unicef
Office of Research-Innocenti

Methodological Briefs
Impact Evaluation No. 7

Randomized Controlled Trials (RCTs)

Howard White, Shagun Sabarwal and Thomas de Hoop

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:
White, H., Sabarwal S. & T. de Hoop, (2014). Randomized Controlled Trials (RCTs), *Methodological Briefs: Impact Evaluation 7*, 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: Dugan Fraser, Simon Hearn, Patricia Rogers, Jessica Sinclair Taylor
Reviewers: Nikola Balvin, Samuel Bickel, Sarah Hague, Sudhanshu Handa, Aisha Yousafzai

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. RANDOMIZED CONTROLLED TRIALS: A BRIEF DESCRIPTION

A randomized controlled trial (RCT) is a way of doing impact evaluation in which the population receiving the programme or policy intervention is chosen at random from the eligible population, and a control group is also chosen at random from the same eligible population. It tests the extent to which specific, planned impacts are being achieved.

In an RCT, the programme or policy is viewed as an 'intervention' in which a treatment – the elements of the programme/policy being evaluated – is tested for how well it achieves its objectives, as measured by a predetermined set of indicators. The strength of an RCT is that it provides a very powerful response to questions of causality, helping evaluators and programme implementers to know that what is being achieved is as a result of the intervention and not anything else.

An RCT is an experimental form of impact evaluation; quasi-experimental and non-experimental forms of impact evaluation are addressed elsewhere in this series in Brief No. 6, Overview: Strategies for Causal Attribution; Brief No. 8, Quasi-experimental Design and Methods; and Brief No. 9, Comparative Case Studies.

The distinguishing feature of an RCT is the random assignment of members of the population eligible for treatment to either one or more treatment groups (who receive the intervention[1] treatment or variations of it) or to the control group (who receive either no intervention or the usual intervention, if the treatment is an innovation to an existing intervention). The effects on specific impact areas for the different groups are compared after set periods of time. Box 1 outlines the difference between random assignment and random sampling – two key features of an RCT.

Box 1. Random assignment vs random sampling

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).

The simplest RCT design has one treatment group (or 'arm') and a control group. Variations on the design are to have either:

- **multiple treatment arms**, for example, one treatment group receives intervention A, and a second treatment group receives intervention B, or
- **a factorial design**, in which a third treatment arm receives both interventions A and B.

In situations where an existing intervention is in use, it is more appropriate for the control group to continue to receive this, and for the RCT to show how well the new intervention compares to the existing one.

[1] RCTs can be used to measure both programme interventions (e.g., nutritional supplements distributed as part of a nutrition programme) and policy interventions (e.g., cash distributed as a result of a cash transfer policy). For reasons of brevity, any such intervention is referred to in this brief simply as a 'programme' or an 'intervention'.

In a simple RCT, the unit of analysis for the intervention and for the random assignment is the same. For example, when evaluating a programme that provides nutrition to individuals, individuals might be randomly assigned to receive nutritional supplements.

For both practical and ethical reasons, however, it is more usual to use a cluster RCT design, in which the unit of assignment contains multiple treatment units. For example, education interventions are usually assigned at the school level, although the intervention takes place at the level of the teacher, classroom or individual child, and effects are measured at the level of the child. Nutrition interventions, for example, can be assigned at the community or sub-district level. Given the kinds of large-scale programmes supported by UNICEF, cluster RCTs are more likely to be used.

Main points

1. An RCT measures the effect of a programme or policy intervention on a particular outcome.
2. The key feature of an RCT is that it uses **random** assignment of an intervention. This design is called an experimental design
3. An RCT is only useful for measuring impact in certain scenarios such as when a large sample is available; the intended impacts of the programme or policy intervention can be readily agreed and measured (e.g., reduction in stunting); and the RCT is planned before an intervention begins.

