unicef
Office of Research-Innocenti

Methodological Briefs
Impact Evaluation No. 10

Overview: Data Collection and Analysis Methods in Impact Evaluation

Greet Peersman

RMIT UNIVERSITY
3ie International Initiative for Impact Evaluation
BetterEvaluation
unicef Office of Research-Innocenti

UNICEF OFFICE OF RESEARCH

The Office of Research is UNICEF's dedicated research arm. Its prime objectives are to improve international understanding of issues relating to children's rights and to help facilitate full implementation of the Convention on the Rights of the Child across the world. The Office of Research aims to set out a comprehensive framework for research and knowledge within the organization, in support of UNICEF's global programmes and policies, and works with partners to make policies for children evidence-based. Publications produced by the Office are contributions to a global debate on children and child rights issues and include a wide range of opinions.

The views expressed are those of the authors and/or editors and are published in order to stimulate further dialogue on impact evaluation methods. They do not necessarily reflect the policies or views of UNICEF.

OFFICE OF RESEARCH METHODOLOGICAL BRIEFS

UNICEF Office of Research Methodological Briefs are intended to share contemporary research practice, methods, designs, and recommendations from renowned researchers and evaluators. The primary audience is UNICEF staff who conduct, commission or interpret research and evaluation findings to make decisions about programming, policy and advocacy.

This brief has undergone an internal peer review.

The text has not been edited to official publication standards and UNICEF accepts no responsibility for errors.

To consult and download the Methodological Briefs, please visit http://www.unicef-irc.org/KM/IE/

For readers wishing to cite this document we suggest the following form:
Peersman, G. (2014).Overview: Data Collection and Analysis Methods in Impact Evaluation, *Methodological Briefs: Impact Evaluation 10*, UNICEF Office of Research, Florence.

Acknowledgements: This brief benefited from the guidance of many individuals. The author and the Office of Research wish to thank everyone who contributed and in particular the following:

Contributors: Simon Hearn, Jessica Sinclair Taylor
Reviewers: Nikola Balvin, Claudia Cappa, Yan Mu

September 2014

UNICEF Office of Research - Innocenti
Piazza SS. Annunziata, 12
50122 Florence, Italy
Tel: (+39) 055 20 330
Fax: (+39) 055 2033 220
florence@unicef.org
www.unicef-irc.org

1. DATA COLLECTION AND ANALYSIS: A BRIEF DESCRIPTION

Well chosen and well implemented methods for data collection and analysis are essential for all types of evaluations. This brief provides an overview of the issues involved in choosing and using methods for impact evaluations – that is, evaluations that provide information about the intended and unintended long-term effects produced by programmes or policies.

Impact evaluations need to go beyond assessing the size of the effects (i.e., the average impact) to identify for whom and in what ways a programme or policy has been successful. What constitutes 'success' and how the data will be analysed and synthesized to answer the specific key evaluation questions (KEQs) must be considered up front as data collection should be geared towards the mix of evidence needed to make appropriate judgements about the programme or policy. In other words, the analytical framework – the methodology for analysing the 'meaning' of the data by looking for patterns in a systematic and transparent manner – should be specified during the evaluation planning stage. The framework includes how data analysis will address assumptions made in the programme theory of change about how the programme was thought to produce the intended results (see Brief No. 2, Theory of Change). In a true mixed methods evaluation, this includes using appropriate numerical and textual analysis methods and triangulating multiple data sources and perspectives in order to maximize the credibility of the evaluation findings

Main points

1. Data collection and analysis methods should be chosen to match the particular evaluation in terms of its key evaluation questions (KEQs) and the resources available.
2. Impact evaluations should make maximum use of existing data and then fill gaps with new data.
3. Data collection and analysis methods should be chosen to complement each other's strengths and weaknesses.

2. PLANNING DATA COLLECTION AND ANALYSIS

Begin with the overall planning for the evaluation

Before decisions are made about what data to collect and how to analyse them, the purposes of the evaluation (i.e., the intended users and uses) and the KEQs must be decided (see Brief No. 1, Overview of Impact Evaluation). An impact evaluation may be commissioned to inform decisions about making changes to a programme or policy (i.e., formative evaluation) or whether to continue, terminate, replicate or scale up a programme or policy (i.e., summative evaluation). Once the purpose of the evaluation is clear, a small number of high level KEQs (not more than 10) need to be agreed, ideally with input from key stakeholders; sometimes KEQs will have already been prescribed by an evaluation system or a previously developed evaluation framework. Answering the KEQs – however they are arrived at – should ensure that the purpose of the evaluation is fulfilled. Having an agreed set of KEQs provides direction on what data to collect, how to analyse the data and how to report on the evaluation findings.

An essential tool in impact evaluation is a well developed theory of change. This describes how the programme or policy is understood to work: it depicts a causal model that links inputs and activities with

outputs and desired outcomes and impacts (see Brief No. 2, Theory of Change). The theory of change should also take into account any unintended (positive or negative) results. This tool is not only helpful at the programme design stage but it also helps to focus the impact evaluation on what stakeholders need to know about the programme or policy to support decision making – in other words, the KEQs. Good evaluation questions are not just about 'What were the results?' (i.e., descriptive questions) but also '*How good* were the results?' (i.e., judging the value of the programme or policy). Impact evaluations need to gather evidence of impacts (e.g., positive changes in under-five mortality rates) and also examine how the intended impacts were achieved or why they were not achieved. This requires data about the context (e.g., a country's normative and legal framework that affects child protection), the appropriateness and quality of programme activities or policy implementation, and a range of intermediate outcomes (e.g., uptake of immunization) as explanatory variables in the causal chain.[1]

Make maximum use of existing data

Start the data collection planning by reviewing to what extent existing data can be used**.** In terms of indicators, the evaluation should aim to draw on different types of indicators (i.e., inputs, outputs, outcomes, impacts) to reflect the key results in the programme's theory of change. Impact evaluations should ideally use the indicators that were selected for monitoring performance throughout the programme implementation period, i.e., the key performance indicators (KPIs). In many cases, it is also possible to draw on data collected through standardized population based surveys such as UNICEF's Multiple Indicator Cluster Survey (MICS), Demographic and Health Survey (DHS) or the Living Standards Measurement Study (LSMS).

