

WHEN TO DO A RIGOROUS IMPACT EVALUATION?

A guidance tool

2021

*Dr Marion Krämer
Lea Jechel
Timo Kretschmer
Elisabeth Schneider*

Authors

Dr Marion Krämer
Lea Jechel
Timo Kretschmer
Elisabeth Schneider

Responsible

Dr Martin Bruder

Editing

Janet King

Bibliographical reference

Krämer et al. (2021), *When to do a rigorous impact evaluation. A guidance tool*, German Institute for Development Evaluation (DEval), Bonn.

© German Institute for Development Evaluation (DEval), 2021

Published by

German Institute for Development Evaluation (DEval)
Fritz-Schäffer-Straße 26
53113 Bonn, Germany

Phone: +49 (0)228 33 69 07-0

E-Mail: info@DEval.org

www.DEval.org

The German Institute for Development Evaluation (DEval) is mandated by the German Federal Ministry for Economic Cooperation and Development (BMZ) to independently analyze and assess German development interventions.

All DEval publications can be downloaded as a PDF-file from the DEval website:

<https://www.deval.org/en/publications>

This guidance tool accompanies the following publication:

Krämer et al. (2021), Rigorous Impact Evaluation: Evidence generation and take-up in German development cooperation, German Institute for Development Evaluation (DEval), Bonn.

CONTENTS

Why do we need this guidance tool?	1
What is this guidance tool based on?.....	1
Guidance tool	2
Step 1: Is it an RIE evaluation question?.....	3
Step 2: When is it feasible to conduct an RIE?	3
Step 3: When is it useful to conduct an RIE?	5
References.....	7

Figures

Figure 1	Three steps of the guidance tool.....	2
----------	---------------------------------------	---

Boxes

Box 1	What is rigorous impact evaluation?.....	1
-------	--	---

ABBREVIATIONS AND ACRONYMS

ADB	Asian Development Bank
BMZ	Federal Ministry for Economic Cooperation and Development
DFG	German Research Foundation
FCDO	Foreign, Commonwealth & Development Office (formerly DFID)
GDC	German development cooperation
J-PAL	The Abdul Latif Jameel Poverty Action Lab
MCC	Millennium Challenge Corporation
ODI	Overseas Development Institute
OECD	Organisation for Economic Co-operation and Development
RCT	randomised controlled trial
RIE	rigorous impact evaluation
ToC	theory of change
UN	United Nations
USAID	United States Agency for International Development

WHY DO WE NEED THIS GUIDANCE TOOL?

There is much research on how best to implement a rigorous impact evaluation (RIE) (see Box 1) (Glennester and Takavarasha, 2013; von Schiller, 2021). However, before addressing what to consider during the implementation of an RIE, decision makers need to clarify if and when an RIE is actually the best choice. Hence, they need to do an **RIE evaluability assessment**.¹ For this purpose, we developed an evidence-based, user-oriented and actionable guidance tool for decision makers, at both project and policy level, who are faced with the question of whether they should initiate an RIE or not.

The guidance tool draws on data collected and analysed within the scope of a BMZ-funded DEval research project on RIE in German development cooperation (GDC). This research project investigated the status quo, barriers for both the initiation of RIEs and the take-up of (rigorous) evidence, and potential measures to increase the initiation of RIE and the use of the evidence it produces within GDC (DEval, 2021). Due to its empirical foundation, the guidance tool tackles the practical need of decision makers working in (German) development cooperation.

Box 1 What is rigorous impact evaluation?

RIE as an evaluation approach comprises impact evaluation designs that allow the causal attribution of a mean change in an outcome of interest (e.g. household income) to a specific intervention (e.g. a microloan). To do so, it is necessary to compare what actually happened with the so-called “counterfactual situation”. This means, for example, comparing incomes of households that received microloans with the incomes of the very same households if they had not received the microloan. Because such an observation is logically impossible (because any given household, cannot both receive and not receive a loan at the same time), the counterfactual condition is approximated using experimental and quasi-experimental study designs.

The guidance tool draws on data collected and analysed within the scope of a BMZ-funded DEval research project on RIE in German development cooperation (GDC). This research project investigated the status quo, barriers for both the initiation of RIEs and the take-up of (rigorous) evidence, and potential measures to increase the initiation of RIE and the use of the evidence it produces within GDC (DEval, 2021). Due to its empirical foundation, the guidance tool tackles the practical need of decision makers working in (German) development cooperation.