2. WHEN IS IT APPROPRIATE TO USE THIS METHOD?

RCTs should be planned from the beginning of the programme

An RCT needs to be planned from the beginning of programme implementation, and participation in the programme needs to be carefully controlled with the experiment in mind. RCTs cannot be undertaken retrospectively.

The exception to this is an 'encouragement design', which does not randomly assign participants to an intervention per se but to receiving promotional materials or additional information about the benefits of the available intervention to encourage participation. Encouragement designs can be used when a programme is universally available but has not been universally adopted.

RCTs need a large sample size

An RCT can only be used when the sample size is big enough to detect effects of the programme with sufficient precision; the study design must have what statisticians call sufficient 'power'.

'Power' is the probability of correctly concluding that an effective programme is working. Part of the process of designing an RCT is to perform power calculations, which indicate the sample size required to detect the impact of the programme (see box 2). Power increases with a larger sample.

In cluster RCTs, it is the number of clusters that determines a study's power rather than the number of observations. For example, a sample of 50 communities with 5 households sampled in each community has far more power (50 clusters) than one involving 25 communities with 10 households sampled in each (25 clusters), even though each has a total sample size of 250 households.

Software packages that perform power calculations (e.g., Optimal Design) are available, but it is best to leave this task to an individual experienced in doing power calculations.

Box 2. Power calculations

Statistical power refers to the probability of detecting an impact when a programme has an impact. To conduct power calculations and calculate the required sample size for an evaluation, evaluators usually use assumptions regarding the expected effect size, the statistical significance level and the intracluster correlation (for cluster RCTs). The intracluster correlation is a descriptive statistic between 0 and 1 that indicates how strongly the groups (e.g., households) or the individuals in the cluster resemble each other. The higher the intracluster correlation, the higher the required sample size. In cluster RCTs, there is usually a greater increase in statistical power when the number of clusters is increased than when the number of individuals or groups within a cluster is increased.

RCTs should be undertaken following formative research or evaluation

Using an RCT to evaluate a programme that has not reached maturity is likely to be inappropriate and, under most circumstances, an RCT should not take place until the programme has been adequately developed. This parallels the process of clinical drug trials, which follow a period of development and initial testing. Formative research or situation analysis should be used to assess the factors behind the problem being addressed by the programme (e.g., poor school performance) and so inform its design. For example, there is no point increasing student attendance alone if teacher absenteeism is rife.

Formative evaluation, in the form of a pilot study or proof of concept study, for example, tests the feasibility of implementing the programme with sufficient take-up and improves the quality of implementation. Results from a pilot study can identify modifications that need to be made to the design of a larger study that might follow the pilot. Although RCTs have sometimes been used to provide a proof of concept, they should not be used at such an early stage. RCTs are costly and if used to evaluate a programme that has not yet been fully developed will waste scarce resources and can generate misleading findings.

RCTs must be appropriate to the nature of the programme being assessed

RCTs are best used for programmes that seek to achieve clear, measurable impacts that can be attributed to a distinct intervention, or set of interventions, and which lend themselves to causal pathway analysis. RCTs are not well suited to programmes that are emergent, or which seek to achieve results that are hard to measure.

Four conditions under which random assignment is undesirable or unfeasible have been identified.[2] These are: when quick answers are needed; when the need for precision is low and the causal question is not the most important goal; in conditions where it is impossible to manipulate assignment such as when the causal question that needs to be answered involves exposure to an undesirable condition; and when sufficient preliminary and empirical work has not been done and the intervention or programme is at a premature stage.

Examples of programmes which may not be amenable to randomization are those where there is a small number of treatment units, for example, institutional support to a single agency, and ones in which programme activities and expected outcomes are not clearly defined in advance. Sometimes, however, a programme that at first sight appears unsuited to randomization may become suitable with a little

[2] Shadish, William R., et al., Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton Mifflin, Boston, 2002.

imagination. For example, a national programme can be evaluated using an encouragement design. In an institutional reform programme, policies for worker pay could be developed through an RCT of different incentive packages facing workers.

