always to artificially create this alternate version of reality, to check what would have been the case if the intervention was not there;* it is to come up with that 'dotted' line in Figure 1. In the following section we discuss the various types of research designs available[1].

2.4.Types of Research Design

There is lack of consistency in classification of different types of research designs. Some classify based on the type of research question being addressed (exploratory, descriptive etc.), others focus on the data collection tools (survey, quantitative, qualitative); Stern et al (2012) classified using the basis for causal inference to categorise different design methods. In this report we follow the structure of Imas & Rist, (2009) while drawing on existing body of literature to ensure there is broad coverage of different designs.

Broadly speaking we can classify research designs in to experimental, quasi experiments and non-experimental designs. These are discussed in the following sub-sections (a useful list of available literature on these various designs are given in the resource section

2.4.1. Experimental

In an experimental design individuals selected from the population of interest are randomly assigned to two groups, one of which is subject to the intervention (referred to as the 'treatment' group) and the other not (referred to as the 'control group'). Generally this assignment is done before the intervention is launched. The experimental design assumes that, since the two groups are drawn from the same population and randomly assigned, they are similar in every aspect except that one group received treatment. Thus, if there is any difference between them, it must be due to the intervention. This difference is known as the **treatment effect.**

Experimental design is the best way to ensure that the treatment group and control group are really comparable. *Without random assignment, individuals receiving the treatment may be systematically different from those not receiving it. This is called* ***selection bias****.* For instance assume that a vocational training program compared career outcomes between students who have been trained

[1] White and Phillips (2012) examines in detail various evaluation approaches that are suitable for small sample.

and those who haven't been trained. In this case, the students who have been trained are the 'treatment group', while those who haven't been trained are the 'control group'. They may not be comparable as student who enrolled themselves to the program might be better to begin with, or more motivated. Alternatively, the institution may choose students who are most likely succeed. Experimental design ensures that groups are similar in both observable and unobservable parameters, because they have been randomly assigned to the control or treatment group.

Experimental designs are often called Randomized control trials (RCTs). There is an extensive literature on how to conduct such experiments and their usefulness. For example, Duflo, Glennerster, and Kremer (2008) produced a succinct introduction to RCTs[2]. J-PAL MIT offers a 5 day course on RCTs titled "Evaluating social programs", currently available in several locations around the world[3]. While some claim randomized control trial to be the 'gold standard' for impact assessment[4], others suggest that evidence from randomized control trials has no special priority[5]. *Unfortunately there is virtually no literature on the viability of using RCTs approaches in M4P or PSD projects; most RCTs based studies are focused on health and education sectors.*

In PSD interventions, where private sector buy-in is the major focus, random assignment before the intervention might be nearly impossible. For instance a PSD intervention might train some of the exclusive retailers of an agricultural input company. The programme aims for the company to buy in to the model and expand it to include all it retailers within its network. In an experimental design, the retailers would be randomly chosen to receive the training so that the outcome is unambiguously attributable to the training. However the fundamental problem with this is that the intervention was designed to elicit buy in from the parent private sector company so that they would expand on their own. In this case, it may be more effective to allow the company to select its trainee retailers based on its own need and priorities, as this would result in greater ownership. However, if they are viable, RCTs are a valuable technique for demonstrating the impact of a project. They can be used as a convincing and influencing tool and a positive RCT can convince the stakeholder to expand the model. RCT studies also come with a very high price tag and require a considerable amount of time from managers. Therefore it is up to the project managers and donors to decide whether RCTs are a worthwhile use of resources.

---

[2] Please look in to the resource section of this report on Randomized control trial for details on the paper

[3] For detail on the course please visit https://www.povertyactionlab.org/j-pal-courses .

[44] Duflo, 2004; Banerjee, 2007; Duflo, Glennerster & Kremer, 2008

[5] Concato et al, 2000; Deaton, 2010; Cartwright, 2007

2.4.2. Quasi Experimental

This design is similar to experimental design, but it does not randomly assign individuals to groups. Instead, the researcher develops a comparison group which is similar to the treatment group but not necessarily equivalent. It may be easier to implement than an experimental design; although the technical expertise required to develop a quasi-experimental design may be just as high as that of developing a RCT. There are various types of quasi experimental designs, and we discuss the most prominent ones below[6]. The objective will be to explain the logic behind these methodologies rather than detailing the mathematics.

Method	Basis for Inference (Attribution strategy)	Examples	Caveats
Propensity score matching (PSM)	The design attempts to select a control group that is as similar as possible to the treatment group, based on observable characteristics. Although it is not randomized, the design attempts to find a control group that is comparable and that does not suffer from selection bias.	To evaluate the success of SME training, you would identify SMEs which have all the relevant pre-treatment features (e.g. scale of operation, length of operation, geographic location etc.) of the treatment SME, but did not receive training. They would act as control. The objective is to then compare the post treatment result between the control and treatment group to evaluate the treatment effect	1. Treatment and control may differ in unobservable factors. For example, in the example to the left, SMEs that received training may be more motivated or experienced. It will be very difficult to take these factors into account when selecting the control group.
Regression discontinuity (RD)	A cut-off or threshold is assigned above or below which an intervention is implemented; by comparing observations lying closely on either side of the threshold, the effect of the intervention is estimated	If all students above a given grade - for example 60% - are given vocational training, then it is possible to estimate the treatment effect by comparing students who received slightly above 60%, to those who received slightly below 60%. This is because natural variation in test scores means that difference between these students is likely to be due to chance, rather than ability, and they are likely to be similar in all other respects. This can be also used to estimate the impact of micro finance or other such schemes if the disbursement uses a sharp measurable cut off point like some index values (e.g. progress out of poverty index)	1. It measures only the local treatment effect i.e. the estimate is only valid for cases around the threshold. 2. There is the possibility of contamination where some people below the threshold end up receiving treatment thus biasing the results.
Instrumental variables (IV)	IV methods use an exogenous instrument(s) which correlates well with the covariates of interest but is uncorrelated with the error term of the regression.	If one wishes to measure the impact of increased visits to retailers on farm productivity, simply running a regression with farm output as dependent variable and number of visits as covariate may bias the result even if it is adjusted for other covariates. There might be general increase in productivity or improvement in road system that can bias the result. Here one can chose distance of farmer from retail store as an instrument as it is unlikely that this can affect productivity through any other mean, other than through number of visits to the retailer outlet.	1. Identifying the right instrument is more art than science 2. The effect it measures is not the average treatment effect but the local average treatment effect i.e. the treatment effect for the subpopulations affected by the observed changes in the instruments

[6] Imbens and Wooldridge (2009) provide a more intensive intro to experimental and quasi experimental designs

Difference in difference (DID)	This methodology first takes the difference between the pre-test and post-test outcome value of a treatment group, and then takes the same from a suitable control group. Rather than comparing the absolute values, between the two groups, it compares the value of the difference between the pre and post test. Thus the control group does not have to be similar to the treatment group, but must have a stable and predictable growth rate.	An intervention might be designed to promote, via a private seed producer, usage of quality seeds among farmers. In such cases where the private sector decides where and how to promote, it might be difficult to conduct an experimental design. However it might be possible to look in to villagers, may be of some nearby villages who were not affected by the program. The idea is to first check if the production/ yield growth between the control and the treatment group before the intervention was similar and stable. Then extrapolate what would have been the case for treatment farmer if they had not participated in the program, based on the production of the control farmers after the intervention.	1. The estimation will be biased if there are other factors that affect the difference in trends between the two groups. 2. Participation in the program may be based in difference in outcomes prior to the intervention. 3. Pre intervention stability in growth relationship between the control and treatment group may be a short term phenomenon.

Since difference in difference methodology is one of the most widely used quasi experimental designs in applied research, we now illustrate the framework in greater detail. Let us assume there are two groups, one treatment and one control and we wish to see whether an intervention has increased the yield of the treatment farmer:

Step 1: Estimate the yield of the treatment group and control group before the intervention. This is the baseline.

Step 2: Estimate the yield of the treatment group and control group after the intervention. This is the end line

Step 3: Estimate the change in yield for both the treatment and control group. This will yield two values. Firstly, you will know the difference between the baseline and end line yield for the control group. Secondly, you will know the same for the treatment group.

Step 4: Compare the change in yield for the control group, with the change in yield for the treatment group. This is the 'difference in difference'. Even if the control and treatment group were initially different, this can still estimate the impact of the intervention.