WHAT IS THIS GUIDANCE TOOL BASED ON?

A **literature review** provided the conceptual basis for the guidance tool. We focused on practice-oriented resources from international development research institutes, and from multilateral and bilateral development organisations (Davies, 2013; Gertler et al., 2016; Gibson and Sautmann, 2021; MCC, 2021; OECD, 2006; Peersman, Gujit and Pasanen, 2015; USAID, 2008; White and Raitzer, 2017). The literature review brought forward numerous criteria on different conceptual levels that can help define when to conduct an RIE. In summary, we find that most criteria are recurring in the literature and that criteria can roughly be categorised into the following: **suitability** of the research question, **feasibility** and **usefulness** of an RIE. The suitability of the research question refers to whether the research under consideration lends itself to an RIE. Feasibility refers to criteria that have to be met for an RIE to be feasible (exclusion criteria), and usefulness to criteria that increase usefulness as the number of met criteria increases (additive criteria).

Secondly, this guidance tool is based on **empirical findings** from our stakeholder interviews and the RIE stocktaking within GDC. Although the criteria presented in this guidance tool appear clearly identifiable and separate from each other, the empirical data revealed a multitude of rather fragmented and variably precise criteria to assess when to conduct an RIE. The most common interviewee response to the question of when to conduct an RIE was the imprecise assumption that it has to be a “suitable project”. Respondents often mentioned certain sectors or implementing institutions as shortcuts to concrete criteria. Yet, feasibility or usefulness of an RIE cannot be generally assigned to specific sectors or stakeholders. Instead, implementers should check the suitability of the research question, the feasibility and usefulness along a list of precise criteria at the level of the individual project.

¹ For more information on evaluability assessment see OECD (2009).

GUIDANCE TOOL

The guidance tool outlines the **three steps** in the decision-making process for or against conducting an RIE (Figure 1). In step 1 we define the type of evaluation questions that can be answered by RIEs. In step 2 we identify criteria for methodological and practical feasibility of an RIE. Whereas methodological feasibility refers to the requirements stipulated by the rigorous design, practical feasibility defines the conditions that have to be met on the ground. If any of the feasibility criteria are not met and also cannot be invoked, a solid RIE cannot be conducted. In step 3 we define criteria for a useful application of RIE. Although an RIE does not necessarily have to fulfil all criteria for usefulness, it will become more useful with an increasing number of criteria being met in this dimension. In sum, in order to confidently launch an RIE, decision makers must **identify** an adequate research question and **ensure** feasibility while **maximising** usefulness.

Each step comprises a number of clearly delineated and precise criteria to decide on whether to conduct an RIE or not. Figure 1 gives an overview of the three steps and the corresponding criteria. The distinction between different decision criteria is illustrated by the use of checkmarks when dealing with exclusion criteria (step 2) and dots when dealing with additive criteria (steps 1 and 3).

Figure 1 Three steps of the guidance tool



Source: own illustration

Step 1: Is it an RIE evaluation question?

The research question is the genuine starting point for choosing the evaluation approach. It is, therefore, the starting point of our guidance tool.

RIEs are rooted in a counterfactual² understanding of impact and they are particularly strong in investigating **causality**. This focus is evident in the questions that RIE can address well:

- Is there a significant change in observed outcome(s) that can be attributed to the intervention?
- How large is this change in the outcome?
- If there are different interventions that might all affect the outcome, which of these interventions causes the greatest change in the outcome?
- If there are different interventions and they cause different changes in the outcome, which of these interventions is the most cost-effective?

In short, if an evaluation is concerned with questions of **cause and effect**, RIE is likely to be the most valid and reliable evaluation design. Also, if questions about the cost-effectiveness of alternative interventions are central, RIE can be extremely helpful. Similarly, if evaluation questions concern not only whether there is an effect of an intervention, but also when and how such an effect occurs, mixed-methods designs (e.g. RIE complemented by case studies) or multi-methods designs (e.g. RIE complemented by mediation analysis), are likely a good choice (Gertler et al., 2016; OECD, 2006; USAID, 2013; White and Masset, 2018).

