



5. Planning and undertaking evaluation

Planning and undertaking evaluation involves several stages, including:

- **Choice of methods:** choosing evaluation areas and methods according to information needs, the type of intervention and organizational capacity and resources.
- **Capacity-building:** identifying and training evaluators.
- **Study design:** selecting the evaluation framework.
- **Sample selection:** deciding who and how many to evaluate.
- **Ethics:** securing consent from participants and ethical approval.
- **Adaptation of methods:** adapting tools to local language and culture, piloting and revising as appropriate.

- **Data collection:** undertaking the evaluation and collecting data.
- **Data analysis:** entering and analysing data and reporting results.

The first stage, choosing evaluation methods, has been discussed in detail in Chapter 4; this chapter is concerned with the remainder. The intention is to raise awareness of the issues associated with each of the stages rather than to cover all aspects in detail. Please note that these stages do not necessarily need to be followed sequentially. For example, the study design affects the choice of evaluation methods and vice versa.

Study design

Some of the methods described in this catalogue were developed as research tools and need to be placed into an evaluation framework (e.g.

baseline survey and follow-up after 6 and 12 months). Other methods have a built-in retrospective evaluation element: i.e. they ask how things have *changed* as a result of an intervention. Methods will need to be chosen and adapted according to the stage of an intervention project or programme and the study design.

Selecting a particular study design depends on a number of factors, including:

- outcomes of interest;
- local conditions;
- available expertise;
- financial and staff resources; and
- stage of project or programme (i.e. planning, underway, completed).

Choosing the most appropriate study design is critical and seeking specialist knowledge may be required. Three key study design options are described in this catalogue; Box 16 describes an additional study design whose complexity goes beyond the scope of evaluation.

For interventions not yet underway:

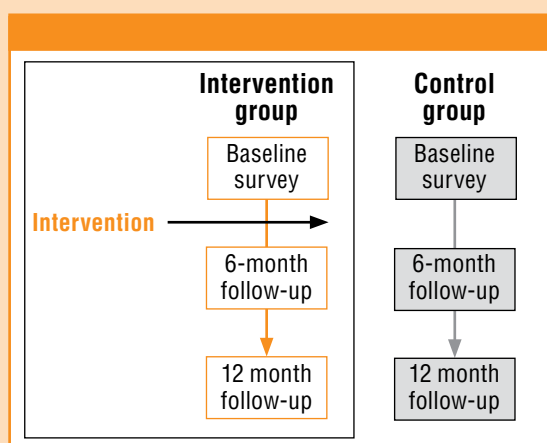
- Before-and-after design with control group (Boxes 11 and 12).
- Before-and-after design (no control group) (Box 13).

For interventions already underway or completed:

- Cross-sectional design (Boxes 14 and 15).

BOX 11 Before-and-after design with control group

This study design involves identifying a control group (ideally similar to the intervention group in every way except that this group does not receive the intervention), and undertaking baseline and follow-up surveys in both groups.



This study design is considered to be very robust, i.e. due to the presence of a control group it is less vulnerable to changes in seasons, economic or political instability and, very importantly, age-related changes in the vulnerability of the study population. It is, in fact, the only option for those wishing to accurately assess the impact of interventions on children's health.

Because of the control group, this study design requires a large sample size resulting in a lot of data to collect and process. The need to identify control homes of comparable composition and with similar socio-economic characteristics and household energy habits can be challenging. It is important not to interfere with the natural adoption process of the intervention, which may lead to some of the control group adopting the

intervention during the study. Finally, there are ethical implications of using a control group which does not benefit from the impacts of an intervention. Depending on the context of the project/programme and choice of control group, this problem can be overcome by providing the control group with access to the intervention at the end of the evaluation study.

BOX 12 Case study – before-and-after design with control group

The Appropriate Rural Technology Institute undertook this evaluation study at two sites in Maharashtra, India to assess the multiple impacts of one- and two-pot improved cooking stoves.

300 homes were studied in total: 150 intervention homes cooking on an improved stove, and 150 control homes from an adjacent community cooking on a traditional open fire. Homes were chosen according to the presence of a woman aged between 15 and 45 years, and of at least one child aged less than 5 years.