2.4.3. Non Experimental Design

A non-experimental design does not compare one group with another but describes the relationship between an intervention (treatment) and its effects on the population of interest. Furthermore it may provide a rich understanding of the contexts, process, event, or situation and explain why results occurred, which may be essential for building result chains. Example of such design includes case studies, longitudinal studies, ethnographic studies etc.

A case study is an intensive analysis of a single unit, whether a farmer, business, or village. Case studies are frequently qualitative, but can also use very quantitative measures, collecting numerical data about the subject of interest. Case studies can be used to showcase program extremes or a typical intervention, or randomly selected to give a broader picture. A scenario where case studies can be used is monitoring for early sign of impact in an intervention, such as the behavioural change of intermediary agents (e.g. service provider). Another example where case studies could be useful is for capturing employment changes in farms. Before we launch a large scale survey to ask farmers about labour usage, we may need in-depth case studies to identify accurately the different steps in the production process (pre harvest, irrigation etc) and associated labour usage. This gives detailed information to base the survey on.

In longitudinal studies individuals or groups of interest are monitored over a period of time and are interviewed at different stages. Sometimes the major objective is to better understand the dynamics of the treatment effect, which can assist in the development of more experimental or quasi experimental design. For example, if one is interested in estimating the consumption pattern of the target group then longitudinal studies can be used to identify major spending channels, and then the research might be supplanted by larger focused survey on those channels only.

In the next section we will look in to various research methods or data collection methods that are available, with particular emphasis on mixed methods.

3. Research Methods

3.1. The spectrum of Qualitative and Quantitative method

Quantitative data collection methods consist of counts or frequencies, rates or percentages, or other numerical data. They often come from surveys, structured interviews, observation checklists, or archival records, such as from government databases. Qualitative data describes problems, behaviours, opinions, experience, attitudes, and beliefs. They are non numerical in nature. Qualitative data are non-numerical in nature, and can come from key informant interviews, focus group discussions, open ended questionnaires, field notes, or personal log or journals. The following table gives a brief overview of the advantages and disadvantages of using qualitative and quantitative research methods.

	Quantitative	Qualitative
Advantage	• Relatively easy to administer, • Can include large number of questions, • Can yield large samples, • Emphasizes reliability	• Captures more depth and provide insights as to the "why" and "how" • Emphasize validity, • Easier to develop
Disadvantage	• Data may not be as rich or as detailed as qualitative methods, • Usually are harder to develop, • May not provide sufficient information for interpretation	• Time consuming to capture and analyze, • More subjective and may be difficult to summarize and compare systematically, • Difficult to have large sample, • Very demanding to administer

As a rule of thumb quantitative methods are used when the researchers wants to conduct statistical analysis, cover large samples, or seek precision while qualitative methods are used when in-depth information is the key, sample size is not an issue. But this sharp distinction between the two methods has begun to erode. For instance qualitative data can now be coded, quantified and even econometrically analyzed using powerful statistical software like NVivo or Atlas.ti. There is a widespread agreement on the benefits of combining various quantitative and qualitative methods; this is known as mixed method research and the following section discusses it in detail.

3.2.Understanding mixed method

Mixed method is a "*type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches for the broad purposes of breadth and depth of understanding and corroboration."*[7] The method combines multiple techniques to enrich the research findings. Triangulation is a term which often associated with mixed methods, however they are not synonymous. There are multiple types of mixed methods and triangulation is but one of them albeit an important one. The following table discusses different types of mixed methods[8].

Types	Definition	Features
Triangulation	Triangulation refers to using different methods to study the same phenomenon. This improves the reliability of results, as unreliability or bias in one method can be balanced out by the use of another.. According to Denzin (1978), three outcomes arise from triangulation: convergence (the results lead to similar conclusions), inconsistency (there are some discrepancies between results), and contradiction (the results don't match at all). Given any of these outcomes, the researcher can construct better explanations of the observed phenomena or expand the research by using different methods to reach further conclusions.	1. Gives equal priority to both quantitative and qualitative data. 2. Collects both the quantitative and qualitative data simultaneously 3. Compares the results from quantitative and qualitative analysis to determine if the two databases yielded similar or dissimilar results. 4. There can be data triangulation (use of multiple source), investigator triangulation (using of multiple researchers), theory triangulation (testing of multiple theories to interpret results)

[7] Johnson, Onwuegbuzie and Turner (2007)
[8] Greene et al, 1989

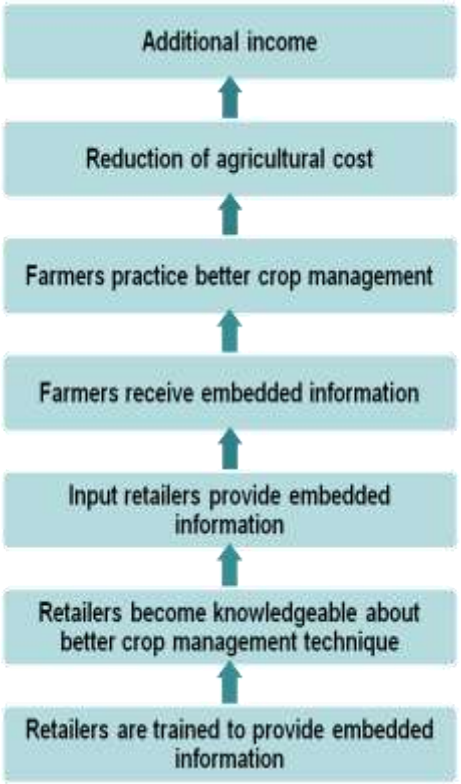
Complementarity (Explanatory)	Seeking elaboration, enhancement, illustration, clarification of the results from one method with results from the other method	1. Priority on quantitative data collection and analysis. 2. Collect quantitative data first in the sequence. 3. Qualitative data is used to refine the results from the quantitative data.
Development (Exploratory)	Using the results from one method to help inform the other method	1. Emphasizes the qualitative data more than the quantitative data. 2. Sequence to data collection, first collecting qualitative data followed by quantitative data. 3. Quantitative data used to build on or explain the initial qualitative findings.

Of the above methods, triangulation is particularly crucial, and should play a part in almost every results measurement system. This is because there is no single perfect method for assessing results in private sector development programmes. Almost every method can be subject to bias, unreliability, or invalidity. By using multiple methods and triangulating the results, it is possible to arrive at a more robust conclusion. The example below gives an example of using mixed methods in practice.

Example: Using mixed method research in Retailer Training program

A program trains input retailers to offer information on better cultivation techniques to farmers (figure 3). Farmers who receive this information from trained retailers are potentially 'direct' beneficiaries of the program. If other farmers receive information from these direct farmers then they are 'copy' or 'indirect' beneficiaries. Often problem arises in trying to estimate the number of reached indirectly. . Here mixed method research can be used. The following suggestions (not exhaustive) are provided:

1. Conduct survey on direct farmers asking them about the number of farmers influenced by them and who adapted the new practice of cultivation. Ask them for a list of five such farmers.
2. Interview a sub sample of this list of copy farmers taken from the survey of direct farmers, to ask them if they have adopted the new practices, and how many other farmers did. This triangulates the results of the original survey, as you are now collecting the same information from multiple sources.
3. Conduct focus group discussions with direct farmers in one region to explore if there are overlaps between indirect farmers (e.g. two separate direct farmers might give the name of the same indirect farmer saying that they individually have influenced him). This can be done before (development) or after (explanatory) the survey. If the former, then results from the focus group discussions can inform the questions asked in the survey. If the latter, then they can deepen and explain the results found in the survey.
4. If the intervention focused in an isolated region then one can ask local extension officer or input sellers to estimate number of farmers who have changed practice after the intervention, as it may be observable and uncontaminated by external factors.



4. Data collection tools

4.1.Types of data collection tools

Data collection tools include focus group discussions, surveys, key informant interviews etc. In the present report we will give a brief overview of the prominent ones along with their advantages and disadvantages. The following table does this for all the major tools, except survey which is dealt with later in greater detail. References to more detailed sources of information on these tools can be found at the end of this document.