It is particularly important to check whether baseline data are available for the selected indicators as well as for socio-demographic and other relevant characteristics of the study population. When the evaluation design involves comparing changes over time across different groups, baseline data can be used to determine the groups' equivalence before the programme began or to 'match' different groups (such in the case of quasi-experimental designs; see Brief No. 8, Quasi-experimental design and methods). They are also important for determining whether there has been a change over time and how large this change (i.e., the effect size). If baseline data are unavailable, additional data will need to be collected in order to reconstruct baselines, for example, through using 'recall' (i.e., asking people to recollect specific information about an event or experience that occurred in the past). While recall may be open to bias, it can be substantially reduced – both by being realistic about what people can remember and what they are less likely to recall, and by using established survey tools.[2]

Other common sources of existing data include: official statistics, programme monitoring data, programme records (which may include a description of the programme, a theory of change, minutes from relevant meetings, etc.), formal policy documents, and programme implementation plans and progress reports. While it is important to make maximum use of existing data for efficiency's sake, the data must be of sufficient quality to not compromise the validity of the evaluation findings (see more below).

Identify and address important data gaps

After reviewing currently available information, it is helpful to create an evaluation matrix (see table 1) showing which data collection and analysis methods will be used to answer each KEQ and then identify and prioritize data gaps that need to be addressed by collecting new data. This will help to confirm that the planned data collection (and collation of existing data) will cover all of the KEQs, determine if there is sufficient triangulation between different data sources and help with the design of data collection tools

[1] Brief No. 1, Overview of Impact Evaluation covers the need for different approaches to evaluating policies rather than programmes.

[2] White, Howard, 'A contribution to current debates in impact evaluation', Evaluation, 16(2), 2010, pp. 153–164.

(such as questionnaires, interview questions, data extraction tools for document review and observation tools) to ensure that they gather the necessary information.

Table 1. Evaluation matrix: Matching data collection to key evaluation questions

Examples of key evaluation questions (KEQs)	Programme participant survey	Key informant interviews	Project records	Observation of programme implementation
KEQ 1 What was the quality of implementation?		✔	✔	✔
KEQ 2 To what extent were the programme objectives met?	✔	✔	✔	
KEQ 3 What other impacts did the programme have?	✔	✔		
KEQ 4 How could the programme be improved?		✔		✔

There are many different methods for collecting data. Table 2 provides examples of possible (existing and new) data sources.[3]

Table 2. Data collection (primary data) and collation (secondary data) options

Option	What might it include?	Examples
Retrieving existing documents and data	• Formal policy documents, implementation plans and reports • Official statistics • Programme monitoring data • Programme records	• Review of programme planning documents, minutes from meetings, progress reports • The political, socio-economic and/or health profile of the country or the specific locale in which the programme was implemented
Collecting data from individuals or groups	• Interviews[4] – key informant, individual, group, focus group discussions, projective techniques • Questionnaires or surveys – email, web, face to face, mobile data	• Key informant interviews with representatives from relevant government departments, non-governmental organizations and/or the wider development community • Interviews with programme managers, programme

[3] More information on each of these and a more comprehensive list of data collection/collation options can be accessed via the 'Collect and/or Retrieve Data' web page on the BetterEvaluation website, at http://betterevaluation.org/plan/describe/collect_retrieve_data.

[4] See Brief No. 12, Interviewing.

	• Specialized methods (e.g., dotmocracy, hierarchical card sorting, seasonal calendars, projective techniques, stories)[5]	implementers and those responsible for routine programme monitoring • Interviews, group discussions (such as focus groups) and/or questionnaires with programme participants
Observation	• Structured or non-structured • Participant or non-participant • Participatory or non-participatory • Recorded through notes, photos or video	• Observations of programme activities and interactions with participants
Physical measurement	• Biophysical measurements • Geographical information	• Infant weight • Locations with high prevalence of HIV infection

Use a range of data collection and analysis methods

Although many impact evaluations use a variety of methods, what distinguishes a 'mixed methods evaluation' is the systematic integration of quantitative and qualitative methodologies and methods at all stages of an evaluation.[6] A key reason for mixing methods is that it helps to overcome the weaknesses inherent in each method when used alone. It also increases the credibility of evaluation findings when information from different data sources converges (i.e., they are consistent about the direction of the findings) and can deepen the understanding of the programme/policy, its effects and context.[7]

Decisions around using a mixed methods approach involve determining:

- at what stage of the evaluation to mix methods (the design is considered much stronger if mixed methods are integrated into several or all stages of the evaluation)
- whether methods will be used sequentially (the data from one source inform the collection of data from another source) or concurrently (triangulation is used to compare information from different independent sources)
- whether qualitative and quantitative methods will be given relatively equal weighting or not
- whether the design will be single level (e.g., the household) or multi-level (e.g., a national programme that requires description and analysis of links between different levels).

The particular analytic framework and the choice of specific data analysis methods will depend on the purpose of the impact evaluation and the type of KEQs that are intrinsically linked to this:

[5] Dotmocracy: collects levels of agreement on written statements among a large number of people. Hierarchical card sorting: provides insight into how people categorize and rank different phenomena. Seasonal calendars: visualize patterns of variations over particular periods of time. Projective techniques: provide a prompt for interviews (e.g., using photolanguage, participants select one or two pictures from a set and use them to illustrate their comments about something). Stories: as personal stories to provide insight into how people experience life.

[6] Bamberger, Michael, 'Introduction to Mixed Methods in Impact Evaluation', Guidance Note No. 3, InterAction, Washington, D.C., August 2012. See http://www.interaction.org/impact-evaluation-notes.