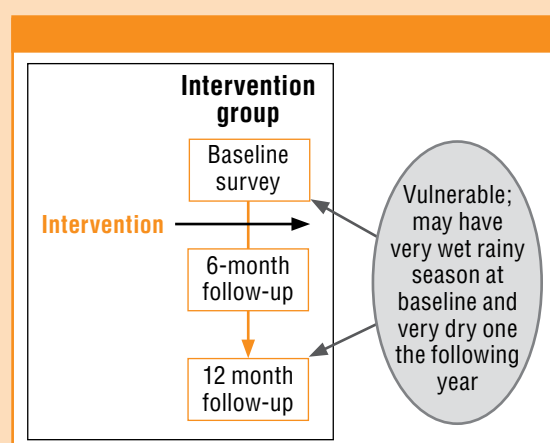
An in-depth survey was conducted to collect data at baseline as well as 6 and 12 months after the intervention was introduced. Brief questionnaires were employed after 3 and 9 months to maintain contact with communities. Moreover, interviews were held with key informants, followed by focus group discussions.

Evaluation areas included:

- pollution levels and personal exposure;
- performance;
- health and safety; and
- time, socio-economic and other impacts.

This case study was kindly provided by ARTI, India.

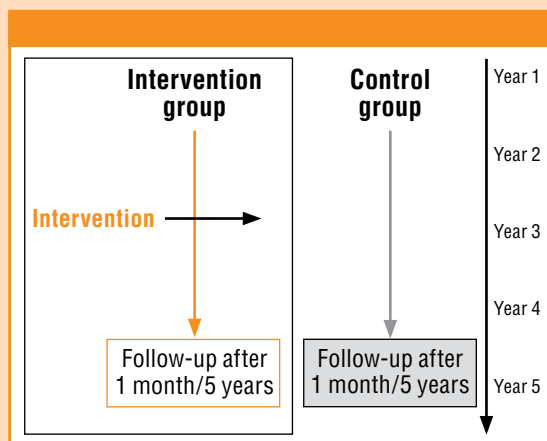
BOX 13 Before-and-after design (no control group)



This study design is similar to the one described in Box 11 with the exception of the control group. Baseline and follow-up surveys are undertaken only with the intervention group.

One of the main advantages of this design is that there is no need to identify and evaluate control homes. The smaller sample size makes the evaluation less resource-intensive and results in less data to collect, process and analyse. This design is suitable where the climatic, economic and political situation is fairly stable. It is vulnerable to any change in factors that affect energy use or health. The absence of a control group means that evaluators will not be able to distinguish changes in variables of interest due to the intervention from changes due to variability in economic, political or climatic conditions. Qualitative enquiry can help determine reasons for observed changes.

BOX 14 Cross-sectional design



This study design involves a one-off study of an intervention group and control group at some point (up to five years) after an intervention project or programme has been implemented.

A main advantage of this design is that it can be undertaken retrospectively. This makes it particularly suitable for evaluating projects which are already underway or completed. Moreover, it does not require a baseline and thus a single visit to each home is sufficient.

Some of the drawbacks of this design include the difficulty of ensuring that the differences between the intervention and control groups are due to the intervention and not other factors. Socio-economic

differences, in particular, may impact both the adoption of an intervention and household energy practices. Furthermore, one-off measurements may not be representative as these often do not reflect the situation at other times of the year (see Box 1). In order to achieve statistically significant results large sample sizes are required.

BOX 15 Case study – cross-sectional design

In 2002, a multidisciplinary team of Chinese and American researchers undertook a large-scale retrospective evaluation of the Chinese National Improved Stove Programme using a cross-sectional design. Three provinces were chosen to represent different adoption rates of improved stoves and prevailing fuels:

- Zhejiang: high adoption and widespread electricity and LPG use;
- Hubei: medium adoption and widespread biomass use; and
- Shaanxi: low adoption and widespread coal use.

In China, IAP levels are determined by a complex combination of different fuels and stoves used for a variety of activities. Moreover, due to space-heating, there are substantial summer-winter differences. The researchers carefully categorized different fuels and stoves to generate meaningful data.

Evaluation areas included:

- pollution levels and personal exposure; and
- health and safety.

The study required large sample sizes:

- A household survey was conducted in nearly 3500 households, amounting to more than 20 in each village including an oversample for women and children.
- 24-hour CO and PM monitoring was undertaken in 400 households, and repeated for a sub-sample during the winter.