Tools	Short Description	Advantages	Disadvantages
Observation	"Monitoring through observation" simply means visiting workshops, events or projects, and watching what happens. Direct observation is undertaken in person while indirect observation takes place when using appropriate technology such as video recording.	§ Collect data where and when an activity occurs § Does not rely on people's willingness to provide information § Directly see what people do rather than relying on what they say they do	§ Susceptible to observer bias § Hawthorne effect – people usually perform better when they know they are being observed § Does not increase understanding of why people behave the way they do
Secondary research	To use pre-existing sources (e.g., documents, data files, log sheet or other written piece) with the intention of collecting independently verifiable data and information. It is usually of three types: 1. Content analysis - focuses on various forms of human communication, like news papers, articles etc 2. Meta analysis- combines and investigates the output of other research concerned with the same or a similar phenomenon 3. Secondary analysis - using quantitative data that were previously collected by other people for a different purpose.	§ Relatively inexpensive § Good source of background information § Unobtrusive § Provides a "behind the scenes" look at a program that may not be directly observable § May bring up issues not noted by other means	§ Information may be inapplicable, disorganized, unavailable or out of date. § Could be biased because of selective survival of information § Information may be incomplete or inaccurate § Can be time consuming to collect, review, and analyze many documents § May not be available for the population of interest to you. § Open-ended or qualitative data usually not available.
Interviews	A purposeful exchange between two people to uncover perspectives, experiences, and insights on a phenomenon. It is useful for collecting in-depth and detailed qualitative data. Data can be	§ Useful for gaining insight and context into a topic § Personal contact with participants might elicit richer and more detailed responses.	§ Susceptible to interview bias § May requires special equipment to record and transcribe interviews. § May seem intrusive to the respondent

	analyzed using content analysis with narrations and quotations. During interviews the interviewee should not lead the respondent(s) with biased, assumption loaded questions.	§ Allows respondents to describe what is important to them	§ Time consuming and expensive compared to other methods
Focus Groups	A specially selected group is interviewed by a moderator. The group is usually composed of six to twelve individuals. Focus groups are useful for exploring norms, beliefs, attitudes, practices and languages. Focus groups require trained moderators.	§ Quick and fairly easy to set up § Group dynamics can provide useful information that individual data collection does not provide § Useful in gaining insight into a topic that may be more difficult to gather information through other methods § Can be effectively used to focus on details regarding issues found through surveys or other data collection methods. § Participants are not required to read or write.	§ Susceptible to facilitator bias § Discussion can be dominated or sidetracked by a few individuals § Data analysis is time consuming and needs to be well planned in advance § Does not provide valid information at the individual level § The information is not representative of other groups

Source: Adapted by author from ACT Evaluation Toolbox

4.2.Surveys

Survey is one of the most important modes of data collection, and one of the most widely used. A survey collects data from a large number of people, using a standardized set of questions. The primary benefit of a survey is that it can give quantified, reliable data on a wide range of issues. This can include opinions and needs of stakeholders, the socio-economic situations of various groups, changes in income, perception, and more. Surveys can gather information that informs the design of programmes, or to evaluate their impact. They also provide data that can be statistically analyzed, and used to make inferences about a larger group of people. However, surveys are often resource intensive. Moreover, as a primarily quantitative tool, they may give a simplistic picture of the situation. They are useful for answering quantitative questions, such as establishing changes in income,

Tips for conducting good survey:

- Keep it simple, clear, easy, and short.
- Find and review similar survey conducted by others
- Do not ask respondents for information that requires them to go to a file or other source
- Conducting follow ups minimizes non-response
- Make sure the questions are well worded
- Avoid double-barreled or double negative questions.
- Use multiple items to measure abstract constructs.
- Do not use "leading" or "loaded" questions.
- PILOT TESTING is a must, it not only improves quality but reduces cost
- If survey is conducted by external enumerators then ensure 1) they are properly briefed and trained. 2) conduct mock interview session with them

employment, or growth rates of firms. They are less useful for uncovering perceptions of the project, sensitive issues, or unanticipated benefits.

Survey can be classified in to 1) cross-sectional surveys where data are collected at one point in time from a sample selected to represent a larger population and 2) longitudinal survey, which involves repeated observations over a period of time. Longitudinal surveys can be further classified in to Trend, Cohort, and Panel survey.

- Trend: Surveys of sample population, i.e. a sub sample of the target population, at different points in time with different respondents in each survey, so the sample differs but the population from which the sample is drawn remains the same.
- Cohort: A cohort is a group that experiences an event (typically birth) in a selected time period. Cohort analysis studies the same cohort each time data are collected, although samples studied may be different. For instance a healthcare program on postnatal caring may use cohort survey to monitor those infants who were born after the program in the intervention region.
- Panel: Collection of data at various time points with the same sample of respondents. Hence the sample is drawn from the population once and is continued throughout the study.

Advantages	Disadvantages
§ Many people are familiar with surveys	§ Items may not have the same meaning to all respondents
§ Reduces chance of evaluator bias because the same questions are asked of all respondents	§ Size and diversity of sample will be limited by people's ability to read
§ Some people feel more comfortable responding to a survey than participating in an interview	§ Given lack of contact with respondent, never know who really completed the survey
§ Can include both close-ended and open-ended questions.	§ Unable to probe for additional details
§ Tabulation of closed-ended responses is an easy and straightforward process	§ Good survey questions are difficult to write and usually take time to develop and hone

Surveys can be conducted face-to-face, over the phone, through mail or online. While online survey is rapidly becoming popular, in developing countries face to face surveys remain the most common way of conducting surveys, followed by telephone surveys. Face to face surveys have two major difficulties. Firstly, there is a substantial cost involved. Secondly, selective hearing on the part of the interviewer may miss information that does not conform to pre-existing beliefs. One innovative use

Example: Multistage survey

A program intervention deals with the establishment of community agricultural information centre, where a service provider sells agricultural information to farmers from an online knowledge bank. Usually it becomes very difficult to trace what kind of information these farmers are using. Multiple farmers may come together and ask for a specific information bundle, in which case website hit rate will underestimate number of farmers accessing the service. Further complication arises if we ask the service provider for the names of accessing farmers, as significant personal bias might be present.

In order to avoid these problems a user tracking survey might be conducted where enumerators are stationed in a sample of such service outlets and note the name and address of accessing farmers. This user tracking survey can be conducted during the season or in a sample of days/weeks to reduce the variability and reduce cost. The user tracking will give a large sample of names of accessing farmers. Then one can conduct a random sampling of farmers from this large pool of names to conduct the impact assessment using a survey.

of survey tool is the multistage survey methodology, where multiple surveys are undertaken sequentially as part of one study. The following example illustrates such a design.

Annex 2 gives advice on how to conduct survey through third parties and what information should be in the Terms of reference. For more information on surveys, consult these links:

- Europe Aid Evaluation Methodology Guide: This is a clearly written and comprehensive guide to survey methodology.
- Research Methods Knowledge Bank: This is a slightly more detailed and theoretical guide to survey methodology.

In the next section we will examine what makes a measure valid and reliable.

5. Characteristics of good measurement

Social research entails measurement of complex constructs such as poverty, nutrition, sustainability, systemic changes etc. As such there is always scope for inaccuracy between what we measure and the true nature of the construct. For example, assume that a project wanted to measure poverty in the local area. They might measure income, the quality of buildings, the number of assets owned, or adopt a participatory approach where a village community decides who in the community is 'poor'. Each of these methods has advantages and disadvantages, and this section explores some of the characteristics of a 'good' measure, whether of poverty or another complex construct. The pertinent questions to ask is "*how do we know that we are indeed measuring what we want to measure?"* and *"can we be sure that if the measurement is repeated we will get the same result?" T*he first question is related to validity and second to reliability. It is important to bear in mind that validity and reliability are not an all or none issue but a matter of degree.

5.1.Reliability

Reliability is the degree to which measures are free from error and therefore can yield consistent results. (Thanasegaran,2009). Reliability is the degree to which a test consistently measures whatever it measures (Gay, 1987).In other words, a reliable test would give the same result if used several times over. The following subsections offer some guidelines as to how to increase reliability in research.

5.1.1. Reliability in quantitative methods

Reliability measures can be built in to research design to improve quality. Some of the most common methods are:

- Test-retest Reliability: Test-retest reliability is the degree to which scores are consistent over time. E.g. IQ tests typically show high test-retest reliability. However this is not useful if the respondent remembers the previous answers and repeats them from memory.
- Equivalent-Forms: Two tests that are identical in every way except for the actual items included in the forms. One example is two equivalent arithmetic tests, which will have different questions but test the same basic concept. A reliable measure should produce similar scores. Unfortunately it is often very difficult to develop such equivalent forms.
- Internal consistency: The degree to which all measures in a test or questionnaire relate to all other items. The split-half method is often used to measure internal consistency. This is done by checking one half of the results of a set of scaled items against the other half.