[7] Ibid.

- **Descriptive questions** require data analysis methods that involve both quantitative data and qualitative data.
- **Causal questions** require a research design to address attribution (i.e., whether or not observed changes are due to the intervention or external factors) and contribution (to what extent the intervention caused the observed changes; see Brief No. 6, Strategies for Causal Attribution).
- **Evaluative questions** require strategies for synthesis that apply the evaluative criteria to the data to answer the KEQs (see Brief No. 3, Evaluative Criteria). Defining up front what constitutes 'success' by constructing specific evaluative rubrics (i.e., standards or levels of performance of the programme or policy) provides a basis on which the collected information can be systematically combined to make evidence based and transparent judgements about the value of the programme or policy (also called 'evaluative reasoning', see Brief No. 4, Evaluative Reasoning).

While an impact evaluation aims to look at the longer-term results of a programme or policy, decision makers often need more timely information and therefore data on shorter-term outcomes should also be collected. For example, it is well known that the results of interventions in education emerge only over a protracted period of time. In the case of the child-friendly schools initiative in Moldova, its evaluation captured the short-term results (such as "increased involvement of students in learning through interactive and participatory teaching methods"[8]) measured during the intervention or shortly after its completion and assumed these to be predictive of the longer-term effects.

Simply determining that change has occurred – by measuring key indicators – does not tell you why it has occurred, however. Information is also needed on specific activities that were implemented, and on the context in which they were implemented. As noted above, having an explicit theory of change for the programme or policy is an essential tool for identifying which measures should be collected, and it also provides direction on which aspects of the programme implementation – and its context – data collection should focus on. By specifying the data analysis framework up front, the specific needs for data collection (primary or new data to be collected) and data collation (secondary or existing data) are clearly incorporated in a way that also shows how data will be analysed to answer the KEQs and make evaluative judgements. The data needs and the data collection and analysis methods linked to each of the KEQs should be described in the evaluation plan alongside specifics about how, where, when and from whom data will be collected – with reference to the strategy for sampling the study population, sites and/or time periods.

Ensure selected data collection and analysis methods are feasible

Once the planning is complete, it is important to check the feasibility of the data collection methods and analysis to ensure that what is proposed can actually be accomplished within the limits of the evaluation time frame and resources. For example, key informants may be unavailable to meet at the time that data are required. It is also important to analyse the equipment and skills that will be needed to use these methods, and assess whether these are available or can be obtained or developed. For example, collecting questionnaire data by mobile phone will require that either every data collector has a mobile phone or that there is a reliable system for sharing mobile phones among the data collectors. Any major gaps between what is available and what is required should be addressed by acquiring additional resources or, more realistically, by adapting the methods in line with the available resources.

Given that not everything can be anticipated in advance, and that certain conditions may change during the course of the evaluation, choices may have to be revisited and the evaluation plan revised accordingly. In

[8] Velea, Simona, and CReDO (Human Rights Resource Centre), *Child-Friendly Schools, External Evaluation Report of the Child-Friendly School Initiative (2007–2011), Republic of Moldova*, Ministry of Education of the Republic of Moldova/UNICEF, 2012. See http://www.unicef.org/moldova/CFS_EN_PRINT.pdf.

such cases, it is important to document what has changed and why, and consider and document any implications that these changes may have on the evaluation product and its use.

3. ENSURING GOOD DATA MANAGEMENT

Good data management includes developing effective processes for: consistently collecting and recording data, storing data securely, cleaning data, transferring data (e.g., between different types of software used for analysis), effectively presenting data and making data accessible for verification and use by others.

Commonly referred to aspects of data quality are:

- **Validity**: Data measure what they are intended to measure.
- **Reliability**: Data are measured and collected consistently according to standard definitions and methodologies; the results are the same when measurements are repeated.
- **Completeness**: All data elements are included (as per the definitions and methodologies specified).
- **Precision**: Data have sufficient detail.[9]
- **Integrity**: Data are protected from deliberate bias or manipulation for political or personal reasons.
- **Timeliness**: Data are up to date (current) and information is available on time.

It is advisable to use standardized data collection tools, which have already been tried and tested in real life situations, and improve these if necessary to maximize data quality. Where adaptations to the local context are necessary, or when data collection tools need to be developed, it is important to conduct a pilot test first (and improve the tool) before using it more generally. Using experienced data collectors, providing training for data collectors on a specific task or tool and/or supervising data collection across multiple data collectors can also help to reduce bias (e.g., inappropriate prompting for answers during interviews) or errors (e.g., misunderstanding which programme elements need to be observed) in the data obtained.

Data collection is not necessarily the sole responsibility of evaluators. The benefits of 'participatory evaluation' are well documented and this approach can go beyond data collection to involve programme staff, participants and/or other stakeholders in setting the agenda for the evaluation; identifying key results and determining what constitutes 'success'; contributing to collecting the data; and analysing and interpreting the results (see Brief No. 5, Participatory Approaches).

Even when data have been collected using well defined procedures and standardized tools, they need to be checked for any inaccurate or missing data. This is known as data cleaning, and it also involves finding and dealing with any errors that occur during the writing, reading, storage, transmission or processing of computerized data.

Ensuring data quality also extends to ensuring appropriate data analysis and presentation of the data in the evaluation report so that the findings are clear and conclusions can be substantiated. This often also involves making the data accessible so that they can be verified by others and/or used for additional purposes such as for synthesizing results from different evaluations (i.e., systematic review, meta-analysis, realist review or other meta-evaluation).

[9] A more specific definition for quantitative measurements is as follows: a measurement is considered 'valid' if it is both 'accurate' and 'precise'. Accuracy is defined as the deviation from the 'true' value and precision is defined as the 'scatter'. In other words, accuracy is about how close the measurement taken is to the actual (true) value; precision is the degree to which repeated measurements under unchanged conditions show the same results.

