Measuring impacts using experimental designs

– and their application at GIZ

Randomised controlled trials (RCTs) for measuring the impacts of development projects and programmes are currently attracting considerable attention in the German media, where they are described as the 'gold standard' for evaluating impacts. However, this debate is not new in professional circles. For several years now, GIZ and its predecessor organisations have been looking at RCTs and the ways in which they can be used to evaluate projects and programmes. Based on these experiences, the Monitoring and Evaluation Unit has drawn up recommendations for the use of RCTs taking account of the realities of projects and programmes on the ground.

Results orientation is one of the key quality features of GIZ's work. Our projects and programmes are regularly evaluated so that we can reliably and credibly substantiate the impacts[1] of our work. What we are primarily interested in here is finding answers to the following questions: What works, how, why and under what conditions?

One particular (experimental) study design for measuring results has been the subject of much attention in the German media since 2011 – randomised controlled trials (RCTs). This debate has been advanced in particular by the publications of Abhijit V. Banerjee and Esther Duflo of J-PAL[2]. But it is by no means new, having continued for many years in international and national evaluation circles both within and beyond the realms of development cooperation. Here, RCTs are seen as an important approach, but are no longer considered to be *the* gold standard for evaluating impacts.

This paper presents the opportunities and limitations of RCTs against the background of the international debate and our own experiences, and identifies recommendations for their use at GIZ. It is aimed firstly at GIZ staff who is involved in measuring impacts, in particular project and programme managers, who have to create the preconditions for subsequent evaluation (especially impacts attribution) at the very start of the projects and programmes. It also informs interested members of the public about GIZ's stance on using RCTs.

What are RCTs?

Randomised controlled trials are based on the idea that the impact of a project can be determined if we know what would have happened had the intervention not taken place (the 'counterfactual' situation). The group that is participating in the project is compared with a control group that is not benefiting from the intervention. People are randomly assigned to one of the two groups *before* the intervention commences. By so doing, it is possible to largely exclude the likelihood that any differences noted after the intervention has been concluded are due to factors other than the intervention itself.

The potential and strengths of RCTs

Evaluations must address the question of whether factors other than the intervention itself could be responsible for the impacts observed. Donors cannot retrospectively assess what might have happened without their intervention. Nor can this be established by comparing people who have benefited from an intervention with those who have not, since the similarities and differences between the two groups must be taken into account. RCTs attempt to resolve this problem. Randomly selecting intervention and control groups establishes a counterfactual situation before the intervention commences, making it possible to determine differences over time.

It is easiest to establish a counterfactual situation in large-scale individual interventions that are directed towards many different entities, such as individuals, households, schools, hospitals, businesses, villages and districts.

1 Following the terminology of the NONIE Guidance on Impact Evaluation (2009), the term impact is used to denote short-, medium- and long-term effects of an intervention. Thus, this paper subscribes to a more comprehensive definition of impact than the OECD-DAC definition does.

2 Abdul Latif Jameel Poverty Action Lab at MIT

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

RCTs are designed to include a control group that is similar to the intervention group. This largely precludes any alternative explanations for the occurrence or extent of the impacts identified in the intervention group. The conclusions drawn from RCTs are therefore assumed to have high internal validity.

The experimental design therefore makes it possible to establish *causal links* and thus measure the contribution made by an individual project, programme or intervention. The main method used in RCTs is the statistical analysis of survey-based data, often using econometric procedures[3].

In 2010, an RCT was used on behalf of GTZ (Evaluation Unit) for the first time in Senegal to investigate the impacts of distributing improved stoves. The aim was to determine the impact on firewood usage, health, time use and financial expenditure. First of all, 253 households were interviewed for the baseline, then they were randomly given either a stove (target group) or a bag of rice (control group). To check on stove usage and any associated technical problems, three interim surveys were carried out. After a year, the usage level was 87%. Standardised questionnaires were used to capture all the socio-economic dimensions of the households, focusing in particular on fuel access and consumption and cooking habits, and triangulated with information from semi-structured interviews with key informants. The results were statistically significant: Consumption of firewood fell by 30%, daily cooking time fell by 70 minutes per day, and eye infections and respiratory problems decreased. On the basis of these results, more stoves are now being distributed.

Since experimental (laboratory) studies have a long scientific tradition, the design is well documented and established. It generates 'objective' data and enjoys a high degree of credibility.