5.1.2. Reliability in qualitative methods

Reliability is often overlooked or misunderstood when it comes to qualitative research. Golafshani (2003) specifically discuss the reliability and validity issues in qualitative research. While it is true that one cannot have a comparable measure of reliability in qualitative research, it is however possible to have similar but alternate dimensions. Lincoln and Guba (1985) identified 1) Credibility, 2) Neutrality or Confirmability, 3) Consistency or Dependability and 4) Applicability or Transferability as the essential criteria for quality. The source used in the qualitative research should be neutral, reputable, credible and trustworthy. Therefore Wikipedia may not be a good research source, since the information is editable by anyone and can change. If however one does find information in such sites then one can look for the source that the information itself refers to, usually an article or news piece, and then cite and/or use that. Similarly blog and other social media may not be appropriate information sources due to doubts regarding trustworthiness and neutrality.

5.2.Validity

A reliable measure isn't necessarily useful. For example, a weighing machine which is always off by 2 kg; will give 'reliable' results, as it will give the same (wrong) result every time. It is also important for measures to be *valid*. Validity is the extent to which a construct truly measures what it was set out to measure. The diagram below illustrates the difference between reliability and validity:

We will now examine four validity criteria which are useful for research design in results measurement.

5.2.1. Types of validity

- **Statistical conclusion validity:** Whether the researchers have used the right statistical approach in measuring the causal relationship. The pertinent aspects to look at -
  - Is there a causal relationship between X and Y?
  - Whether the study is sufficiently sensitive to pick up on the correlation?
- **Internal validity:** Even if a change is observed, it is important to understand how confident we can be that the intervention contributed to it. A strong internal validity implies that the conclusion that the intervention did or did not cause the observed results is robust. We should ask whether the study has been conducted so as to rule out other possible causes for the observed effect. The internal validity is related to the choice of research design. However a strong internal validity may sometimes be for a very specific context, and therefore have low external validity (i.e. not valid for other contexts).
- **External validity:** This refers to the extent to which the findings of the study are generally applicable. It addresses the problem of generalization, whether the intervention will work "somewhere else".
- **Construct validity:** The degree to which a test measures an intended hypothetical constructs. It stresses the importance of precise definitions and concepts. For instance, if we develop an index to measure social status, we would expect social status to positively correlate with education and monthly income, and negatively with criminal record. If this is not the case, then it may be that the construct/index for social index was invalid.

5.2.2. Threats to validity

- **Threats to internal validity:** Threats to internal validity are factors which may prevent us from deducing a causal link between the intervention and effects. Different types of research designs aim to address these threats to internal validity. The major threats to internal validity includes
  - **History effect**. External events may affect the outcome of intervention. Consequently, any observed change may be due to the external events, rather than the intervention itself.
  - **Maturation effect**. The group that is being tested is likely to change over time. Two tests over a period of time may produce different results because of the maturation effect, rather than because of the intervention.
  - **Repeated testing effect**. Behaviour of the individuals may change because they are participating in the research. For example, direct beneficiaries may wish to emphasise the benefits that they received from the programme, in order to receive more funding in future.
  - **Selection bias**. Bias due to comparison between self selected individuals in a program and those who chose not to participate. If those who chose not to participate form a control group, they may not be comparable. For example, they may have a higher base level of skill, or a lower level of motivation than those who participated in the programme,
  - **Attrition effect**. Effects of drop outs from the intervention on the outcome measured. For example, a programme might try to compare a pre-test to a post-test following a training session. However, if some of the participants in the training dropped out, they would have participated in the pre-test but not the post-test. Since the dropouts are likely to differ from the rest of the group (they might be less motivated, less able – or even potentially more able) this will affect the validity of the test.
  - **Contamination effect**. Members of control group may benefit from the intervention indirectly. This is especially true in PSD interventions where programmes may aim to catalyse systemic change. In this case, a 'control group' who were not directly affected by the programme may still have indirectly benefitted. Consequently, it is impossible to compare them with a treatment group to see the impact of an intervention.

- **Threats to external validity:** High degree of context specificity implies lack of generalizability. That implies that even if the research design is robust enough to exclude threat to internal validity the result may be valid only for the local area or region where it was carried out. Additionally a result from a specific small scale intervention may not necessarily hold when it is scaled up since new variables may come in to play.

5.3.Degrees of Evidence

The Degrees of Evidence framework[9] for monitoring and evaluation has been developed by the PSD-IAI and is gaining much traction among development practitioners. It is a set of principles that allows one to gauge the rigor of research. Methodological rigor is determined by the extent to which the research adheres to the following principles:

- **Methodological validity** which refers to internal, external, construct and statistical conclusion validity
- **Triangulation** focusing on mixed method research design and emphasizing that the measurement is more credible if it is supported by multiple sources of evidence.
- **Methodological transparency** ensuring research methodology and designs are well documented and therefore traceable.
- **Sound data collection methods** focusing on appropriate use of data collection method in line with good practice.
- **Methodological appropriateness** examines whether the research methodology is appropriate to answer the research question(s).

For each of these aforesaid criteria there exists a continuum from 'sound' to 'unsound.' The extent of rigor is determined by how well the research scores across all of the criteria. In the next section we will examine available sampling strategies.

[9] A two page summary can be found here: https://beamexchange.org/resources/347

6. Sampling Strategy

Sample selection strategy is a crucial part of the research design. Consequently, the DCED has produced specific guidelines on how to select sample sizes, and an accompanying sample size calculator. This guidance provides a simple, practical tool to help programmes using the DCED Standard to select the appropriate sample sizes for quantitative surveys.

7. Conclusion

These guidelines have been prepared to give an overview of different aspects to keep in mind while conducting research in line to established good research practices. However it is very important to keep in mind that the DCED Standard for results measurement calls for programmes to be pragmatic in results measurement. It is important to balance the line between 'what is good enough' and 'what is practical' because if the two don't meet, programmes are stuck with a system that is not used, that is unaffordable or not resource efficient. Hence sometimes programmes might choose to select only a few key interventions where they spend more resources to validate results; while for the rest they try to find more cost efficient alternatives. In the end when it comes to research design and choosing methods, programmes need to find the options that suit them best, instead of trying to fit a one-size solution for all impact assessment. Annex three shows three examples from three different programmes which are working towards integrating the different elements of the Standard in their own results measurement systems. They exemplify how they use different research methods for their impact assessment that fits to their organizational need and resources that are available.

Annex I: Outlier Analysis

Quantitative data often have outliers which need to be filtered before a meaningful interpretation is possible; sometimes the outliers themselves may be of research interest. The outliers may originate out of measurement error, survey error, or simply because the population of interest has outliers No matter what the case it is important that the dataset is adjusted for outliers. There are various methods of outlier analysis; Barnett and Lewis (1984) provide a good overview of the literature.

The outlier algorithm that will be discussed in this section is called "outlier labelling rule" and draws on the paper by Iglewicz and Banerjee (2001). The method is very easy to deploy and does not require prior knowledge of the number of outliers nor in which are they located. That rule declares observations as outliers if they lie outside the following interval: $\mathbf{[\ Q_1 - \beta\ (Q_3\text{-}Q_1)]}$, and $\mathbf{[Q_3 + \beta\ (Q_3\text{-}Q_1)]}$, where Q_1 and Q_3 stand for 25th and 75th percentile or 1st and 3rd quartile value respectively. Literature shows that β value should be between 1.5 and 3 although most research uses either 1.5 or the more conservative estimate of 2.2. In the following we provide a step by step guideline for outlier estimation of a dataset:

Step 1: Order the data in ascending or descending order.

Step 2: Calculate the 25th percentile (1st Quartile). This can be estimate by taking the value of the number which falls at the given percentile. For instance if there are "n" numbers in an ordered set, then the 25th percentile or 1st quartile is the value of the number in the position (n+1)/4. Unfortunately there is no universal agreement on choosing the quartile values and some may suggest using the formula (n+2)/4 if the numbers in the ordered set is even. Fortunately in excel one can use the function "QUARTILE.EXC" to derive the quartile value. For 1st quartile this will be QUARTILE.EXC ("array", 1), where array refers to the column of data in the worksheet.