4. SPECIFIC ISSUES IN ENSURING QUALITY IN DATA COLLECTION

As noted above, ensuring quality in data collection is *part* of good data management. In this section, a few particularly pertinent issues are touched upon in more detail.

Make the sampling strategy explicit

A particularly important issue that needs to be addressed and fully described in the impact evaluation plan and report is the sampling strategy used. Start by defining the 'population of interest' from which to draw the sample. This is the group or 'units' of interest (such as all married couples) in the geographical area of interest (such as a particular village) during the time of interest (such as registered since 1995). Sampling is the process of selecting units (e.g., individuals, groups) from the population of interest to study these units in detail with the aim of drawing conclusions about the larger population (note that the generalizability of the evaluation findings will depend on how the sampling is done). Sampling error exists within any sample. In other words, no sample will yield exactly the same information as if all people in the population of interest were included in the data collection. Hence, information collected from a sample is used to make *estimates* about the population of interest. Table 3 provides an overview of the three basic clusters of sampling options.

Table 3. Overview of sampling options with illustrative methods

Cluster of sampling options	What are some specific methods?	Risk of introducing bias
Probability: Use random or quasi-random methods to select the sample, and then use statistical generalization to draw inferences about that population	• Simple random • Stratified random • Multi-stage • Sequential	This cluster has specific rules about selection of the sampling frame, the size of the sample and managing variation within the sample
Purposive: Study information-rich cases from a given population to make analytical inferences about the population. Units are selected based on one or more predetermined characteristics and the sample size can be as small as one	• Confirming and disconfirming • Critical case • Maximum variation • Outlier • Snowball • Theory-based • Typical case	This cluster encourages transparency in case selection, triangulation, and seeking out of disconfirming evidence
Convenience: These sampling options use individuals who are available or cases as they occur	• Readily available • Volunteers	This type has the lowest credibility but requires less time, money and effort

Different ways of sampling will introduce different types of bias when assessing the results of a programme or policy. As such, it is important to clearly describe the strengths and weaknesses of the sampling strategy

used and to address these, where possible, in the data analysis and/or describe how they affect the conclusions drawn about the programme or policy tested.

While three clusters of sampling options are referred to above, 'convenience' sampling is best avoided, as it does not allow for appropriate generalization of the findings (i.e., making estimates about the total population from the findings about the sample studied). In probability sampling, each unit has the same or a known chance of being selected, which increases the likelihood that the information collected about the sample is representative of the entire population of interest, especially if conditions about sample size are met (see Brief No. 7, Randomized Controlled Trials). Probability sampling allows for 'statistical generalization' in which a certain margin of error is set (typically, 5 to 10 per cent) that defines the level of acceptable risk of being wrong about the programme's results. Purposeful sampling is not based on representation or randomness and is used for selecting those units from which you can learn a great deal about the issues that are important to the evaluation (such as 'key informants')[10]. It does allow for 'analytical generalization' (i.e., making projections about the likely transferability of findings based on a theoretical analysis of the factors producing the outcomes and the effect of context.

Address dropouts and refusals

While appropriate sampling can help to reduce 'selection bias', bias can also be introduced when there are high dropout rates (or substantial 'loss to follow-up') – many people leave the programme and are no longer available to collect data from. Similarly, a low response rate when asking people to fill in a questionnaire or a high number of people refusing to participate in the data collection effort may introduce 'measurement bias'. Ways of addressing these issues include: making a special effort to increase participation; collecting information from several sources, which can then be triangulated; and determining how those who participate are different from those who drop out or refuse to participate at all, and also how participants compare to the population of interest (and, in turn, how this affects the interpretation and generalization of results). Again, transparency and in-depth discussion are warranted.

5. SPECIFIC ISSUES IN ENSURING QUALITY IN DATA ANALYSIS

As with data collection, ensuring quality in data analysis is *part* of good data management. A few particularly pertinent issues are touched upon in more detail in this section.

Each data source has its strengths and limitations, which should be explicitly described (i.e., in terms of data quality, see below) and clearly understood to ensure that sources are used and interpreted appropriately (see section 7 below). For this reason, purposefully integrating different data collection methods and/or triangulating different data sources to answer the KEQs will overcome the weaknesses inherent in each data source when used alone. A mixed methods approach also improves the credibility of the findings when information from different data sources converges (i.e., they are consistent about the direction of the findings) and can deepen the understanding of a programme or policy, its effects and its context. While mixing methods can be done post hoc (i.e., after data have already been collected) it is advisable, and often much easier, to do this up front (i.e., during evaluation planning) so that data collection can be specifically geared towards ensuring that the right mix of evidence is available at the right time to answer the KEQs. Table 4 highlights the key purposes of combining data sources.

[10] Taylor-Powell, Ellen, *Sampling*, University of Wisconsin, Milwaukee, 1998.

Table 4. The value of mixing methods: Key purposes of combining data sources

Enriching	Using qualitative data to identify issues or obtain information about variables that cannot be obtained by quantitative approaches
Examining	Generating hypotheses from qualitative data to be tested through the quantitative data (such as identifying subgroups that should be analysed separately in the quantitative data, e.g., to investigate differential impact)
Explaining	Using qualitative data to understand unanticipated results from quantitative data
Triangulating (confirming or rejecting)	Verifying or rejecting results from quantitative data using qualitative data (or vice versa)

Analysing data in order to summarize them and look for patterns is an important part of every evaluation. Strategies for the analysis of the data and how the data will be synthesized should be decided at the evaluation design stage, appropriate to the type of KEQ and described fully in the evaluation plan.

For descriptive KEQs, a range of analysis options is available (see table 5). The options can largely be grouped into two key categories: options for quantitative data (numbers) and options for qualitative data (e.g., text).

