FOUNDATIONS OF IMPACT EVALUATION

Impact evaluation is a methodology designed to determine whether a program or intervention actually *caused* observed changes in outcomes. It goes beyond measuring "what happened" to answer: *Did the intervention make the difference? How? And why?*

This brief introduces the core concepts of impact evaluation, including types of evidence, selection bias, randomized controlled trials (RCTs), and theory-based evaluation (TBIE). It features practical examples from the education sector, with tools and strategies for using evidence to make smarter decisions, and design better programs.

Designed for practitioners, policymakers, and researchers, this guide offers a **practical and rigorous approach to understanding what works—and why—in education and social development.**

This brief is part of a series of five brief guides produced by the Queen Rania Foundation, based on materials developed by Prof. Howard White (presenting the Research and Evaluation Center (REC)) for the training course titled "Impact Evaluation: Methods, Advocacy, and Scalability". The training was funded by the Education Endowment Foundation (EEF), and the BHP Foundation, and the Queen Rania Foundation.

Why Does Evidence Matter?

Evidence-based decision-making has gained prominence across all sectors, including health, education, and social policy. Reliance on evidence is most advanced in the health sector. Clinical trials are a prerequisite before approving treatments. World Health Organization (WHO) Guidelines are based on systematic reviews of studies of effectiveness.

However, in other areas such as education and social policy, many interventions remain untested. The evidence revolution promotes the use of systematic research in designing policies and programs, drawing from different types of evidence including primary studies, literature reviews, and systematic reviews.

Research evidence is categorized into primary and secondary research:

- **Primary research** involves direct data collection through fieldwork, utilizing either qualitative or quantitative methods.

- **Secondary research** uses existing data from reports and studies.

Different Research Types and What They Tell Us?

Social science research can be categorized into several key areas, each offering insights for decision-making:

- **Prevalence Studies** : These studies measure how common a condition, behavior, or issue is within a population at a specific time. For example, a study might examine how many people in a city experience food insecurity. Prevalence studies help identify priorities and inform advocacy work by showing the scale of the programme.

- **Risk Factor Analysis** : This research identifies factors that increase the likelihood of a specific outcome. For instance, a study might explore how poverty and education levels increase the likelihood of child labor. Risk factor analysis can help identify target groups for an intervention.

- **Social drivers and risk factors**: Research in this area examines the underlying causes or drivers for social phenomenon, such as inequalities in education access and outcomes. Understanding these drives can inform intervention selection.

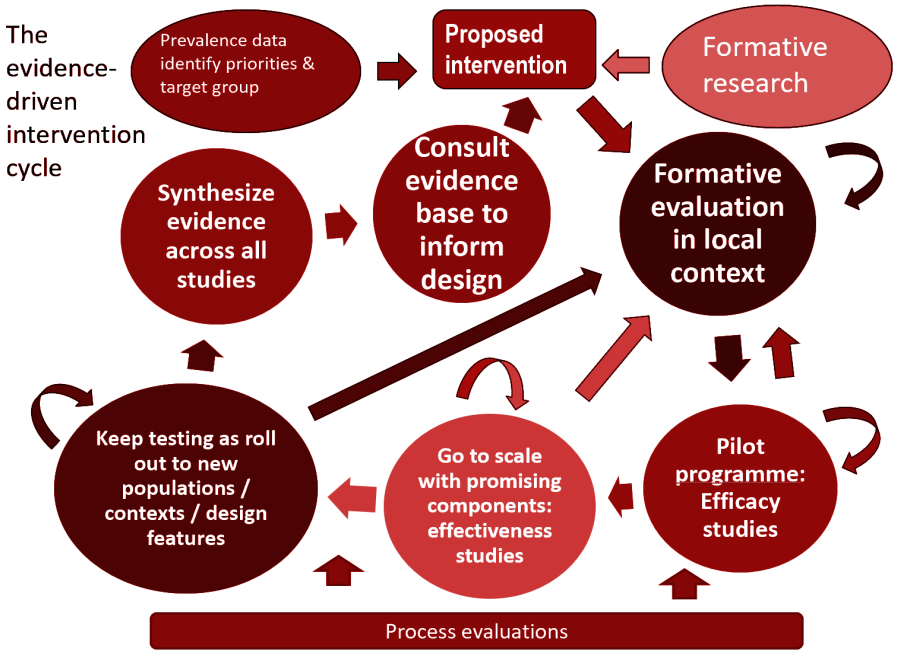
- **Intervention Research:** This type of research largely comprises evaluations, both impact evaluation and process evaluation. Impact evaluations test the difference an intervention makes. For example, a study might evaluate whether disaster preparedness training improves community response to floods. Process evaluations examine the implementation of the intervention, not its outcomes.