Step 3: Estimate the 75th percentile (3rd Quartile) values using the same method as above but in this case the formula is (3n+3)/4 or in case of even (3n+2)/4. We can use the QUARTILE.EXC function to estimate the 3rd quartile and in that case the function should be QUARTILE.EXC ("array", 3).

Step 4: Estimate the outlier interval based on the aforesaid formula using the two quartile values.

Step 5: Keep the data that are equal or below the interval values determined in step 4 and filter out the remaining outliers.

Annex II: Writing a terms of reference for external research

This section specifically focuses on what should be taken in to consideration while developing a terms of reference (TOR) for a field research to be outsourced to a third party. The rationale for the study can be multifarious and may refer to a baseline study, sector mapping, impact assessments etc. Thus the section will offer succinct guidelines that can assist organizations to develop such TORs. The resource section of the report provides additional reference to secondary source that discusses how to develop TORs in detail. The following elements should be present in a TOR:

1. **Background to the Study:** This can be at most 2 paragraph (10-15 lines) providing a brief background of the organization and specifically focusing on the motivation of the study. For instance in case of baseline of a value chain sector this can discuss what prior studies were undertaken and why they were inadequate or what were the information gaps etc.
2. **Objective of the study:** This section will primarily focus on the key objective of the study and can be in bullet points. However being specific is the key focus in this section. Thus in case of impact assessments it is not sufficient to mention 'the study will try to assess the increase income of farmers'; it might be better formulated by stating 'the study will assess the increase in income of direct and indirect farmers due to usage of better quality seed'. One important thing to avoid is to overload this section with too many details. An alternative option could be to have a table with a list of specific indicators to be measured or assumption to be validated.
3. **Scope of the study:** This focuses on both on technical and geographic scope of the study. This section sets the boundary or the limits of the study. Hence it may list the geographic coverage of the study area or the scale of operation (fixed number of enumerators or sample size).
4. **Methodology:** This section refers to the data collection instruments to be used like FGDs, Survey etc and their numbers. It should also discuss about the sample size and the methodology of selecting individuals from the population to create the sample. It might happen that the contracting organization has already developed its own sampling plan and geographic coverage in that case this section will provide this in detail. The section should also discuss efforts taken to ensure quality of the research. Following is a list of steps that can be taken to improve quality of such external research :

a. Pre testing: All questionnaires should be pre tested before launching of the major survey. Often translation from English to local language may transform the meaning which may be picked up in such pre testing. Formulation is also something that may change after questionnaires are pre tested.

b. Training of enumerators & supervisor: In a third contract enumerators are usually drawn from a pool and therefore are not directly part of the third party organization, as a result quality is an issue. Although by looking at the CVs and previous experience one can filter and improve the quality of the enumerators, training still remains essential. The training should ideally focus on four things 1) Background to the study 2) Rapport building with interviewee 3) Explanation of the questionnaire 4) Mock in-house interview sessions.

c. Random checking of completed questionnaires: Once hardcopies of the questionnaire are received it is prudent to randomly check, either through additional field visits or phone calls, 2-5% of the questionnaires.

d. Double entry for data input: This implies two individuals will simultaneously enter the data in to the system and that the final dataset will be developed once all inconsistency between the two sets of data are resolved. While this increases the cost but in case of large scale survey this is indispensable.

5. **Technical direction:** The management staff or the team that will provide the technical direction and represent the contractor for the particular assignment.

6. **Deliverables:** This is one of the most crucial sections of the TOR and should be detailed. It should list all items that are expected to be delivered by the contracted organization and the due dates (time line). The following gives a list of items that may be included in a survey type assignment (not exhaustive):

    a. Sample Size and Plan (geographically disaggregated)
    b. CVs of enumerators
    c. Draft Questionnaires
    d. Finalized Questionnaires (after pre testing and in local language)
    e. Participants list of enumerator and supervisor training
    f. Codified cleaned database (in SPSS & Excel Format)
    g. Hard copy of the filled up questionnaire
    h. Draft Report
    i. Final Report

7. **Budget and Invoice:** Usually financial details are provided in the contract however a small explanation may be provided in relation to when an invoice is due and how they will be processed.

Annex III: Case studies

Case Study 1

Impact assessment of promoting the use of appropriate soil nutrients by palm oil producing farmers in Thailand with T-G PEC[10]

Overview of TG-PEC

Description of the program: The Thai-German Programme for Enterprise Competitiveness (T-G PEC) was a programme funded by the German Federal Ministry for Economic Cooperation and Development (BMZ) and implemented by GIZ. The objective of the programme was to improve the competitiveness of small and medium sized enterprises in the Thai agro-industry sector. The programme specifically focused on improving the institutional environment for specific value chains, like palm oil, fresh fruits and vegetables, using a market development approach.

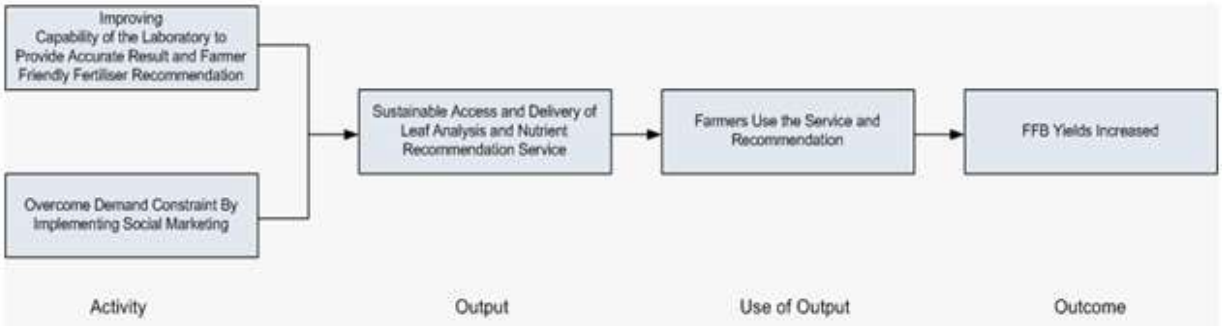
Impact assessment of the leaf analysis intervention

Background: Palm oil is the second most consumed vegetable oil globally, and is the 2nd most important cash crop in the southern part of Thailand with around 70,000 SMEs involved in palm oil production in the country. However Thailand hardly contributes to this global market. One of the key reasons for this is due to low productivity resulting from poor farm management[11]. T-GPEC worked on number of interventions focusing on improving the competitiveness of these small palm oil producers and one of them was to facilitate the growth of a market for leaf analysis services for fertilizer recommendation. The idea being that judicious use of fertilizer would result in optimal cost effective yield of Fresh Fruit Bunches (FFB). TG-PEC has worked with the company Vichitbhan Palm Oil to turn its leaf analysis laboratory into an accurate and cost-effective facility for the Thai palm oil industry. The aim of this intervention was to provide about 1,000 plantations with expert recommendations on the economic use of fertilizer. The following diagram gives a simplified view of the result chain of the intervention[12].

[10] Special thanks to Phitcha Wanitphon for his assistance in preparing this case study.

[11] For more detail on the sector and T-G PEC's contribution read: *Significance of the Model for Thailand – Role of the State, Constitutional Values, and Planning Models by Visoot Phongsathorn*

[12] From the presentation *"The Thai Palm Oil Sub Sector"* by Jim Tomecko



Impact assessment: The program used a difference in difference methodological framework for estimating impact of this intervention to compare difference between treatment group (farmers who received recommendation on fertilizer use) and control group (farmers who did not receive such recommendation). The impact assessment guideline TG-PEC stipulates that the programme will use a before and after methodology usually with a control group for this type of embedded services. The guideline suggests that at least 30 treatment and 10 control farmers should be taken per intervention during an assessment. While this may not be statistically significant but given the modest size the sample it is possible to conduct in-depth interviews with semi-structured questionnaire, which allows for evaluation of more qualitative indicators like satisfaction with the services, which is crucial for the sustainability of the intervention. Furthermore it recommends that 2 interviewers conduct each interview together; this reduces interviewer bias, and reduces error.

In case of the present intervention, 50 treatment and 12 control farmers were interviewed. A baseline was conducted before the service was launched; the treatment farmers were selected non-randomly[13] based on the training invitee list and the control farmers were chosen based on their similarity (age, plantation size, geographic spread) with the treatment group. From the assessments it was found that for the small farmers accessing the service, yield increase was about 20% and net income increase was about 26%.