The decision about the most appropriate evaluation design should be taken against the backdrop of a well-developed theory of change (ToC). A ToC outlines causal linkages between distinct activities, outputs, outcomes and impacts of an intervention. In the case of projects with activities on several levels, it is important to identify and clearly delineate the intervention (or combinations thereof), so it becomes possible to formulate a meaningful evaluation question about causality. In addition, a ToC helps to identify the assumptions underlying the causal linkages (Noltze, Leppert and Harten, 2018; White and Raitzer, 2017).

Step 2: When is it feasible to conduct an RIE?

Feasibility of an RIE can be assessed along an unambiguous **checklist** of criteria that all have to be fulfilled. We differentiate between methodological and practical feasibility criteria.

a) Methodological feasibility

F1. Valid counterfactual

The core element of RIE are control or comparison groups that constitute a counterfactual for the treatment group. In an experimental design, randomisation before the start of the intervention assigns participants to a control or treatment group. This increases the likelihood of a valid counterfactual. In a quasi-experimental design, this deliberate random group-assignment before the start of the intervention did not take place and statistical methods are employed to construct (often retrospectively) valid treatment and comparison groups. Both counterfactuals must be *valid* with regard to three aspects:

- On average, the observable and unobservable baseline characteristics of treatment and control/comparison groups must be as similar as possible.
- There must be no direct or indirect spill-over between the treatment and the control/comparison group.

² A “counterfactual” measures what would have happened in the absence of the intervention. Impact is hence estimated by comparing counterfactual outcomes to those observed under the intervention, while eliminating potential confounding factors.

- The common trend assumption must be met, i.e. that the comparison group would respond to the treatment in the same way as the actual treatment group (Gertler et al., 2016).

Forming comparison groups can be easier or more difficult, depending on the type of intervention, conditions on the ground, or the target group. However, interviewees gave examples of apparently complex development cooperation projects that found solutions to the formation of a comparison group through innovative approaches. One commonly mentioned solution was the so-called phase-in approach.³ Other examples included the use of satellite data or matching methods that allowed the construction of a valid comparison group even once the intervention had started.

F2. Sufficiently large sample size

The required sample size depends on a number of factors, including the expected effect size of the intervention, the variance of the outcome indicator and the number of treatment arms. The smaller the expected effect size, the larger the required sample size. Overall, large sample sizes improve the outcome estimates of an RIE. Yet, large datasets cause high costs in the data collection process. A statistical power calculation helps to estimate the minimum sample size required to detect difference in outcomes between the treatment and comparison groups. Thus, power calculations prevent actual impacts from remaining undetected (Type II error) and at the same time contribute to cost efficiency (Gertler et al., 2016; USAID, 2013).

b) Practical feasibility

F3. Financial resources

RIEs often require substantial financial resources, mostly because of high costs of large n primary data collections (e.g. wages of enumerators). In cases where existing data, such as from monitoring or administrative data, can be used, the cost of the RIE can substantially be reduced. A sample of projects from the World Bank's Strategic Impact Evaluation Fund showed that, on average, about 6% of the total project budget was used to conduct RIEs (Gertler et al., 2016).

F4. High-quality data

High-quality data is a precondition for any reliable and valid empirical analysis. Secondary data can be used, such as monitoring or administrative data, or own primary data can be collected. Access to data and the quality of the data depend on factors such as cooperation with local stakeholders or well-trained enumerators.

F5. RIE-methodological and practical implementation knowledge of the evaluation team

Successful implementation of an RIE requires both specialised technical RIE knowledge (e.g. knowledge of possible RIE designs and their implementation and data analysis skills) and practical understanding of the sectoral and country-specific challenges of project implementation and data collection (e.g. defining adequate outcome measures).

F6. Time horizon

Three temporal parameters are necessary to conduct an RIE. First, for treatment outcomes to be valid, the stability of the intervention should be assured. This means that as long as the data collection is running, the treatment should not change. Second, there must be a plausible amount of time between the intervention and the collection of the data to generate a visible impact: a measurable outcome data. What is considered

³ The phase-in approach serves to delay inclusion in the intervention for parts of the target group so that those who join later serve at first as a comparison group (Funk et al., 2018).

a “plausible amount of time” can vary widely, depending on the nature of the intervention and its intended outcomes, as reflected in the underlying theory of change (Gertler et al., 2016; USAID, 2013). Third, in the case of randomised controlled trials (RCTs) randomisation must take place before the start of the project intervention.