Table 5. Options for analysing numeric and textual data

Numeric analysis Analysing numeric data such as cost, frequency or physical characteristics. Options include:
• Correlation: a statistical technique to determine how strongly two or more variables are related. • Crosstabulations: obtaining an indication of the frequency of two variables (e.g., gender and frequency of school attendance) occurring at the same time. • Data and text mining: computer-driven automated techniques that run through large amounts of text or data to find new patterns and information. • Exploratory techniques: taking a 'first look' at a data set by summarizing its main characteristics, often through the use of visual methods. • Frequency tables: arranging collected data values in ascending order of magnitude, along with their corresponding frequencies, to ensure a clearer picture of a data set. • Measures of central tendency: a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. • Measures of dispersion: a summary measure that describes how values are distributed around the centre. • Multivariate descriptive: providing simple summaries of (large amounts of) information (or data) with two or more related variables. • Non-parametric inferential: data that are flexible and do not follow a normal distribution. • Parametric inferential: carried out on data that follow certain parameters. The data will be normal (i.e., the distribution parallels the bell curve); numbers can be added, subtracted, multiplied and divided; variances are equal when comparing two or more groups; and the sample should be large and randomly selected.

- Summary statistics: providing a quick summary of data, which is particularly useful for comparing one project to another, before and afterwards.
- Time series analysis: observing well defined data items obtained through repeated measurements over time.

Textual analysis

Analysing words, either spoken or written, including questionnaire responses, interviews and documents. Options include:

- Content analysis: reducing large amounts of unstructured textual content into manageable data relevant to the (evaluation) research questions.
- Thematic coding: recording or identifying passages of text or images linked by a common theme or idea, allowing the indexation of text into categories.
- Narratives: construction of coherent narratives of the changes occurring for an individual, a community, a site or a programme or policy.
- Timelines: a list of key events, ordered chronologically.

For causal KEQs, there are essentially three broad approaches to causal attribution analysis: (1) counterfactual approaches; (2) consistency of evidence with causal relationship; and (3) ruling out alternatives (see table 6). Ideally, a combination of these approaches is used to establish causality (see Brief No. 6, Overview: Strategies for Causal Attribution).

For evaluative KEQs, specific evaluative rubrics linked to the evaluative criteria employed (such as the OECD-DAC criteria) should be applied in order to synthesize the evidence and make judgements about the worth of the programme or policy (see Brief No. 4, Evaluative Reasoning).

Table 6. Data analysis approaches for causal attribution

Counterfactual approaches	Developing an estimate of what would have happened in the absence of the programme or policy; this implies the use of a control group or comparison group.
Consistency of evidence with causal relationship	Identifying patterns that would be consistent with a causal relationship, and then seeking confirming and disconfirming evidence.
Ruling out alternatives	Identifying possible alternative causal explanations, and then seeking information to determine if these can be ruled out.

A powerful tool for both the analysis and communication of the evaluation findings is data visualization. This is the process of representing data graphically in order to identify trends and patterns that would otherwise be unclear or difficult to discern. There are many options to consider[11], and selecting which type of visualization to use will depend on the nature of the data (e.g., relational, comparative, time-based). In addition, the type of visualization used to facilitate data analysis (e.g., spikes in website traffic related to a social media campaign) may not necessarily be the best type for communication purposes. When communicating data, the visualization may need to be simplified and key areas emphasized in order to

[11] For further information on specific methods of data visualization, see 'Visualise Data', web page, BetterEvaluation, http://betterevaluation.org/plan/describe/visualize_data.

attract the attention of readers and stakeholders. The conclusions of the impact evaluation should be firmly grounded in the data and presented in a manner that is transparent and verifiable by those reading the evaluation report.

6. ETHICAL ISSUES AND PRACTICAL LIMITATIONS

Ethical issues

All evaluations should follow the United Nations Evaluation Group (UNEG) *Ethical Guidelines for Evaluation*,[12] to which UNICEF adheres. These guidelines set out both a code of conduct for evaluators as well as ethical obligations to participants in the evaluation in relation to data collection and storage (see also Brief No. 1, Overview of Impact Evaluation). The latter include: respect for dignity and diversity; rights; confidentiality; and avoidance of harm. These obligations must be addressed explicitly in the design of the evaluation and throughout the evaluation process (see the Ethical Research Involving Children website).[13]

In most cases, official approval for how the data are collected, managed and used in the evaluation will be required from an institutional review board. This is to ensure that these practices comply with the ethical obligations for all those from whom data are collected.

The Moldovan evaluation of child-friendly schools[14] provided some good examples of strategies to address ethical issues during data collection and analysis:

- During the preparatory phase, stakeholders (UNICEF and implementing partners) were consulted and the data collection tools then tested on a small group of teachers, children and parents who participated voluntarily after being informed about the testing process. This was done to identify anything that could possibly be interpreted as causing nuisance, creating awkward situations or affecting the rights of participants.
- During the data collection stage, respondents were assured that all data were provided confidentially and would be used exclusively for evaluation purposes. This was also stated in the introductory text of the questionnaires. Questionnaires did not contain any fields for filling in personal data that could lead to the identification of the respondents.
- During the focus group discussions, the names of the participants were not recorded nor were the sessions taped. Participation was based on informed consent: participants received information about the evaluation approach, their role in it and the benefits to them personally, both directly (as a learning experience) and indirectly (by contributing to a better understanding of the educational system).
- During the interviews, the interactions among the participants themselves as well as between the participants and the evaluators were based on mutual respect and trust.

The same evaluation also provides good examples of safeguards to ensure the confidentiality of data:

- During data processing, no precise references were made to particular schools or to statements made by particular students, teachers or parents.

12 United Nations Evaluation Group, *Ethical Guidelines for Evaluation*, UNEG, New York, 2007.

13 The websitewww.childethics.com is a repository of evidence-based information, resources and links to journal articles collated by the Ethical Research Involving Children project to guide and improve research involving children and to provide a platform for further critical reflection and dialogue.

14 Velea, Simona, and CReDO (Human Rights Resource Centre), *Child-Friendly Schools, External Evaluation Report of the Child-Friendly School Initiative (2007–2011), Republic of Moldova*, Ministry of Education of the Republic of Moldova/UNICEF, 2012.