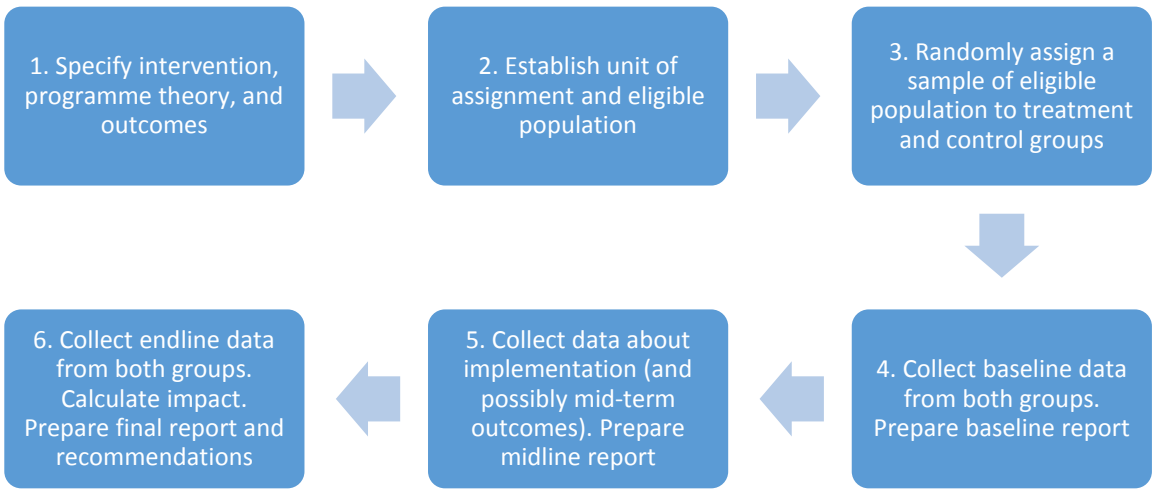
In situations where the use of an RCT is inappropriate, the relevant decision makers need to be informed that this is the case. To undertake an impact evaluation in such situations, a quasi-experimental design (such as propensity score matching or PSM) or a rigorous non-experimental design (such as process tracing) could be used instead. (See Brief No. 8, Quasi-experimental Design and Methods and Brief No. 6, Strategies for Causal Attribution.)

Alternatively, if it is too soon to implement an impact evaluation, it might be possible and useful to do an evaluation that focuses lower down the causal chain on outputs or outcomes. An outcome evaluation collects information about outcomes (shorter-term results) that are closely associated with the impacts of interest. If the causal chain is well understood and certain, this can be appropriate.

In a vaccination campaign, for example, it might be more appropriate to report on immunization status in an evaluation rather than wait to collect data about subsequent morbidity or mortality. As another alternative, a process evaluation could be undertaken to improve implementation and make a subsequent impact evaluation more useful.

3. HOW TO CONDUCT A RANDOMIZED CONTROLLED TRIAL

Figure 1. Overview of conducting an RCT



1. Specify intervention, programme theory, and outcomes

As with any impact evaluation, an RCT should start by clearly specifying what is being evaluated and why, and the outcomes and impacts of interest. A programme theory of change that articulates what the programme intends to achieve and how – what the change processes will be and how the programme

activities will produce these – is extremely helpful in this regard (see Brief No. 2, Theory of Change). Analysis of the theory of change assists in the identification of evaluation questions related to the causal chain and helps to determine which impacts should be evaluated. It is also useful at this stage to clarify how the evidence generated through the RCT will be used.

2. Establish the relevant population and unit of assignment

As noted earlier in this brief, it is important when designing an RCT that the eligible population and the unit of assignment for randomization purposes are clearly identified and that consistency is ensured. An important decision that must be made relates to the unit of assignment – i.e., whether individuals should be randomized to treatment and control or whether groups of individuals such as schools or villages should be randomized to treatment and control. Evaluators should also decide upon the potential subgroups of interest at the beginning of the study, so it can be ensured that the study is sufficiently powerful to conduct the subgroup analyses of interest.