- **Consequences**: Research to identify the consequences of a social phenomenon. For example, how child labor affects the health and life chances of a child. Understanding consequences can be useful for advocacy, or inform choice of interventions to ameliorate adverse consequences.

- **Formative Research:** Designing effective interventions requires more than good intentions—it requires an understanding of the context. This is where formative research plays a vital role. It helps:
    - Assess the current services available
    - Identify likely barriers to implementation
    - Understand the interests and capacities of the target population

This research is foundational when determining what kind of intervention is both needed and feasible.

Fig.1: The Role of Formative Research In Intervention Design

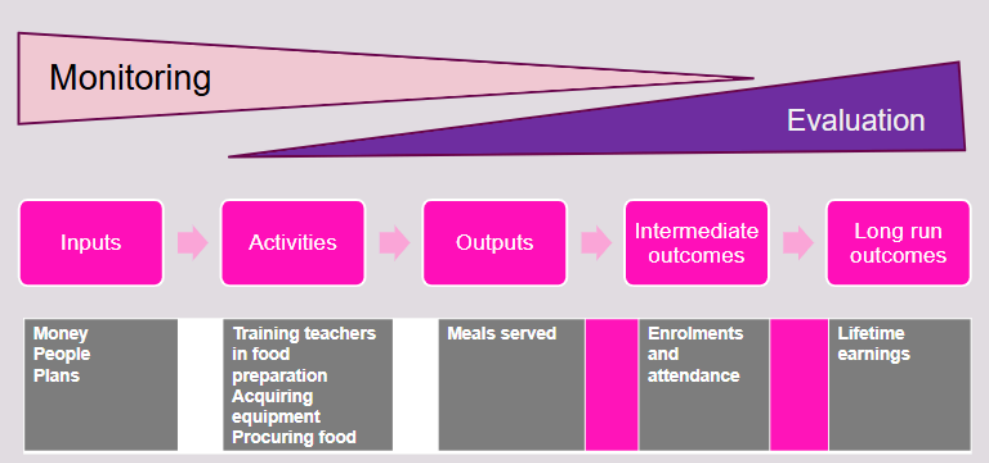


Types of Evaluation and the Role of Impact Evaluation

Monitoring and evaluation are instrumental functions for the delivery of effective interventions. Monitoring tracks progress, usually against predetermined indicators. Evaluations assess interventions using explicit criteria. Impact evaluations try to show whether the intervention caused the observed outcomes.

Figure 2 shows an example of a simple log frame for a school feeding programme and the relative roles of monitoring and evaluation. The log frame goes from inputs to long run outcomes. The bottom row shows selected indicators at each stage.