This case study was kindly provided by Kirk Smith, UCB.

BOX 16 Randomized controlled trial

A randomized controlled trial (RCT) is the 'gold standard' approach to measuring the impact of a given intervention on a given health outcome, such as childhood pneumonia. This research design involves the detailed investigation of a randomized selection of clusters (e.g. villages) or individuals who have received an intervention as well as a control group that has not received the intervention. RCTs, such as the trial undertaken in Guatemala (see Chapter 2), are complex and costly and go much beyond the scope of an evaluation.

Sample size and sample selection

Choosing the right size for an evaluation study is critical. The sample size is the number of homes/individuals/stoves (i.e. sample) being evaluated. If sample sizes are too large, time and money will be wasted gathering and analysing unnecessary data. If sample sizes are too small, findings may not be representative or statistically significant, and may therefore lack credibility.

Organizations intending to contribute evidence to the international knowledge base must ensure that their results are statistically significant. Even basic evaluations whose results are not intended for a wider audience need to consider sample sizes to obtain meaningful results and in the interest of careful use of resources. Small-scale evaluations can produce 'false positives' or 'false negatives', where data gathered for too small a sample size do not allow the evaluator to draw firm conclusions.

Calculating optimal sample sizes is complicated and dependent on many factors including:

- prevalence (e.g. of symptoms);
- typical levels and variation (e.g. pollution);
- expected change attributable to the intervention; and
- study design.

Organizations may choose to seek specialist help or form partnerships with, for example, universities for this aspect of evaluation planning.

Sample selection

Choosing participants, homes or devices can follow many different approaches, including:

- **Random selection.** This approach is suitable for quantitative evaluation and can be undertaken in several ways. The most straightforward approach involves assigning a unique number to households that are eligible to participate in the evaluation, and choosing a sample according to random numbers generated by a computer programme.
- **Theoretical sampling.** This approach is suitable for qualitative evaluation. Discussing issues or asking questions to individuals is repeated until no more (or very few) new

responses are recorded. This point, where further study no longer contributes anything new on a particular subject, is called saturation.

Some organizations, such as Practical Action, have adopted a sampling method based on the natural adoption process to explore impacts of interventions during scaling up. This approach monitors only those people who have adopted interventions by choice. It provides evaluators with data based on what is happening in a real-life situation, and is deemed appropriate for market-based projects/programmes and behaviour change interventions.

It is vital that the right people are asked the right questions (Box 17). Practical considerations may also impact sample selection, such as location and accessibility.

BOX 17 Consulting the right people in evaluation

Sometimes those most willing and able to participate in studies are not necessarily those to whom evaluators want to listen. The poor and marginalized are often those least articulate or willing to speak.

For example, an improved stove project is implemented in rural Africa and a survey plans to investigate the socio-economic impacts of the intervention on women. Conducting the survey during daytime means that the majority of women are not at home but working in the field. Therefore, it will be difficult for fieldworkers to locate the appropriate group of women. If the fieldworkers were to ask the women who remained at home (who may not be the poorest, or those who have adopted the intervention) or other family members, they may obtain a false picture of the intervention impact.

Dropout and loss to follow-up

Sometimes participants will refuse to complete studies (dropout) or not be available during follow-up visits because they have moved away or cannot be traced (loss to follow-up). It is not possible to give exact figures on how many people will refuse to cooperate throughout an evaluation study, but many projects or programmes use sample sizes around 20% larger than statistically required so that, accounting for losses, results will still be valid (see Box 18).

BOX 18 Accounting for loss to follow-up in Guatemala

HELPS International evaluated the Onil stove in a small community of 48 families in Guatemala.

A before-and-after design was used to evaluate IAP levels and stove performance in homes using the Onil stove, and in homes cooking on a traditional open fire on a raised wooden box filled with earth.

Calculations based on an 80% reduction of IAP levels result in a required sample size of 30. To allow for 'loss to follow-up' 36 homes were monitored.

This case study was kindly provided by HELPS International, Guatemala.

Adapting evaluation methods

Most evaluation methods described in this catalogue have been developed for a specific geographical and cultural context, and most will need to be adapted to local conditions. All methods will need pilot testing prior to use (Box 19).