F7. Conducive project context

Gauging the commitment of stakeholders (e.g. partners, project team, funder) and their demand for and openness to an RIE can constitute a practical, people-centred approach to assessing whether implementing an RIE will be feasible. This includes involving staff at all levels: from decision-makers at management level to enumerators in the field. Further, taking account of the socio-political context (e.g. political stability, access etc.) helps to assess whether a successful RIE implementation can be expected.

F8. Compliance with ethical standards

Criticism of RIEs is often based on ethical concerns. In fact, there are a number of ethical considerations when conducting an RIE, given that human beings are the subject of study. In their Ethical Guidelines for Evaluation, the UN Evaluation Group defines four areas of ethical behaviour that must be complied with: integrity, accountability, respect, and beneficence (UNEG, 2020). Ethical compliance also includes the understanding that in a sovereign nation state, support and permission by local authorities to collect data must be requested. Furthermore, to address the complexity and diversity of ethical compliance, an ethical clearance process by an ethics committee or Institutional Review Board (IRB) is required to ascertain the ethical feasibility of the RIE (DFG, 2021; Gault, 2020; Kaplan et al., 2019).

Step 3: When is it useful to conduct an RIE?

The usefulness of an RIE comprises a number of **different criteria**. Not all of these have to be present at the same time to make an RIE useful. Still, the higher the number of criteria met, the more useful the RIE is likely to be.

U1. Function of the evaluation

An RIE will be more useful if there is demand for it. Demand emerges when the RIE has one or several explicit functions. The existing literature distinguishes different functions that can be fulfilled by evaluations, including RIEs. Like any evaluation, RIEs can serve the functions of learning, accountability and legitimacy, with all three building on knowledge generation. Learning can concern different stakeholders on different levels (e.g. partners, project team, funder) and serve different specific purposes (e.g. project steering, future project design, strategic policy adjustments). The implementation of an RIE can also be useful if it serves the function of accountability vis-à-vis donors or taxpayers. Finally, RIE as part of a comprehensive evaluation system can contribute to the legitimacy of the policy field as a whole.

U2. Characteristics of the intervention

An RIE will be more useful if the intervention offers high potential for learning or accountability. Project characteristics can serve as indicators to help decide whether this potential exists.

- **Pilot projects** intended to be scaled-up in the future, which are often characterised by **innovative** approaches, can warrant an RIE. In this scenario, an RIE allows the generation of insights into the impact of a new approach on a small scale before replicating the approach and **scaling it up**.
- If an intervention is **untested** in the sense that little is yet known about its effectiveness or the evidence is inconclusive, an RIE is useful. The RIE fills an existing **evidence gap** and can contribute to the **global RIE evidence base**.
- Alternatively, if a project or intervention follows a standardised approach that is **frequently applied** (and may hence allow for synthesis across contexts) or it is **financially large** but yet untested, an RIE can support both learning (in improving the approach) and accountability.

U3. Strategic relevance

An RIE will be more useful if it is aligned with strategic goals of the implementing organisation and its funders.

- An RIE is strategically relevant if the rigorous testing of interventions aligns with an overall organisational strategy of **efficacy and evidence-orientation**.
- An RIE is strategically relevant if it promotes the **visibility** of the organisation vis-à-vis peers or donors (e.g. with the potential of opening up future financing opportunities).

U4. Economic considerations

Finally, an RIE will be more useful if economic considerations support its implementation.

- It can be beneficial to conduct an RIE if it has the potential to increase **cost-effectiveness** by comparing different treatment arms (BMZ, 2011; Gertler et al., 2016; USAID, 2013). This can either follow the goal of a) optimising the design of an intervention to achieve greater impact under given resources (= maximum principle) or b) achieving unchanged impact with fewer resources (= minimum principle).
- It will be useful to implement an RIE if the relation of **costs and benefits**⁴ of the evaluation is deemed appropriate. This implies that there is already an understanding of the intended benefit of the RIE, including its value as global public good. Ideally, the benefit of an RIE (for instance: benefit from learning, benefit for target group or partners) should outweigh the costs (e.g. financial, personnel, time, potential loss in flexibility of project adaptation).