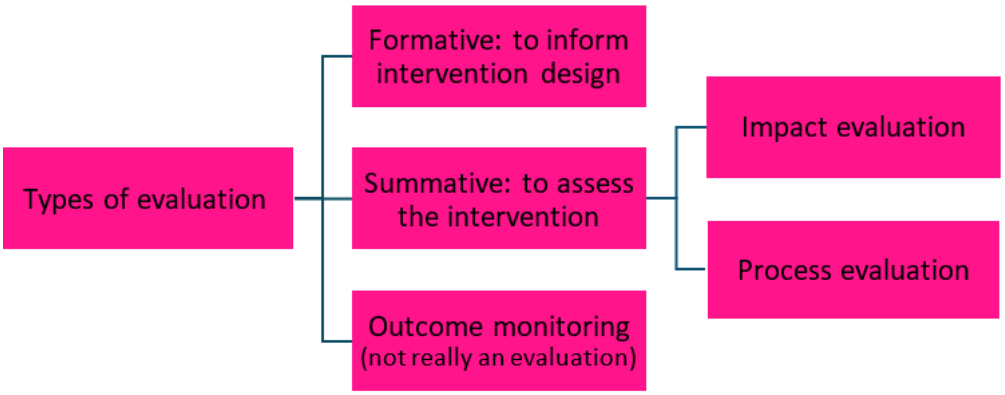
Fig. 2: Example logframe for a school nutrition programme



There are various types of evaluations (Figure 3), including:

- **Formative evaluation**: Conducted in the early stages of implementing an intervention to test feasibility and acceptability.
- **Process evaluation:** Examines the implementation process, assessing whether the intervention is executed as planned and identifying barriers to successful implementation.
- **Outcome evaluation:** Measures changes in indicators but does not establish causality.
- **Impact evaluation:** Utilizes counterfactual analysis to determine whether observed changes are attributable to the intervention.

Fig.3: Types of Evaluations



Impact evaluations should employ mixed methods, incorporating both qualitative and quantitative approaches to enhance understanding. While factual analysis describes implementation and participation patterns, counterfactual analysis establishes causality by comparing intervention and non-intervention groups. The need for impact evaluations arises from the limitations of outcome monitoring, which fails to capture the actual contribution of an intervention to observed changes in outcomes.

Quantitative methods provide precise numerical evidence of whether an impact occurred.
Qualitative methods, on the other hand, help explain the underlying mechanisms and context—they reveal why an intervention worked (or didn't).
Combining both approaches—mixed methods—gives us a more complete picture: Did an impact occur? Why did it happen? And can it be replicated in other settings?

Table.1: Summary of factual vs counterfactual analysis

	Qualitative	Quantitative
Factual	Analysis of participatory processes, barriers to adoption, power relations	Analysis of targeting, adoption of new approaches, construction quality, skills acquisition
Counterfactual	Small sample (n) analysis impact analysis of causality at all levels of the causal chain	Large sample impact analysis of causality, usually intermediate and final outcomes

The Impact Evaluation Problem and Selection Bias

One of the primary challenges in impact evaluation is the issue of selection bias. Selection bias occurs when the individuals or communities participating in an intervention are systematically different from those who do not participate. This can arise due to:

- Program placement: Targeting specific groups based on need or performance.
- Self-selection: Individuals who choose to participate may already be more motivated or better off and so have better outcomes regardless of the intervention.

Box 1: Examples of Selection Bias

Organisation XYZ is implementing a school meal program to increase education outcomes such as literacy, enrolment, and school retention. The table shows possible sources of selection bias.

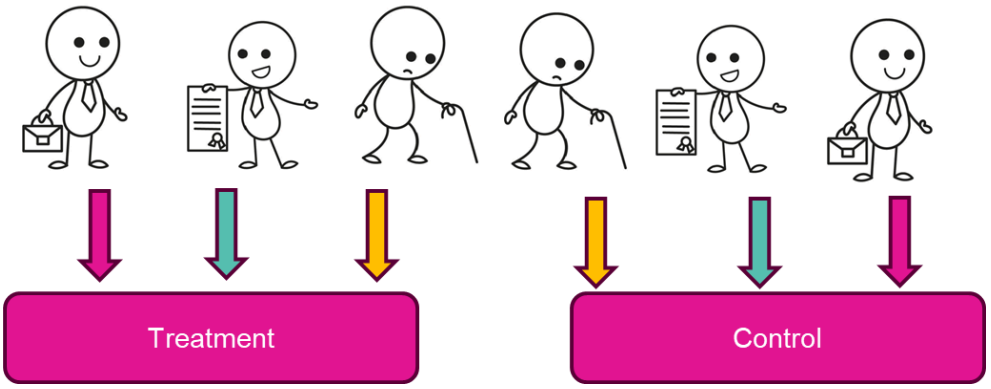
Source of bias	Example
Program placement	If the program is implemented in schools or communities with higher pre-existing levels of poverty or malnutrition, students in these areas might have lower baseline educational outcomes compared to other regions. These adverse differences may remain even after the intervention.
Self-selection	Some families are by default more motivated to improve their children's education, therefore might be more likely to send their children to schools participating in the program. These households may already place a higher value on education, skewing results as such children may perform better regardless of the program.
Attrition	Children in non-program schools might drop out at higher rates, particularly in food-insecure areas. Comparisons between participants and non-participants may then reflect attrition patterns rather than the program's true impact.