Examples of culture- and geography-specific aspects of methods include:

- local language and terminology (e.g. specific words for illnesses/symptoms, use of the Hindi word *chultha* for cooking stove);
- climatic and geographic conditions (e.g. need for space heating);
- cultural taboos (e.g. the reluctance of women to speak about coughing symptoms because of cultural stigmatization related to tuberculosis);
- cultural practices (e.g. the use of a traditional sauna or *temescal* in Guatemala); and
- locally specific cooking devices and practices (e.g. alcohol-brewing in Nepal which significantly contributes to IAP and thus impacts on the choice and effectiveness of a given intervention).

Data collection

Evaluation is undertaken for many reasons but, ultimately, it is undertaken to ensure that people (i.e. beneficiaries) have been well served, and to inform decisions about how to serve people better in the future. Data collection is at the

BOX 19 Pilot testing tools

Any newly developed or adapted evaluation tool will need to be pilot tested in the field. These are some of the key considerations in pilot testing:

- Does the flow of the questions work?
- Are the words understood? Are they too difficult, too simple, ambiguous (e.g. wheezing, stove names)?
- Do the response categories in quantitative surveys capture all options (e.g. plastic used as fuel in South African slums)?
- Are there any cultural sensitivities in relation to specific questions (e.g. asking about cough symptoms in India)?
- Are the questions interpreted in the same way by different respondents? (This is referred to as reliability.)
- Do they measure what they are supposed to measure? (This is referred to as validity.)
- Are the questions answered in the same way if repeated with the same respondent? (This is referred to as reproducibility.)

heart of evaluation: both interviewer and interviewee play a key role. This section addresses some of the issues related to choosing the right evaluator and designing and asking questions in the best possible way.

Choice of evaluator

The person that facilitates discussions or administers surveys can have a profound impact on data collection. Choosing the most appropriate evaluator depends on what information is being collected and from whom. For example, in some cultures it would be considered inappropriate for a male researcher to speak with female cooks. In other situations, respondents may feel intimidated or under pressure to respond positively if the project manager of the intervention project/programme is asking the questions.

It is important that evaluators are well-trained and that they and, where applicable, their translators are aware of the aims of the research. It is also vital that they appreciate that refusals to respond or negative responses are valuable out-

BOX 20 People-centred evaluation

Evaluation implicitly involves interaction with people: gaining access to their homes to monitor IAP levels or test the stove, or asking (sometimes personal) questions relating to their health, time use and socio-economic status. Evaluation is by nature intrusive, and it is important for evaluators to be sensitive to the expectations and specific needs of participants.

One of the key principles of participatory approaches is reversing roles. For evaluation this means that the development professionals become the learners and listeners, and the participants become the teachers and informers. Users of stoves are the experts, and evaluators need to recognize their knowledge on how the intervention has worked.

comes, and will not be viewed as failure on their part by the evaluation coordinators.

Designing and asking questions

In designing qualitative and quantitative materials it is important to minimize suggestion in questions, and to phrase questions in an open-ended way. For example:

‘Can you tell me about any difference in how much wood the two stoves use?’ ✓ *open question*

‘Is there any difference in how much wood the two stoves use?’ ✗ *closed question*

‘The new stove uses more wood, doesn’t it?’ ✗ *leading question*

People tend to respond more honestly to open questions than to closed or leading questions. It is also more difficult to answer an open-ended question if it has not been properly understood, whereas answering a closed question only requires a simple yes or no.

To aid data analysis open questions can include coded responses, for example:

‘Which stove do you use most of the time?’

- A. Single pot
- B. Double pot
- C. Three stone fire
- D. Other

In the interest of eliciting accurate and meaningful responses, participants should be made aware that:

- the purpose of the study or survey is to enable improvements to be made to the work;
- interviewers are equally interested in whether the situation is worse, the same or better; and
- their answer will not disadvantage them in terms of future assistance.

Be aware of language issues – particularly when using translators – as certain words (e.g. wheeze) may be difficult to translate accurately (see Box 19). If necessary, demonstrate words and concepts to make sure people have understood.

Feeding back to participants

Evaluation information voluntarily provided by participants should be made available to them upon request (see Box 20). Many organizations choose to share evaluation findings, partly as a show of appreciation for participation but also as a promotional tool. For example, results showing that an improved cooking stove adopted by some families in the community has resulted in increased disposable income, lowered IAP levels and reduced coughing in children could convince many more families to purchase one.