[13] It was done purposively using age and size of plantation as criteria. This ensured sufficient heterogeneity within the sample

Case Study 2

Impact assessment of Minipack seed intervention with Katalyst[14]

Overview of Katalyst

Description of the Programme: Katalyst is a multi-donor market development programme in Bangladesh. It is implemented under the Ministry of Commerce (MoC) of the Government of Bangladesh by Swisscontact and GIZ International Services. Katalyst is currently active in various agricultural sectors including maize, vegetable, fish, prawn, fertilizer, and seed. In its first phase (2003-2008), Katalyst was set up as a Business Development Services (BDS) programme and has gradually evolved into a more general market development/Making Markets Work for the Poor (MMW4P) programme. In its second phase (2008-2013), Katalyst aims to boost the income and competitiveness of 2.3 million farmers and small businesses[15].

Impact assessment of Seed mini pack intervention

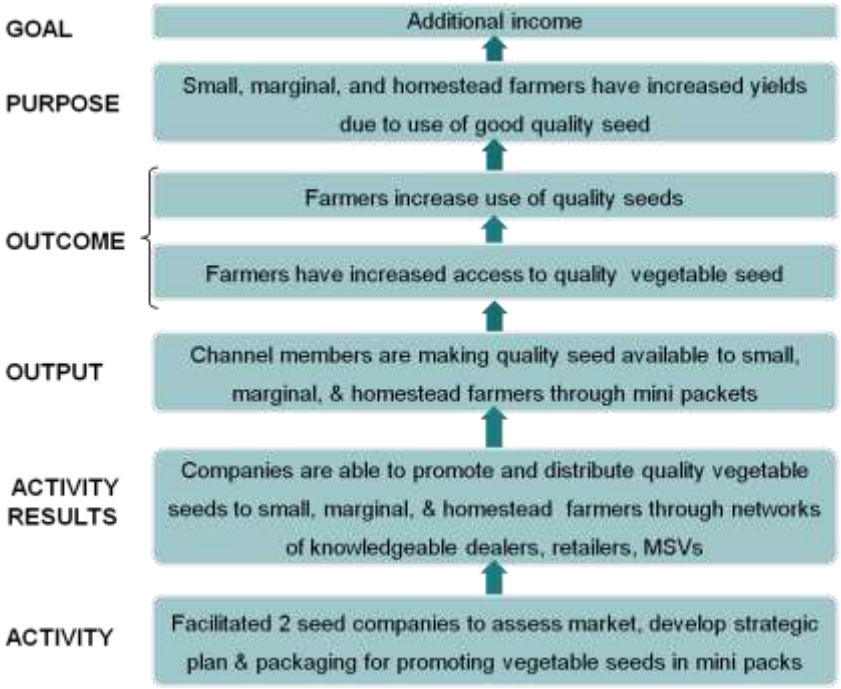
Background: Katalyst's vision in the seed sector is to *'increase productivity and competitiveness of jute, vegetable and potato farmers through facilitation of increased availability and usage of quality seeds'.* In this intervention Katalyst and Action for Enterprise (AFE) facilitated private seed companies A. R. Malik & Company (Pvt.) Limited and Lal Teer Seeds Limited (LTSL) to expand their client base by including small and marginal farmers through facilitating introduction of wide selection of customized mini packs of quality vegetable seed[16]. The idea being farmers who use these high quality minipack vegetable seed

[14] Special thanks to Shovan Chakraborty and Markus Kupper for all of their assistance in preparing this case study.

[15] For more information on Katalyst, visit www.Katalyst.com.bd

[16] Small and marginal farmers' increased access to mini packs of quality seed, News Issues 40, Katalyst Bangladesh; Reference : https://www.enterprise-development.org/wp-content/uploads/Katalyst_A.pdf

would experience increase in yield and thus income. The following figure gives the simplified result chain of the intervention.



Impact assessment: It was decided that the research design for the assessment was going to be similar in nature to that of difference in difference methodology. This meant that Katalyst would estimate the income and yield impact by differencing the 'before' and 'after' income and yield value of control (non user of quality minipack seed) farmers and treatment (user of quality minipack seed) farmers. The data collection tool was chosen to be structured questionnaire survey with primary emphasis on quantitative data. Challenge arose in identifying user farmer because retailers (service provider) were unable to recall who bought these fast moving low cost mini packet of seeds (usually costs $0.25 per packet). So while one could get the total volume of sales of such packets from the companies, tracing the packets to individual user farmer became impossible. Therefore an innovative new approach was designed:

Step 1. Track users by appointing enumerators (15) with retailers and *MSVs* at bazaars & *haats (35 sub-districts)* during peak time of seed sales for over a week

Step 2. Choose User samples from the list (over 600) above for subsequent user survey

Step 3. Conduct 'before' survey for the user group (13 sub-districts, 6 minipacks seed varieties)

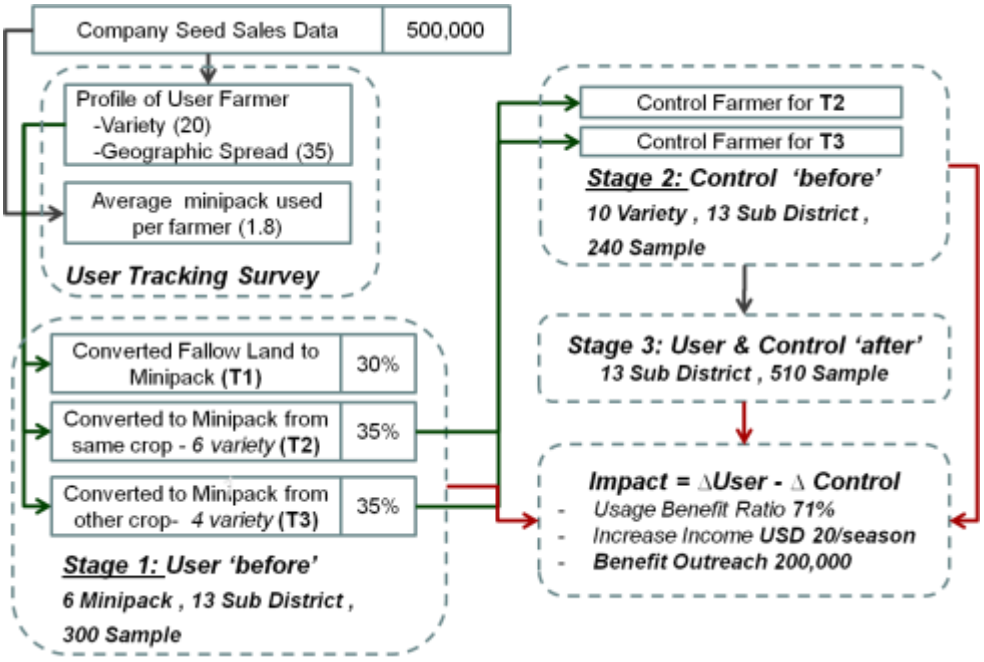
Step 4. Identify control group based on the profile of the user group

Step 5. Conduct 'before' survey for the control group

Step 6. Conduct 'after' survey for both control and treatment groups.

Step 7. Estimate the difference in difference in yield and income between control and treatment group.

It was crucial to conduct steps 3, 4 and 5 in close succession before the cultivation season started in order to ensure farmers remember what they cultivated in those fields prior to this season. In one sense this used retroactive baseline however it would have been quite unfeasible to conduct baseline prior to the launching of the intervention since one could not predict a priori where the sales would take place. The following figure diagrammatically shows the stages and types of control group that were selected *for computing the plausible attributable impact of Katalyst intervention on farmers.*



As can be seen from the figure above that the user sample was reduced to 6 varieties of minipacks from the initial tracking survey, which included 20 varieties. This was done because it was found that these 6 varieties provided significant number of samples with sufficient geographic spread. In case of control 10 varieties were used because some of the user farmers switched from a different crop to the specific minipack variety. The income impact is for per farmer in one season (usually there are two vegetable seasons in a year).

Case Study 3

Impact assessment of EACFFPC Training Course on Freight Forwarder Performance in Rwanda with TMEA[17]

Overview of TMEA

Description of the Programme: TradeMark East Africa (TMEA) is provides technical and monetary support to the East African Community (EAC) Secretariat[18], national governments, private sector and civil society organisations so that there is greater integration of markets and improved trade within the East African region[19]. TMEA is a not for profit organisation that receives funding from the governments of Belgium, Denmark, Netherlands, Sweden and United Kingdom. TMEA projects include infrastructure, business environment, public sector organisational development and private sector and civil society advocacy for greater regional integration.