The limitations and weaknesses of RCTs

One of the inherent methodological weaknesses of RCTs is their limited external validity – in other words, the extent to which the findings of one study can be transferred to another context is unclear. Implementers are often guilty of improper generalisations, but it is questionable whether problems in India can be resolved by drawing on the findings of an experiment in Africa.

In using predominantly quantitative methods (notably standardised surveys), there is a risk of perceiving reality through a severely restricted filter. Interviewees are sometimes unable to raise aspects of particular importance to them if these have not been included in the questionnaire, making it difficult or even impossible to detect unexpected impacts.

An associated criticism is that with many econometric procedures, it is impossible to analyse how and why impacts are generated or are not generated (the 'black box'). They are therefore unable to ascertain the reason for the absence of the intended impacts, for example planning or implementation errors. So RCTs often do not tell us why something is effective or ineffective.

Other challenges relate to the feasibility of conducting RCTs under the practical conditions of development evaluation. In practice, it is usually not possible to adhere to the scientific requirement that experiments must be 'triple blind'[4] or at least 'double blind'[5] in order to produce valid findings. And it is often not possible to adequately control 'spill-over' effects on the control group.

Certain features of GIZ-assisted projects and programmes also frequently make the use of RCTs impossible or extremely difficult: Target groups are often intentionally selected (for example, to contain particularly motivated people) and are therefore not directly comparable with people who are not taking part in a programme; many projects and programmes operate at a national level, with everybody often benefiting simultaneously from certain interventions (e.g. changes in legislation) – in this situation it is not possible to set up a control group; macroeconomic issues cannot be addressed using control groups; in many projects and programmes, the high level of complexity makes it difficult to measure impacts across several different interventions; conversely, it does not always make sense to consider individual interventions in isolation; it is harder to quantify impacts in some sectors than in others (e.g. good governance vs. vocational training). So the necessary preconditions for RCTs are usually not present in projects and programmes.

Other factors to be considered include the considerable costs and effort required to conduct RCTs, as they involve the extensive collection of primary data. It is important to carefully weigh up the costs and benefits of RCTs. For this reason, the literature contains very few large-scale impact evaluations.

Some critics of RCTs raise ethical concerns, saying that the random allocation of people to one of two groups, only one of which benefits from an intervention, is untenable.

Position and recommendations of the Monitoring and Evaluation Unit

RCTs are an important approach to attributing impacts more precisely in evaluation practice. They have certain strengths that other approaches cannot offer. In this respect, there is still untapped potential within GIZ that

[3] Econometrics is a branch of economics that combines economic theory with mathematical methods and statistical data to empirically verify economic models and quantitatively analyse economic phenomena.

[4] Neither the members of the intervention and control groups, nor the project staff, nor the evaluating experts know who is in the control group and who is in the intervention group.

[5] Neither the members of the intervention and control groups nor the project staff know who is in the control group and who is in the intervention group.

should be exploited to improve how we substantiate the impacts of our work. Nevertheless, opportunities for using RCTs within a GIZ context remain limited due to the project and programme characteristics mentioned above (type and scope), which prevent or impede the formation of control groups, or mean this is not expedient.

The unit is generally of the view that RCTs should not be considered superior to other designs, but are just one of many possible approaches to evaluating impacts. In its publication *Guidance on Impact Evaluation* (2009), NONIE (the Network of Networks on Impact Evaluation) stresses the importance of rigorous quantitative methods for the causal attribution of impacts, but recommends using a mix of methods that combines the strengths of a variety of different quantitative and qualitative methods. The literature contains many pointers in this respect. For example, in his book *Utilization-Focused Evaluation* (2008), Patton presents a comprehensive selection of different evaluation designs that can be used for a wide variety of evaluation issues. In their publication *Real World Evaluation* (2nd edition 2011), Bamberger/Rugh/Mabry also address the challenges of impact evaluations when there are time and cost constraints and a shortage of key data.

The Monitoring and Evaluation Unit has also investigated concepts that could be used to more accurately measure impacts in its independent evaluations. The focus here was on practicality, bearing in mind the methodological requirements of rigorous impact evaluation. The Unit has been able to develop an approach that can be implemented within acceptable financial limits and timescales and takes account of the reality of GIZ-assisted projects and programmes. By carrying out ex-ante evaluations, the Unit also intends to investigate the necessary preconditions for RCTs at the project planning stage, jointly examine with the Sectoral Department which projects and programmes would be suitable for conducting a pilot RCT, and advise and support these projects and programmes in using RCTs in connection with decentralised evaluations.