- Data from the questionnaires were statistically processed and the qualitative data reviewed as a whole, in the absence of any personal information. (Specific comments about specific, identified schools were reported, however.)

Practical limitations

Common practical limitations in data collection may include:

- Relevant key informants, participants and field sites may be unavailable/unreachable if
  - there is no comprehensive information about programme sites and/or participants
  - they are not contacted early enough before or during the data collection period
  - the evaluation takes place after the programme has already been completed.
- Existing relevant data sources (such as indicators for progress reporting by implementing partners; survey data) are limited in time (e.g., no baseline, no time series available) and/or content (e.g., important indicators are missing) and/or quality (i.e., there are significant data quality concerns).
- There is often limited information about the quality of services provided by different implementers in the case of a large-scale programme.

If the resources and time frame for the evaluation allow, some of these limitations in data collection may be rectified (e.g., extending the data collection period or scope) but it is likely that conclusions will have to be drawn based on imperfect data. Hence, it is important to be transparent about the limitations and describe how these may have affected the findings, conclusions and recommendations of the evaluation.

7. EXAMPLE OF GOOD PRACTICES

The evaluation of the child-friendly schools[15] referred to above provided some good examples of well developed thinking around and execution of data collection and analysis options.

Identifying data needs through appropriate unpacking of key concepts:

Child-centredness was defined as a methodological approach that could be 'unpacked' as consisting of the following specific components:

- knowing students' learning styles, problems and aspirations
- involving and valuing all students, regardless of ethnicity, social status, disability, religion or gender
- shifting the focus from teaching to learning and making the entire teaching process child friendly
- involving the students in establishing goals, learning new study methods and in performance evaluation (establishment and communication of criteria) and self-assessment
- ensuring that students are taught how to learn, and that what they learn has relevance to their daily lives, integrating into the classroom the positive learning experiences acquired by students outside of school
- diversification of cases of learning situations and teachers' roles: conveying information, facilitating their own search for information, guidance, moderation

[15] Velea, Simona, and CReDO (Human Rights Resource Centre), *Child-Friendly Schools, External Evaluation Report of the Child-Friendly School Initiative (2007–2011), Republic of Moldova*, Ministry of Education of the Republic of Moldova/UNICEF, 2012, pp. 47–48.

- cultivating good communication with parents, and encouraging parental involvement in their child's education.

The unpacking of this central concept then led the evaluators to identify the need for a range of data and, with this, a range of data sources. Questionnaires were one of the ways in which relevant data were collected. The complexity of child-centredness did not allow for the use of just a few simple indicators to assess achievements; instead, this complexity necessitated the bringing together of several measurements (obtained through a range of specific questions in the data collection instrument; see footnotes in table 7, below, reproduced from the evaluation report) linked to key result areas (such as favourable school conditions for equal treatment; see items listed in table 7).

Table 7. Students' opinions on various teaching-related indicators. (Averages; column comparison, lower score = better situation of the given indicator.) Reproduced from the evaluation report on Child-Friendly Schools in Moldova, Table 8, p. 47.

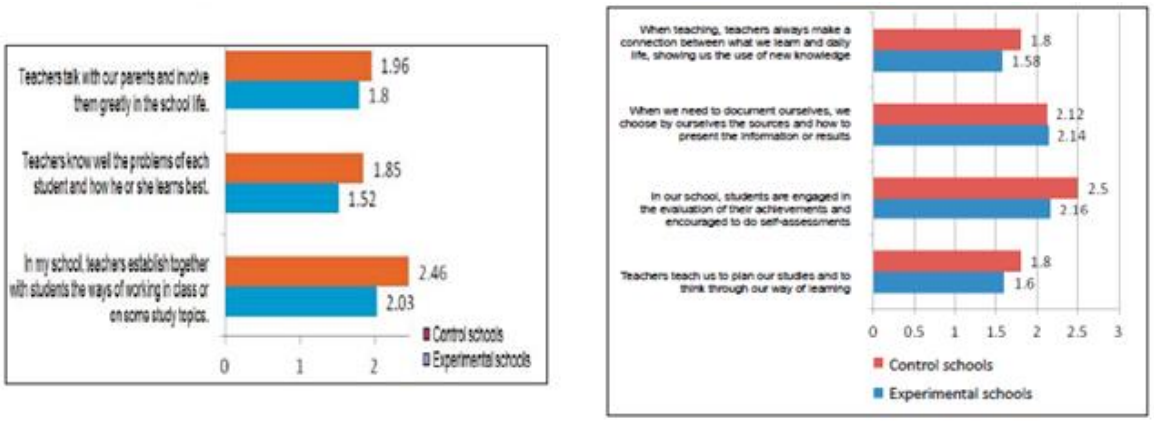
	Items	Experimental schools	Control schools
1.	Encouraging thinking in students, and self-assessment of their own learning styles[31]	3.77	4.33
2.	Setting hours and homework assignments together with the students[32]	15.48	16.89
3.	Working with children who have disabilities, and those with learning difficulties [33]	7.74	8.51
4.	Working with children of ethnic minorities [34]	6.7	7.33
5.	Treating other teachers and all students with equal respect [35]	2.79	3.05
6.	Creating school conditions favourable to equal treatment [36]	6.62	7.3
7.	Encouraging communication and dialogue with other teachers and with parents and students [37]	25.02	26.24
8.	Management of knowledge (correct learning, use of information technologies, etc.). [38]	5.1	5.31

31 Student questionnaire, Item 1, questions 10.4, 11.4, 11.12.
32 Ibid., Item 2, questions 9.1, 9.4, 9.3, 9.7, 9.8, 11.1, 11.7, 12.3, 12.4.
33 Ibid., Item 3, questions 11.3, 10.6, 11.12, 11.17.
34 Ibid., Item 4, questions 10.4, 11.4, 11.12.
35 Ibid., Item 5, questions 10.7, 11.7.
36 Ibid., Item 6, questions: 11.9, 11.11, 11.12, 11.17.
37 Ibid., Item 7, questions: 9.1–9.9, 11.1, 11.2, 11.11.
38 Ibid., Item 8, questions: 12.5, 12.1, 11.8.