3. Randomly assign a sample of the eligible population to treatment and control groups

RCTs can be designed in various ways. There are different ways of implementing an RCT for a programme and the RCT design can be selected according to the programme characteristics. Three common designs are described below.

Pipeline randomization means that all units of assignment will receive the programme, if it is found to be effective, over time. So here it is the time of entry to the programme that is randomly assigned. Implementing agencies often roll out a programme in stages, making it possible to randomly select the order in which the participants receive the programme. For example, if budgetary and logistical constraints prevent the immediate nationwide roll-out of a programme, it may be possible to randomly select units that will receive the programme during the first stage. A well known example of this approach is Mexico's Progresa/Oportunidades conditional cash transfer (CCT) programme.[3] In its initial phase, the programme was a pilot for 506 communities, half of which received the programme first and half of which acted as a control group for two years. Hence, the communities were randomly allocated into the two groups to receive the programme in either year one or year three (i.e., those receiving the programme in year three served as a control group for two years).

(Raised) threshold randomization can be used when the eligible population is larger than can be served with the budgetary resources available. Since the programme will not be made available to all of those who are eligible, random assignment into the programme can be the fairest and most transparent means of deciding who gets in. Such cases often occur when a threshold determines eligibility, e.g., the poverty line. If the eligible population can be covered in its entirety by the available budget, then raising the threshold slightly can make randomization possible. For example, if the eligibility criterion for a nutrition programme is households with children aged up to 24 months, this threshold could be raised to 30 months. An analogous approach can be used geographically. A programme planning to work in 50 communities can first identify 100 communities and then randomly select 50 communities from this total to enter the programme. In this latter case, the technique of matched pair randomization – that is, putting communities into pairs and randomly assigning one community of each pair into the treatment group – would increase the strength of the design.

Encouragement designs are used for programmes and policies that are universally available but not universally adopted. The treatment group is provided with an encouragement to take up the intervention, but this encouragement should not be something that in and of itself affects the intervention. One example of a suitable encouragement is the conducting of information campaigns for an ongoing programme in

[3] See Secretaría de Desarrollo Social (SEDESOL), Oportunidades, www.oportunidades.gob.mx.

certain villages but not others. The villages where the campaigns will be conducted are selected at random from among all of the villages where the programme has been implemented. The researchers then measure the impact of the programme on outcomes of interests by comparing the outcome between the control and treatment villages (in this case, villages exposed to the information campaign). This approach allows an impact estimate to be made because of the differential take-up rates between those villages exposed to the information campaign and those not.

It is possible to randomly assign population groups to the treatment and control groups in several ways, including:

- **Simple randomization –** Individuals or sites are listed and then assigned to the treatment and control groups using random numbers, for example, issued by a random number generator.
- **Matched pair randomization** – Individuals or clusters are grouped into pairs based on having similar observable characteristics. One unit in each pair is randomly assigned to the treatment group and the other to the control group. This initial matching helps to ensure balance and reduces the required sample size.
- **Stratified random assignment** – For key variables likely to influence results, for example, income or education, participants are divided into groups (strata) such as low, medium and high income and then randomization is conducted for each group. This ensures an equivalent distribution of key variables across the treatment and control groups.

The random assignment process must be followed, and checks made that it is being adhered to, throughout the evaluation (see below).

4. Collect baseline data from both groups

Before or after the random assignment of participants, evaluators usually conduct a baseline survey to generate the data that will be used as the basis for endline (and perhaps subsequent) comparisons. Baseline data are also used to assess the equivalence of baseline characteristics between the treatment and control groups. This assessment of equivalence, also known as balance checks, ascertains whether the mean of the treatment group and the mean of the control group are similar for different observable variables. The main reason for doing this is that it helps to confirm that randomization was successful. In cases where important differences are found (or anticipated) the use of stratified random assignment may be warranted.