Impact assessment of the training course

Background: The Federation of East African Freight Forwarders Associations (FEAFFA) is an apex body of national associations of clearing and forwarding agents in the EAC. Their primary responsibility entails training of clearing and forwarding agents, advocacy, transport logistics information dissemination, and membership development. In 2007, with support from USAID, the association launched a training program called the East Africa Customs and Freight Forwarding Practising Certificate (EACFFPC). The EACFFPC is a joint program between East Africa Revenue Authorities (EARAs) and the national freight forwarding associations affiliated to FEAFFA. The training is expected to increase the competencies of customs and freight forwarding agents. Improved knowledge and skills can help the agents to make fewer deliberate and inadvertent errors in completing import and export documentation. This knowledge also allows them to more easily identify inconsistencies or mistakes in information submitted from importers. If documentation is correct, this is expected to reduce import and export processing time delays, and finally reduce transaction cost of doing trade and business in EAC region.

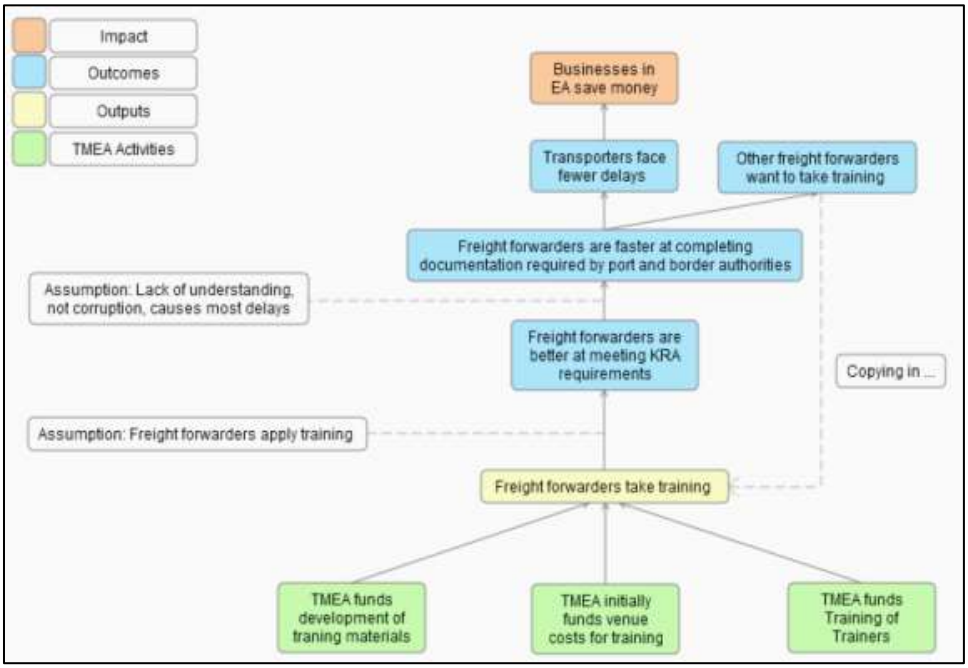
Since 2011, with assistance from TMEA, the training program has been reviewed, curriculum improved and training capacity expanded in order to achieve a critical mass of 4,500 trained customs agents by end of 2013. From 2014, it is expected that the possession of the certificate will become a precondition

[17] Special thanks to Adam Kessler and Donna Loveridge for all of their assistance in preparing this case study.

[18] East African Community (EAC) is an intergovernmental organisation comprising five countries in East Africa - Burundi, Kenya, Rwanda, Tanzania and Uganda. TMEA also operates in South Sudan which has yet to join EAC.

[19] For more information on TMEA, visit : www.trademarkea.com

for acquiring all agent operating licenses within the EAC. The Rwanda Revenue Authority has already started implementing this requirement. The following figure shows the impact logic of the intervention.



Result chain of the EACFFPC training program intervention[20]

Impact assessment: The assessment in this present case focused on impact of EACFFPC training course on freight forwarder performance in Rwanda. It addressed this by exploring the impact of the training course on the number of errors made by freight forwarding companies in Rwanda. This was measured by means of a proxy indicator, which was the number of modifications made on a lodged document and recorded on the Rwandan Revenue Authorities' electronic system. The assessment covered the period between 2009 and 2011, and data were collected for all 97 freight forwarders operational during that period. It examined the efficacy of the existing course and therefore tested the assumptions outlined in the results chain. . Following the implementation of further training, the assessment can be re-run and a comparison made between the effectiveness of the previous course and the updated course. Further assessments could also be expanded to cover more countries.

Methodology and findings: The study used a difference-in-difference research design with quantitative data using available secondary sources, namely data from Revenue Authorities' electronic system and

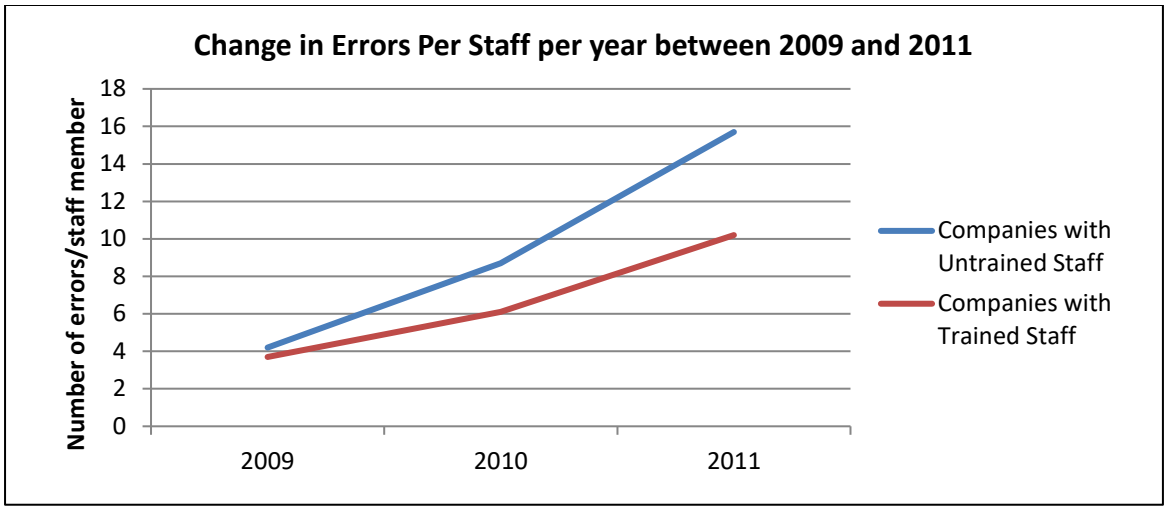
[20] Case Study on TradeMark EA's Experience with the DCED Standard, *op.cit.*

the list of trained staff from the training institute's reports. The study compared errors made by companies with trained staff (treatment) with those with untrained staff (control) at two points from 2009 and 2011. If training caused trained staff to make less errors then the difference between companies with no trained staff and those with trained staff would gradually increase in time as companies had an increasing number of trained staff and they would make fewer errors. The results were further strengthened by triangulation[21] using various statistical analyses. If the training is useful then one should expect a statistically significant difference in errors between the control and treatment group. The following analyses were carried out to ascertain the impact of the training:

- The percentage of trained staff in an organisation was correlated with the number of errors each staff made. A correlation test is used to see the statistical relationship between a dependent and independent variable. For this case the percentage of trained staff was the independent variable and the errors per staff was the dependent variable.[22] A weak, albeit statistically significant, relationship was found that showed that an untrained staff member makes nine errors per year, while a trained staff member is likely to make just one error per year.
- The finding was further strengthened by the fact that no such relationship was found when a similar analysis was carried out with number of errors per staff against percentage of staff failed in the training. Which meant just by participating in the course did not result in improved performance, one had to pass the test implying internalize the learning.
- The following figure shows that between 2009 and 2011 the difference in errors per staff is increasing between the treatment and control companies. It might seem that the total number of errors has gone up for both the groups. Consultation with experts suggested that there could plausibly be two explanation for this : 1) Increased volume of traffic over this period naturally placed greater burden on freight forwarders hence for both treatment and control group the trend is upward sloping; 2) Employee numbers were only provided for 2012, and therefore it was assumed that this remained constant over all years. However, this is unlikely and it is more probable that employee numbers were less in previous years and as a result errors per staff are likely to be understated in 2009 and 2010. If errors per staff are actually greater, the difference between the treatment and control groups would be higher.