Selection bias distorts the estimated effects of an intervention, leading to inaccurate conclusions. Overestimating or underestimating the impact can result in inefficient resource allocation and misinformed policy decisions. Addressing selection bias requires the use of robust evaluation designs that create comparable treatment and control groups. One of the most effective methods for mitigating selection bias is the use of Randomized Controlled Trials (RCTs).

Introduction to Randomized Controlled Trials (RCTs)

RCTs randomly assign eligible participants into treatment and control so that differences in outcomes are likely due to the intervention itself, and not external factors. By eliminating selection bias, RCTs provide the most reliable evidence on intervention effectiveness. Randomization should not be confused with random sampling. See box 2.

Box 2: Random assignment should not be confused with random sampling. Random sampling refers to how a sample is drawn from one or more populations. Random assignment refers to how individuals or groups are assigned to either a treatment group or a control group. RCTs typically use both random sampling (since they are usually aiming to make inferences about a larger population) and random assignment (an essential characteristic of an RCT).



Why is randomisation important?

Randomization ensures baseline balance between the treatment and control groups, meaning that observable and unobservable characteristics are equally distributed. This means that with any post-intervention differences in outcomes measure the impact of the intervention.

However, conducting RCTs requires careful planning, including defining eligibility criteria, determining sample size, and addressing ethical considerations. The feasibility of randomization depends on factors such as program design, political acceptance, and logistical constraints.

RCT Designs and Calculating Impact

There are various RCT designs used to evaluate interventions. The choice depends on the nature of the intervention and logistical considerations::

- **Simple randomization:** Assigns individuals or units randomly to treatment and control groups, ensuring comparability.

- **Cluster RCTs:** Randomize entire groups, such as schools or communities, to treatment or control, which is a logistically feasible approach to work with schools and reduces the risk of contamination between groups.

- **Stratified randomization (including matched pairs):** Ensures that treatment and control groups are balanced across key characteristics, enhancing statistical power.

- **Pipeline randomization (stepped-wedge design):** Used when interventions are rolled out gradually, allowing for phased implementation while maintaining a control group for comparison.

- **Encouragement designs:** Apply randomization to encourage participation in an intervention without restricting access, making them useful for evaluating programs with voluntary uptake.

Calculating impact in an RCT involves comparing outcomes between treatment and control groups. The difference in means between these groups represents the intervention's effect.

Despite their advantages, RCTs also present challenges. Some people raise ethical concerns, though these can be countered, and argue that RCTs are expensive, which is the case for any evaluation design requiring extensive data collection. The valid concern is ensuring compliance with randomization protocols, preventing attrition, and managing spillover effects. However, when properly implemented, RCTs generate robust evidence that can guide decision-making, making them the preferred method for impact evaluation in many fields.

RCTs in Education

In education, RCTs assess interventions such as scholarship programs, technology-assisted learning, and teacher training initiatives. By randomly assigning participants to treatment and control groups, RCTs ensure that observed differences in outcomes can be attributed to the intervention rather than external factors. Below are a few examples:

Computer-Assisted Learning (CAL) in India

- This study evaluated the impact of a technology-driven intervention implemented by Pratham in municipal schools in Vadodara, India, on student performance.
- Fourth-grade students played educational games designed to enhance their math competencies.
- Findings: Computer-Assisted Learning (CAL) increased math test scores by 0.37 standard deviations but had no effect on language skills. This shows the importance of intervention specificity in achieving desired educational outcomes.