With regard to using RCTs within GIZ, the Monitoring and Evaluation Unit makes the following recommendations:

1.) **RCTs should be used if it is possible to do so and if it makes sense in terms of content, strategy and funding. However, their use across the board is neither expected nor required.**

In principle, RCTs are *possible* if a control group has been set up at the start of the project or programme, if the intervention in question is aimed at individuals, and if a sufficiently large number of suitable entities exists to enable appropriate statistical analysis to be carried out.

We would consider RCTs to be *advisable* primarily for examining the impact of individual large-scale interventions, provided these interventions can be considered in isolation from the overall project/programme context.

In view of the high *costs* and effort levels associated with RCTs, we would advise adopting a *strategic approach* when selecting projects or programmes in which to use RCTs (e.g. where new or innovative interventions are to be used). RCTs can then be used during the project/programme to establish what is effective and what is not, with this information then being used for example to decide whether to *scale up* the interventions in question. RCTs are thus being used not merely as an accountability tool after the end of projects and programmes, but also to fulfil the important function of evidence-based steering at ongoing projects.

2.) RCTs are very good at establishing which interventions work and which don't. But they need to be embedded within a broader, hypothesis-based evaluation design that also addresses the question of why a particular intervention is effective in a specific context or not. It is generally preferable to use a mix of methods rather than individual ones, as different methods have different strengths and weaknesses. **RCTs should therefore be used in combination with qualitative methods wherever possible** (in order to answer the 'why' question, provide the opportunity to identify unexpected impacts, record impacts that are difficult to measure and check for spill-over effects).

3.) Where it is not possible to define control groups at the start of projects/programmes or where randomly selecting them does not make sense, **we recommend taking into consideration the creation of comparison groups,** so that a quasi-experimental design can be used in any subsequent evaluation. Comparison groups should be 'constructed' using specific matching procedures to generate the greatest possible similarity with the intervention group. In such circumstances, it is possible for quasi-experimental designs to generate 'robust' data and provide evidence-based information regarding intervention impacts. If a comparison group is not established at the start of a project or programme, this can be done retrospectively at the time of evaluation, although this is a much less satisfactory arrangement.

In such circumstances, the initial situation (baseline) will also have to be reconstructed – a time-consuming and less accurate process.

4.) The prerequisites for impact measurement and the general evaluability of projects and programmes should be established in the planning phase: setting up a comparison or control group, conducting a baseline survey of both groups, and implementing an M&E system that enables time series data to be captured in order to continuously determine changes over time. A suitable evaluation design can then be selected based on these parameters. **When appraising and planning projects and programmes, greater attention should be paid to the above prerequisites so that evidence-based statements can subsequently be made regarding the impacts of interventions.**

5.) **It is possible to mitigate any ethical concerns regarding the use of RCTs by using a 'phasing-in' process.** GIZ-assisted projects and programmes often start with a pilot region or group in the first phase, then expand to other regions or groups in subsequent phases. These other regions or groups can initially act as control groups, but then subsequently benefit when they themselves become the project's intervention region or group.

Conclusion

There are still only a few situations in a GIZ context where RCTs can be used and where it is feasible to use them from a methodological and financial point of view. So we still need to use alternative approaches for evaluating impacts. It is important to explore whether a counterfactual situation can be established and if so how, so that impacts can be causally attributed with the greatest possible accuracy. Furthermore, there are statistical approaches of impact evaluations that do not involve a counterfactual design, and also theory-based approaches that use a generative perception of causality to establish how programmes function and why. The Monitoring and Evaluation Unit intends to test out these approaches in depth and also encourage projects and programmes to make use of the full range of existing impact evaluation tools.

Further Information

- http://pooreconomics.com/
- http://www.realworldevaluation.org/
- http://www.giz.de/de/ueber_die_giz/97.html

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

Produced by:
Dr Stefanie Krapp, Sabine Dinges
Monitoring and Evaluation Unit
June 2012

Contact:
Sabine Dinges
Dag-Hammarskjöld-Weg 1-5
D-65760 Eschborn, Germany
T +49 (0) 61 96 / 79 - 1633
F +49 (0) 61 96 / 79 80 - 1633
E sabine.dinges@giz.de
I www.giz.de