[21] Theory triangulation: use of multiple perspectives and theories to interpret the results of a study.

[22] A regression line was drawn with % trained staff as an independent variable and errors per staff as dependent variable



- From the graph it can be seen that the difference in errors per staff between treatment and control companies is around 5.5 in 2011. The Rwandan Revenue Authority charge $10 for each modification (classed as an error for the purpose of the assessment) to lodgement forms. Notwithstanding other costs such as time delay, loss of customers etc, $10 is the minimum cost of one error and as such companies with trained staff spent $55 less per staff member on correcting errors in 2011. . The study found that firm had six or seven staff, on average, resulting in unnecessary costs of $385 a year, which is a substantial cost for the company.
- It could be argued that companies who train their staff are better performers to begin with and hence there might be a case of selection bias. However, no correlation was found between company ability[23] and number of staff trained, meaning better performing companies were not more likely to send their staff on training. Therefore, the relationship between training taken and reduced number of errors is likely to be causal.

[23] Number of errors made in 2009 was taken as a proxy for company ability.

Resources

Data Collection Tools –

EuropeAid Evaluation Methods, https://europa.eu/capacity4dev/evaluation_guidelines/wiki/en-methodological-bases-and-approach-0

Research Methods Knowledge Base, http://www.socialresearchmethods.net/kb/

Evaluation Toolbox - http://evaluationtoolbox.net.au/index.php?option=com_content&view=article&id=51&Itemid=5

Causal Inference –

Angrist, Joshua D. and Jörn-Steffen Pischke (2010) "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics,"

Selection bias -

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1996), "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," Proceedings of the National Academy of Sciences, Vol. 93, No. 23, pp. 13416-13420.

Randomized Control Trial-

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2008), "Using Randomization in Development Economics: A Toolkit," in T. Schultz and John Strauss (eds.) Handbook of Development Economics, Vol. 4, pp. 3895-3962.

Critical examination of RCT –

Deaton, A, 2010. "Instruments, Randomization, and Learning about Development," Journal of Economic Literature, American Economic Association, vol. 48(2), pages 424-55, June.

Instrumental Variable –

Heckman, James J., Sergio Urzua, and Edward Vytlacil, 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." Review of Economics and Statistics, 88(3): 389-432.

Regression Discontinuity-

Imbens, Guido W. and Thomas Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics, Vol. 142, No. 2, pp. 615-635.

Propensity Scoring (Matching) –

Caliendo, Marco and Sabine Kopeinig (2008), "Some practical guidance for the implementation of propensity score matching," Journal of Economic Surveys, Vol. 22, No. 1, pp. 31-72.

Difference in Difference –

Meyer, Bruce D. (1995), "Natural and Quasi-Experiments in Economics," Journal of Business and Economics Statistics. Vol. 13, No. 2, pp. 151-161.

Overall research design -

Imbens, Guido W., and Jeffrey M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature, 47:1, 5-86.

M&E framework –

The 2008 Reader - Private Sector Development: Measuring and Reporting Results (2008). Eighth Annual BDS Seminar - Chiang Mai, Thailand, ITC, ILO.

Imas Linda G. Morra and Rist Ray C, 2009, "The Road to Results: Designing and Conducting Effective Development Evaluations", World Bank Publications

Elliot Stern, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, Barbara Befani (2012), "Broadening the range of designs and methods for impact evaluations", DFID Working Paper 38

Kaplinsky, Raphael, and Mike Morris. "A handbook for value chain research. Vol. 113". IDRC, 2001.

Miehlbradt Alexandra and Linda Jones, "Market research for value chain initiatives - Information to Action: A Toolkit Series for Market Development Practitioners", Mennonite Economic Development Associates

Mixed method research –

Johnson, R. Burke, and Anthony J. Onwuegbuzie. "Mixed methods research: A research paradigm whose time has come." Educational researcher 33, no. 7 (2004): 14-26.

Developing TOR-

Dawn Roberts, Nidhi Khattri, and Arianne Wessal "Writing terms of reference for an evaluation: A how-to guide", IEG Blue Booklet series, World Bank (2011)

Outlier Analysis-

Iglewicz, Boris, and Sharmila Banerjee. "A simple univariate outlier identification procedure." In Proceedings of the Annual Meeting of the American Statistical Association. 2001.

References

Banerjee, Abhijit V., 2007, “Making aid work”, Cambridge, MIT Press.

Barnett, Vic, and Toby Lewis. 1984, "Outliers in statistical data." Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed. 1 .

Caracelli VJ, Green JC and Graham WF (1989) Toward a Conceptual Framework for Mixed-Method Evaluation Designs. Educational Evaluation and Policy Analysis. 11 (3).

Cartwright, Nancy, 2007, “Are RCTs the gold standard?” Biosocieties, 2, 11–20.

Concato, John, Nirav Shah and Ralph I. Horwitz, 2000, “Randomized, controlled trials, observational studies, and the hierarchy of research designs,” New England Journal of Medicine, 342(25), 1887–92.

Deaton, A, 2010. "Instruments, Randomization, and Learning about Development," Journal of Economic Literature, American Economic Association, vol. 48(2), pages 424-55, June.

Duflo, Esther, 2004, “Scaling up and evaluation,” Annual World Bank Conference on Development Economics 2004, Washington, DC. The World Bank.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, 2008. “Using Randomization in Development Economics Research: A Toolkit.” In Handbook of Development Economics, Volume 4, ed. T. Paul Schultz and John Strauss, 3895-3962. Amsterdam and Oxford: Elsevier, North-Holland.

Elliot Stern, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, Barbara Befani (2012), Broadening the range of designs and methods for impact evaluations, DFID Working Paper 38

Evans, Bill (2008), Difference in difference models, ECON 47950, Fall 2008, Oakland University Bamberger M, Rao V and Woolcock M (2010) Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development, Policy Research Working Paper, 5245: The World Bank.

Gay, L., 1987. Educational research: competencies for analysis and application. Merrill Pub. Co., Columbus.

Golafshani, Nahid. "Understanding reliability and validity in qualitative research." The qualitative report 8, no. 4 (2003): 597-607.

Guest, Greg; Bunce, Arwen & Johnson, Laura (2006). "How many interviews are enough? An experiment with data saturation and variability". Field Methods, 18(1), 59-82.

Holland, Paul W. (1986), “Statistics and Causal Inference,” Journal of the American Statistical Association, Vol. 81, No. 396, pp. 945-960.

Hume, David (1848), An Enquiry concerning human understanding. Section 7.2.

Iglewicz, Boris, and Sharmila Banerjee. "A simple univariate outlier identification procedure." In Proceedings of the Annual Meeting of the American Statistical Association. 2001.

Imas Linda G.Morra and Rist Ray C, 2009, “The Road to Results: Designing and Conducting Effective Development Evaluations”, World Bank Publications

Imbens, Guido W., and Jeffrey M Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” Journal of Economic Literature, 47:1, 5-86.

Johnson, R. Burke, and Anthony J. Onwuegbuzie. "Mixed methods research: A research paradigm whose time has come." Educational researcher 33, no. 7 (2004): 14-26.

Kusek, Jody Zall, and Ray C. Rist. 2004. Ten Steps to Building a Results-Based Monitoring and Evaluation System. Washington, DC: World Bank.

Lewis, David (1973), “Causation,” Journal of Philosophy, Vol. 70, No. 17, pp. 556-567.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.

Mason, Mark. "Sample size and saturation in PhD studies using qualitative interviews." In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, vol. 11, no. 3. 2010

Onwuegbuzie, Anthony J., and Kathleen MT Collins. "A typology of mixed methods sampling designs in social science research." The Qualitative Report 12, no. 2 (2007): 281-316.

Osborn, David, and Ted Gaebler. 1992. Reinventing Government. Boston: Addison-Wesley Publishing.

Thanasegaran, Ganesh. "Reliability and Validity Issues in Research." Integration & Dissemination 4 (2009): 35-40.

White, Howard, and Daniel Phillips. Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. Working Paper 15, International Initiative for Impact Evaluation. New Delhi: 3ie. https://www.3ieimpact.org/evidence-hub/publications/working-papers/addressing-attribution-cause-and-effect-small-n-impact

Yin, R.K. Case study research design and methods. 4th ed. Thousand Oaks, CA: Sage publications Inc.; 2009