Secondary School Scholarships in Ghana

- A large-scale RCT examined the impact of four-year scholarships on school enrollment and subsequent life outcomes.
- Findings: Scholarship recipients had significantly higher completion rates, improved cognitive skills, and better employment outcomes, particularly among female students who experienced a greater benefit in terms of career opportunities and delayed childbearing.

Vocabulary Learning and Task Types in Iran

- This study assessed the impact of different learning task types on English as a Foreign Language (EFL) students.
- Findings: Writing tasks yielded the greatest impact on both receptive and productive vocabulary acquisition, supporting the theory that deeper cognitive engagement enhances learning.

Reporting Standards for RCTs

RCTs require complete and accurate reporting to ensure credibility and replicability. Key reporting standards include:

- Unit of Assignment: The level at which randomization occurs (e.g., individual students, classrooms, or schools).
- Unit of Treatment: The entity receiving the intervention (e.g., students receiving tutoring, teachers undergoing training).
- Unit of Analysis: The primary level at which outcomes are measured (e.g., test scores at the student level, school-wide performance improvements).
- Data Collection and Attrition Rates: Addressing missing data and response bias is crucial for maintaining the validity of findings.
- Ethical Considerations: Transparency, informed consent, and data protection measures ensure compliance with ethical standards.

Theory-Based Impact Evaluation (TBIE)

Theory of Change

A Theory of Change (ToC) is a useful framework for understanding how and why an intervention achieves its intended outcomes. By mapping the causal chain from inputs to outcomes, it identifies the assumptions underpinning success and highlights potential bottlenecks. A Theory of Change helps plan and evaluate programs by showing how each step leads to results. It helps stakeholders design, implement, and evaluate programs, ensuring resources are directed toward successful interventions.

Concept

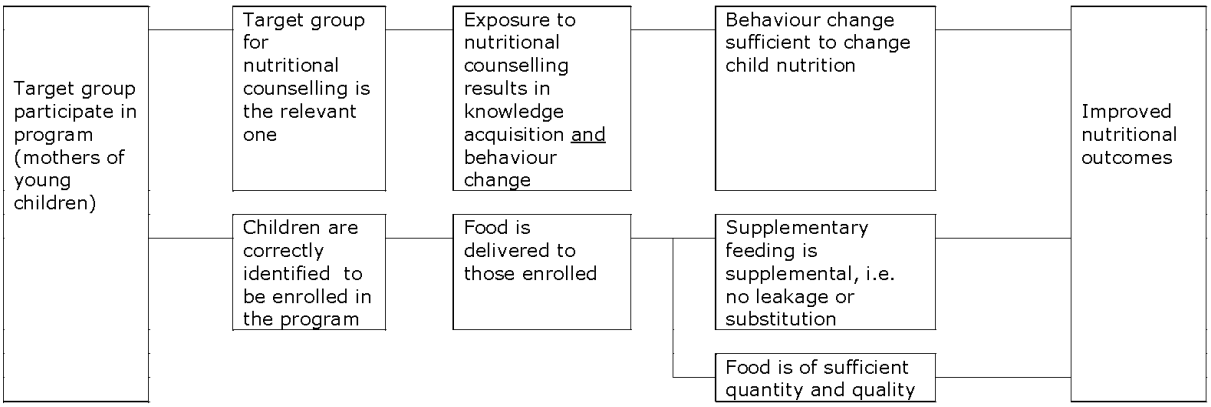
A Theory-Based Impact Evaluation (TBIE) explores the causal mechanisms of an intervention rather than just measuring its effects. It maps out the steps through which an intervention is expected to create change and tests the underlying assumptions. Unlike RCT only analyzing outcomes, TBIE provides insights into why an intervention works or fails.