Use of different data collection methods with the intent to triangulate different perspectives, and clear presentation of results:

The evaluation not only collected data about child-centredness from teachers but it also collected data about teaching practices from those at the receiving end – the students. While this may seem an 'obvious' data source to some evaluators, 'asking programme participants' is not necessarily systematically included in all impact evaluations; where it is included, it is often treated in a 'tokenistic' manner (see also Brief No. 5, Participatory Approaches). The child-friendly schools study presented complex data in a very accessible manner through good data visualization (see figures 15 and 16 below, again reproduced from the evaluation report).

Figure 1. Students' opinions regarding teaching practices relevant to child-centredness (figure on left) and students' opinions regarding teaching practices (figure on right). (Mean ratings; 0 = to the greatest extent, 3 = to the smallest extent). Reproduced from the evaluation report on Child-Friendly Schools in Moldova, Figure 15 (left) and Figure 16 (right), p. 48.



Assessing differential impacts of the programme

In addition to reporting on 'averages', the evaluators also explored differential impacts of the programme. These are essential to providing a better understanding of what worked for whom under which conditions.

For example, the evaluators pointed to positive results in student behaviour as reflected by greater and more frequent participation in school activities, especially for children with disabilities or those from vulnerable groups (as defined in the study). Also, more children obtained better academic results, with notable progress made by low achievers (as defined in the study).

8. EXAMPLES OF CHALLENGES

Common challenges in data collection and analysis can relate to *poor choices of methods* as well as *poor implementation of methods*. Some examples are provided below.

Poor choices of methods include:

1. Choosing the 'same old' methods regardless of their suitability to the specific evaluation.

   As is clear from this brief, there is a range of data collection and analysis methods available for use in impact evaluations. The most appropriate combination of methods depends on the nature of what is being evaluated and the context of the evaluation. In other words, what are the evaluation questions that need to be answered; by when must the answers be available to support decision making by different stakeholders; what resources are available for the evaluation; and what other conditions may constrain the evaluation (e.g., the political environment).

2. Choosing methods that are not specifically related to answering the KEQs.

   Constructing an evaluation matrix (see box 1, above) can help to make sure that all of the KEQs are covered.

3. Choosing just one method to answer a KEQ.

   It is not just about employing several methods, but about choosing them in such a way that they complement one another, thereby reducing bias and ensuring triangulation. That is, the strength of any single method selected will address (at least some of) the weaknesses in another selected method, and different perspectives about the intervention – in terms of key stakeholders – are included.

Poor implementation of methods includes:

1. Poor execution of sampling.

   For example, using a convenience sample or having a systematically low response rate without systematically checking and making transparent how the respondents may have been different from the non-respondents.

   The key concern in sampling is ensuring that the data from the sample can be appropriately generalized – either statistically or analytically. Significant under-representation in the selected sample may lead to spurious or incorrect conclusions about the programme or policy under investigation – for example, where there is significant under-representation in terms of certain ages, gender balance or other relevant demographics, or of certain characteristics thought to be linked to the outcomes or impact of interest (such as the exclusion of low performers in the case of looking for the impact on test pass rates).

2. Poor quality data collection.

   For example, poorly constructed questionnaires; failing to give people the opportunity to really talk during interviews; lack of good translators; and failure to record the data accurately. Ways to avoid or minimize these often common mistakes include: obtaining expert input into the design or adaptation of data collection instruments; ensuring good quality data collection through the selection, training and supervision of data collectors (including ensuring gender balance in interview teams); providing private spaces for the interviews so that people feel more able to express their views, or using technology to provide more ‘privacy’ in responses and to avoid obtaining ‘socially desirable’ responses; and using certified and experienced translators.

3. Poor quality data analysis.

   For example, drawing conclusions that are unsupported by the data, or focusing only on the average effect of the programme or policy. Ways to support a good quality data analysis may include: ensuring ‘initial’ conclusions are challenged by holding feedback sessions with key stakeholders or intentionally searching for disconfirming evidence; looking for patterns; and ensuring ‘cross-cutting’ issues such as equity are addressed.

4. Poor quality reporting and presentation.

   For example, focusing reporting according to the different data collection instruments (i.e., going through items in the questionnaire one by one) rather than answering the KEQs, or using complicated graphs that need a great deal of explanation in order to be understood. Good data visualization is an important tool for communicating key messages, for example, by creating a visual ‘map’ of key ideas for textual data and adding specific quotes in boxes to illustrate summary findings, or by showcasing key findings through carefully selected spreadsheets (i.e., there are a lot of options to choose from once you decide to go beyond the usual defaults).

9. KEY READINGS AND LINKS

Bamberger, Michael, 'Introduction to Mixed Methods in Impact Evaluation', Guidance Note No. 3, InterAction, Washington, D.C., August 2012. See http://www.interaction.org/impact-evaluation-notes.

BetterEvaluation, 'Analyze Data', web page, BetterEvaluation, http://betterevaluation.org/plan/describe/look_for_patterns.

BetterEvaluation, 'Collect and/or Retrieve Data', web page, BetterEvaluation, http://betterevaluation.org/plan/describe/collect_retrieve_data.

BetterEvaluation, 'Combine Qualitative and Quantitative Data', web page, BetterEvaluation, http://betterevaluation.org/plan/describe/combining_qualitative_and_quantitative_data.

BetterEvaluation, 'Manage Data', web page, BetterEvaluation, http://betterevaluation.org/plan/describe/manage_data.

Evergreen, Stephanie, D.H., Presenting Data Effectively, Communicating Your Findings for Maximum Impact, Sage, Thousand Oaks, 2013.