Ethical considerations

Wherever research involves human subjects, ethical issues need to be considered carefully (Box 21). Organizations intending to publish evaluation results need to take particular care, and those monitoring IAP levels, personal exposure and health outcomes may need to seek approval from an ethical review panel or institutional review board.

A range of example informed consent forms developed by WHO have been included in the accompanying CD-ROM. Further examples and information can be downloaded from www.who.int/rpc/research_ethics/informed_consent/en/

BOX 21 Three basic principles of ethics in research

Beneficence: 'the duty to do good'

- Research should cause no harm to participants, whether intentionally or by failing to anticipate and avoid harm.
- The research design should maximize benefits and minimize harm.

Respect for persons

Research should uphold the following principles:

- Autonomy or self-determination.
- Voluntariness, including the choice to opt out of activities at a later stage.
- Duty to protect persons with limited autonomy (e.g. children, refugees, women).
- Confidentiality, anonymity and privacy:
 - Numeric codes (instead of address/name) should be used on all forms/databases.
 - All records (e.g. health status, socio-economic status) should be stored in locked rooms with only study staff having access.

Justice

- Research should not create injustices, whether in relation to risks and discomforts or in relation to benefits.

Data management and data analysis

Data management

It is important to be systematic about data storage and management. Many of the methods included in this catalogue, for example the methods for testing stove performance and assessing IAP levels, are accompanied by data collection forms. These serve as a template for data entry and storage. Some survey questionnaires, for example those related to impacts on symptoms, time and socio-economic status, are lengthy and require many data entries for each participant. Similarly, qualitative methods, such as focus group discussions, can generate large volumes of data which require careful and skilled management.

Data analysis

Analysing data can be simple or complicated, depending on the methods employed and the type of outcome to be reported.

Producing descriptive statistics is usually relatively straightforward. It requires familiarity with general applications in software packages, such as Microsoft Excel or other spreadsheet programmes, and a basic understanding of mathematics. The outcomes produced tend to be numbers or percentages, such as the proportion of households using different types of stoves or fuels based on an adoption survey, the proportion of women reporting cough symptoms or the weekly amount of different fuels used for cooking.

Computer software is essential for downloading, analysing and storing data from digital IAP monitors, such as the HOBO CO monitor or the UCB particle monitor. Much or all of the required data processing, management and analysis can be performed in Excel or another spreadsheet programme.

Establishing relationships between the intervention and changes in an outcome of interest (e.g. pollution levels, health outcomes) is more difficult. It requires at least a basic understanding of statistics and epidemiology and experience with software packages, such as Microsoft Excel, EpiInfo, Stata, SAS or SPSS. An important question in evaluation research is how to distinguish between changes brought about by the intervention and changes due to other factors or chance.

An awareness of concepts, such as statistical significance and confounding, is therefore critical. Statistical significance allows differences that are meaningful to be distinguished from differences that are not meaningful and brought about by chance or small sample sizes. Adjusting for confounders (i.e. influences on the outcome of interest other than the intervention) is important as these may confuse, distort or mask true associations.

Some organizations may not have the skills to analyse the data they are well-equipped to collect. In this case, it is necessary either to develop the skills within the organization through specialist training (e.g. courses in statistics) or partner with organizations that already have the required knowledge (e.g. universities).

Reporting evaluation results

Communicating the results of an evaluation is critical. Reporting difficulties may help others avoid repeating the same mistakes. Sharing suc-

cesses may enable others to replicate these in different settings. Important communication channels include organization-specific reports, articles published through organizations engaged with household energy and health monitoring (some of which are listed in Chapter 7) and peer-reviewed articles in the scientific literature.

As with any report, the description of an evaluation should be as detailed as necessary while as concise as possible. Even technical experts en-

joy reading an interesting report, and findings should therefore be communicated in a clear and simple way. Usually it is not necessary to report every finding: key messages should be selected based on the target audience for the report.

Many evaluations will combine qualitative and quantitative methods including IAP monitoring, stove tests, questionnaires, discussions and observations. Data presentation should reflect this diversity and include the use of text, graphs, tables, quotations, photographs and even sketches.