⁴ “Cost-benefit analysis estimates the total expected benefits of a programme, compared to its total expected costs. Cost-effectiveness analysis compares the relative cost of two or more programmes or programme alternatives in reaching a common outcome.” (Gertler et al., 2016, p. 18)

REFERENCES

- BMZ (2011)**, “BMZ evaluation methods”, https://www.bmz.de/en/what_we_do/approaches/evaluation/Evaluation/methods/index.html (accessed 17 August 2021).
- Davies, R. (2013)**, “Planning evaluability assessments. A synthesis of the literature with recommendations”, *Working Paper*, No. 40, Department for International Development, Cambridge.
- DFG (2021)**, “FAQ: Informationen aus den Geistes- und Sozialwissenschaften”, https://www.dfg.de/foerderung/faq/geistes_sozialwissenschaften/index.html (accessed 17 August 2021).
- Funk, E. et al. (2018)**, “Lessons learnt from impact-oriented accompanying research: Potentials and limitations to rigorously assessing the impact of governance programmes”, *Discussion Paper*, No. 28, German Development Institute/Deutsches Institut für Entwicklungspolitik (DIE), Bonn.
- Gault, S. (2020)**, “Ethics of randomized evaluations”, Presentation, J-PAL Global, https://www.povertyactionlab.org/sites/default/files/IRC_ethics_lecture_0.pdf (accessed 17 August 2021).
- Gertler, P. J. et al. (2016)**, *Impact Evaluation in Practice*. World Bank, Washington, D.C.
- Gibson, M. and A. Sautmann (2021)**, “Introduction to randomized evaluations”, The Abdul Latif Jameel Poverty Action Lab (J-PAL), <https://www.povertyactionlab.org/resource/introduction-randomized-evaluations> (accessed 20 October 2021).
- Glennerster, R. and K. Takavarasha (2013)**, *Running Randomized Evaluations: A Practical Guide*, Princeton University Press, Princeton, doi:10.2307/j.ctt4cgd52.
- Kaplan, L. et al. (2019)**, “Ethics in development research: ‘Doing no harm’ when conducting research in low and middle-income countries”, Presentation, German Development Institute/Deutsches Institut für Entwicklungspolitik (DIE), https://www.die-gdi.de/fileadmin/user_upload/pdfs/projekt/project_ethics_in_research_introduction.pdf (accessed 17 August 2021).
- MCC (2021)**, “Millennium Challenge Corporation – Evaluation”, Millennium Challenge Corporation, <https://www.mcc.gov/our-impact/independent-evaluations> (accessed 19 October 2021).
- Noltze, M., G. Leppert and S. Harten (2018)**, “Impact assessment in complex evaluations”, *Rural21*, Vol. 52/1, p. 3.
- OECD (2006)**, “Outline of principles of impact evaluation”, Presentation, International Workshop on Impact Evaluation for Development, 15 November 2006, Hosted by the World Bank and the DAC Evaluation Network – OECD, OECD, 2006, p. 9.
- OECD (2009)**, *Glossary of Key Terms in Evaluation and Results Based Management*, Organisation for Economic Co-operation and Development (OECD), Paris.
- Peersman, G., I. Gujit and T. Pasanen (2015)**, *Evaluability Assessment for Impact Evaluation*, Method Lab Publication, Overseas Development Institute (ODI), London.
- von Schiller, A. (2021)**, *Applying Rigorous Impact Evaluation in GIZ Governance Programmes: Results of a GIZ Initiative on Impacts in Governance*, German Development Institute/Deutsches Institut für Entwicklungspolitik (DIE), Bonn and Deutsche Gesellschaft für internationale Zusammenarbeit (GIZ), Eschborn.
- UNEG (2020)**, *Ethical Guidelines for Evaluation*, United Nations Evaluation Group, <http://www.unevaluation.org/document/detail/2866> (accessed 18 October 2021).

USAID (2008), “Planning for cost effective evaluation with evaluability assessment”, *Impact Assessment Primer Series*, No. 6, United States Agency for International Development (USAID), Washington, D.C.

USAID (2013), “Technical note: Impact evaluations”, *Monitoring and Evaluation Series, Technical Guide*, US Agency for International Development (USAID), Washington, D.C.

White, H. and E. Masset (2018), “The rise of impact evaluations and challenges which CEDIL is to address”, *Journal of Development Effectiveness*, Routledge, Vol. 10/4, pp. 393–399.

White, H. and D. Raitzer (2017), *Impact Evaluation of Development Interventions. A practical guide*, Asian Development Bank (ADB), Manila.