Example: Bangladesh Integrated Nutrition Project (BINP)

- BINP was designed to improve child nutrition through growth monitoring, nutritional counseling, and supplementary feeding.
- Challenges Identified:
    - Knowledge gaps due to deeply ingrained traditional norms and household decision-making structures so that the child's mother was not the sole or main decision-maker on child feeding practices..

- Misidentification of malnourished children by field staff, leading to improper targeting of beneficiaries.
- Food distribution faced leakage (supplement sold or given to someone else) and substitution effects (supplements replace regular meals rather than supplementing them).

Fig. 4: Causal chain for nutrition project: nutritional counselling and supplementary feeding



Source: White (2009)

Applications of TBIE

- **Causal Chain Mapping:** Identifies weak and missing links in implementation by tracking expected vs. actual pathways of change.
- **Assumption Testing:** Evaluates whether the expected mechanisms (e.g., increased knowledge leading to behavioral change) function as intended.
- **Contextual Analysis:** Assesses external factors influencing outcomes, such as socio-economic conditions and policy environment.

Case Study: School Voucher Program

Background

A national school voucher program aimed to increase access to quality education by subsidizing private school fees for low-income families.

Evaluation Design

- Method: Randomized Control Trial (RCT)
- Sample: 10,000 students across 200 schools
- Outcomes Measured:
  - Enrollment rates
  - Student learning outcomes (test scores)
  - loyment

Findings

- Enrollment Impact: Voucher recipients had an 80% enrollment rate, compared to 50% in the control group.
- Learning Outcomes: Math test scores were 0.5 SD, and reading scores by 0.3 WHOSD, higher in the treatment group compared to the control.
- Employment Outcomes: Recipients had a 10% higher employment rate post-graduation.

Key Lessons Learned

- The program was most effective for low-income households.
- Variability in private school quality affected learning outcomes.
- Supplementary interventions, such as teacher training, could enhance program success.

Further Reading

RCTs and Quantitative Causal Inference

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

White, H., Sabarwal S. & de Hoop, T. (2014). Randomized Controlled Trials (RCTs), Methodological Briefs: Impact Evaluation 7, UNICEF Office of Research, Florence.

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

Angrist, J.D. (2014). Mastering 'Metrics: The Path from Cause to Effect. Princeton University Press.

Cunningham, S. (2021). Causal Inference: The Mixtape. Yale University Press.

Heckman, J. J. (1979). Selection Bias as a Specification Error. Econometrica, 47(1), 153–161.

Mixed Methods and Theory-Based Evaluation

White, H. & Yang, T. (2023). Mixed methods in education RCTs. In R. Tierney, F. Rizvi, & K. Ercikan (Eds.), International Encyclopedia of Education (4th ed., pp. 599–607). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.11060-7

White, H. (2009). Theory-Based Impact Evaluation: Principles and Practice. Working Paper 3, 3ie. https://www.3ieimpact.org/sites/default/files/2017-11/Working_Paper_3.pdf

White, H. (2013). Using a causal chain to make sense of the numbers. https://www.3ieimpact.org/blogs/using-causal-chain-make-sense-numbers

White, H. (2017). How to build a theory of change for an impact evaluation [Video]. https://www.youtube.com/watch?v=pWutrZwzP18

Evaluation Systems and Evidence Use

White, H. (2019). The twenty-first century experimenting society: the four waves of the evidence revolution. Palgrave Communications, 5(1), 1–7. https://www.nature.com/articles/s41599-019-0253-6

White, H. (2020). The global evidence architecture in health and education: A comparative scorecard. In Getting Evidence into Education (pp. 20–33). Routledge. https://www.researchgate.net/profile/Howard-White/publication/340700178

WHO (2014). WHO Handbook for Guideline Development (2nd ed.). https://www.who.int/publications/i/item/9789241548960

Evaluation Theory and Practice

Scriven, M. (1991). Evaluation Thesaurus. Sage Publications.

White, H. (2023). Ten Common Flaws in Evaluation. https://www.gdn.int/ten-common-flaws-evaluations