As noted above, the survey sample size is determined on the basis of a power calculation. Usually, the baseline survey takes place at the individual or household level. Data are collected on household characteristics, socio-economic situation, education, health and all of the other characteristics that are potentially associated with the programme to be evaluated and the impacts it seeks to achieve.

5. Collect data about implementation (and possibly mid-term outcome data)

Data that provide information on implementation should be collected, possibly through a mid-term survey, which will usually focus on process aspects. Such a survey may also be used to provide initial estimates of programme impact if it is not premature to do so.

It is important to check that people in the control group do not suffer from 'contamination', either through a similar intervention being carried out in the control areas or through self-contamination, where participants of the study cross over from one arm of the study to the other, thereby contaminating the initial randomization process. It is also important to check attrition (when participants of the study drop out from the sample between one data collection round and another) between the groups as this can produce misleading results. For example, if fewer people in the control group provide outcome data than in the

treatment group, this would skew the results (as those participants who drop out are excluded from the analysis).

6. Collect data on impacts

Following the implementation of the programme, an endline survey is conducted. When the endline should be depends on the theory of change as to how long it will take for the expected impacts to occur. Where programmes continue for a longer period of time, endline data can be collected after the passing of a reasonable period, following which it can be expected that any change in outcomes due to the intervention will begin to be seen.

For example, for a nutrition intervention involving iron fortification, researchers should ensure that a sufficient amount of time has elapsed for the programme to be rolled out and for participants to have had adequate exposure to iron fortification for iron to have been absorbed in the body and for anaemia incidence to have started to decrease (due to increased consumption of iron). The endline data are used to calculate impact estimates.

If randomization was not very successful (as determined by the balance checks done earlier) then the difference-in-differences (DID) method can be used. As outlined in Brief No. 10, Overview: Data Collection and Analysis Methods in Impact Evaluation, difference-in-differences measures the difference in the change in outcome between the treatment group and the control group. It is also possible to assess heterogeneous and differential effects for subgroups using difference-in-differences.

The effect of the programme may differ according to different groups such as men versus women, rich versus poor and educated versus uneducated. Comparing outcomes across these different categories, according to control and treatment groups, can help in estimating the impacts for these subgroups. This can be done by comparing difference-in-differences in effects for the various subgroups of the eligible population, for example, men, women and children.

Depending on the nature of the programme, the evaluators may also be interested in a post-endline survey to estimate longer-term effects of the intervention. The timing of the follow-up surveys would depend on how quickly the intervention is likely to yield results. For example, food transfers can result relatively quickly in nutritional benefits, while interventions that aim to change existing attitudes, norms and behaviours usually take a longer time to achieve results.

4. ETHICAL ISSUES AND PRACTICAL LIMITATIONS

RCTs face a range of ethical and practical concerns similar to those faced by all evaluations (see Brief No. 1, Overview of Impact Evaluation) but this section highlights those issues that are specific to RCTs.

There are particular ethical concerns around RCTs that relate to their experimental nature and which make it important for participants in the trial to be consulted and their wishes identified and addressed, and for the associated risks and benefits to be balanced. The ethical concerns around experimentation become even more salient in the case of RCTs that involve a control group that does not receive any intervention. The potential for disadvantage to consequently occur makes it very important that randomization is a transparent process, especially when randomizing at the individual level. It is the evaluator's responsibility to ensure that no tensions exist between the treatment and control groups.

One of the ways of mitigating this possibility is by clearly explaining the purpose of randomization. A pipeline design can allay these ethical concerns if it can be ensured that the intervention will be rolled out to the control group later if it is found to be effective.