Measure Evaluation, 'Data Quality Assurance', web page, Measure Evaluation, http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/data-quality-assurance-tools/data-quality-assurance-tools.

Patton, Michael Quinn, Qualitative Research & Evaluation Methods, third edition, Sage, Thousand Oaks, 2001.

Perrin, Burt, 'Linking Monitoring and Evaluation to Impact Evaluation', Guidance Note No. 2, InterAction, Washington, D.C., April 2012. See http://www.interaction.org/document/guidance-note-2-linking-monitoring-and-evaluation-impact-evaluation.

GLOSSARY

Activity	*Actions taken or processes through which inputs are mobilized to produce specific outputs. For example, counselling sessions that adhere to quality standards.*
Baseline data	*Initially collected data which serves as a basis for comparison with data which is acquired at a later stage. For example, data collected before an intervention is implemented for later comparison with data collected after the intervention is implemented.*
Cleaning data	*Process to check data for adherence to standards, internal consistency, referential integrity, valid domain, and to replace/repair incorrect data with correct data. To "clean" a data file is to check for wild codes and inconsistent responses; to verify that the file has the correct and expected number of records, cases, and cards or records per case; and to correct errors found.*
Effect size	*A statistical term of the magnitude of the relationship between two variables.*
Impact	*Positive and negative primary and secondary long-term effects produced by the intervention, whether directly or indirectly, intended or unintended.* [16]
Indicator	*A verifiable measure that has been selected by programme or policy management to make decisions about the programme/policy. For example, the proportion of students achieving a passing grade on a standardized test.*
Input	*The financial, human and material resources used in a programme or policy. For example, training materials produced.*
Key informant	*Research/evaluation participants who are likely to be well informed about an issue, and willing to answer without bias.*
Maximum variation sampling	*Used in qualitative research to purposively selecting a wide range of cases.*
Meta-analysis	*The process or technique of synthesizing research results from two or more studies by using various statistical methods to retrieve, select, and combine results from previous studies which all address the same question.*
Meta-evaluation	*A review/assessment of the quality of an evaluation. The term is also used to describe the effort to aggregate findings from a series of evaluations in order to judge their quality and/or to assess the performance of the evaluators – this is also known as an evaluation synthesis.*
Multi-stage	*A probability sample that involves several stages (and frequently a cluster sampling stage), such as randomly selecting clusters from a population, then randomly selecting people from each of the clusters.*

[16] Organisation for Economic Co-operation and Development – Development Assistance Committee, *Glossary of Key Terms in Evaluation and Results Based Management*, OECD-DAC, Paris, 2010. See http://www.oecd.org/development/peer-reviews/2754804.pdf.

Outcome	*"Something that happens" to people, communities, the environment or the economy, which is at least partially caused by the programme or policy being evaluated. Outcomes may be positive or negative, intended or unintended, changes that happen or changes that are prevented from happening. In the context of evaluation, use of the term outcome implies that the programme/policy contributed to, helped to cause and/or catalysed (speeded up) the change.*
Outlier	*An observation in a data set which is far removed in value from the others in the data set. It is an unusually large or an unusually small value compared to the others.*
Output	*The immediate effects of programme/policy activities, or the direct products or deliverables of programme/policy activities. For example, the number of vaccines administered.*
Participatory evaluation	*An approach to evaluation which involves stakeholders, particularly the participants in a programme or those affected by a given policy, in specific aspects of the evaluation process, including design, implementing data collection and analysing data, or overseeing the work of technical experts.*
Pilot test	*A small scale study conducted to test the reliability of a data collection tool.*
Quasi-experimental design	*A research/evaluation design in which participants are **not randomly** assigned to treatment conditions, but in which comparison groups are constructed by statistical means. It differs from the (classic) controlled experiment by not having random assignment of the treatment/intervention.*
Realist review	*An explanatory and iterative type of research synthesis which tests and builds middle-range theories. It provides an exploratory analysis to answer what works, for whom, under what circumstances, why and how.*
Sampling error	*The degree to which the results from the sample deviate from those that would be obtained from the entire population, because of random error in the selection of respondent and the corresponding reduction in reliability.*
Sequential sampling	*A sampling method which can be both probability and non-probability based. As a probability sampling method, sequential sampling requires an ordered frame of sampling units from which the sample is selected using an explicitly stated probability or expectation, for example selecting every tenth case from a list.* *As a non-probability sampling technique, sequential sampling begins with the sampling of a single observation or group of observations which are then tested against the hypothesis. If the null hypothesis is not rejected, another observation or group of observations is sampled and the hypothesis tested again, and so on.*

Simple random sample	*A specific type of probability sample in which subjects are selected from a population list using a table of random numbers or a random number generator. (A random sample requires a list of population members in which each member can be assigned a discrete number.) The assignment of subjects to different treatments, interventions, or conditions according to chance, rather than systematically. Random assignment of subjects increases the probability that differences observed between subject groups are the result of the experimental intervention.*
Snowball sample	*A non-probability sample that is created by using members of the group of interest to identify other members of the group (for example, asking a subject at the end of an interview for suggestions about who else to interview).*
Stratified sample	*A probability sample that is determined by dividing the population into groups or strata defined by the presence of certain characteristics and then random sampling from each of the strata. This is a good way to make sure that a student sample is racially diverse (for instance).*
Systematic review	*A research method that provides a framework for drawing together and critically assessing the results from existing evidence on a focused question or topic. The evidence-base is selected according to clear criteria and the review is conducted through a standardized protocol. After the aggregation of all relevant literature is finalized, an appraisal of the quality of the studies and reports using the clearly defined criteria is made, and conclusions are drawn only from the data that meet the quality standards.*
Theory of change	*Explains how activities are understood to produce a series of results that contribute to achieving the final intended impacts. It can be developed for any level of intervention – an event, a project, a programme, a policy, a strategy or an organization.*