Another ethical concern surrounds the need for an RCT in the first place. When there is no reasonable doubt about the benefits and cost-effectiveness of a programme, then there is no need for an in-depth

evaluation (of any kind) and impact monitoring may be more appropriate to assess whether the programme continues to have the intended results over time. If there are questions about a programme's effectiveness, and only limited resources available for its implementation, however, it may be considered most ethical to randomly assign the participants to the programme – with the intention of rolling out the programme to the whole population if it is found to be effective.

It is also important to be sensitive about the data collection involving the control group. Evaluators need to take into consideration that they are using the time of non-recipients appropriately. It is sensible to compensate the respondents in a survey for their time, although it should be done in such a way that it does not affect the results (e.g., by encouraging individuals to respond in a particular way).

5. WHICH OTHER METHODS WORK WELL WITH THIS ONE?

An RCT is a research design aimed explicitly at answering questions around causality and attribution. The RCT design has to be located within an overall evaluation plan that must also include methods for data collection (such as interviews, observations or direct measurements) and analysis.

It is recommended that formative research or evaluation is undertaken before doing an RCT to assess the feasibility of the implementation of a programme and to improve implementation.

It is also recommended that an RCT either includes a component that will review implementation processes or that it is complemented by a process evaluation. While the RCT addresses the question of the counterfactual, a process evaluation will address questions about how the programme was implemented, usually drawing on both quantitative and qualitative data. A process evaluation thus addresses questions along the causal chain and is helpful in explaining the reasons for programme impact. The data collected will complement, and can be used to verify (or disprove), the data from the programme's own monitoring system.

6. PRESENTATION OF RESULTS AND ANALYSIS

An RCT requires quality assurance in order to guarantee the quality of the study, and it is important to provide sufficient detail when writing up the methodology and findings. It is also important to focus not only on the methodology when describing an RCT but also to describe the intervention being evaluated. This information can come from the corresponding process evaluation. A detailed description of the intervention allows for the theory of change to be linked with the analysis of the findings. When reporting the findings of an RCT, a detailed description of the theory of change should also be provided.

It is recommended when detailing the methodology that the sampling is described as well as the random allocation method. In this description, it is important to describe both the number of clusters and the number of households and/or individuals in the treatment group and the control group. The report should also include tables on balance checks (described above).

The impact estimates can be reported using difference-in-differences analysis. The findings from the difference-in-differences analysis can be reported for the entire sample as well as for subgroups to analyse heterogeneous effects.

The findings then need to be linked to the theory of change. Does the analysis support the theory of change? If not, which assumption behind the theory of change was unfulfilled? Further findings from this study should be explained. What possible reasons, from both within and outside the theory of change, could have led to the results? This analysis can help the evaluators to identify concrete policy-relevant findings. These findings should be presented in the conclusion of the report and should be explicitly linked to the data analysis. In most cases, it is also important to include a discussion on whether the results can be extrapolated to different settings, and, if so, which ones.

7. EXAMPLE OF GOOD PRACTICES

The Pakistan Early Child Development Scale Up Trial[4] funded by UNICEF aimed to evaluate the effectiveness and feasibility of integrating early childhood interventions to strengthen and improve early child health outcomes. This trial used the existing lady health worker (LHW) programme to deliver the various components of the intervention. The study had a cluster-randomized factorial design and took place in rural Pakistan.

Different clusters (defined as LHW catchment areas) were randomized to these different groups: the **control group**, which received basic health and nutrition services from the LHW programme; the **enhanced nutrition group**, which received nutrition counselling, responsive feeding advice and a nutrition supplement (Sprinkles®) from 6 to 24 months of age; the **early childhood development group**, which received stimulation and care for development advice, including coached practice integrated with routine monthly home visits and group meetings; and, finally, the **third treatment arm**, which received a combination of both the early childhood development and the enhanced nutrition interventions. Interventions were delivered by the LHW to every family in her catchment with a young child under 24 months old.

The children were measured on a number of outcomes and data recorded for the children and their families at various points until the children reached 24 months of age. The data collection team was separate from the intervention support team and blind to the intervention the child received. This minimized any bias in assessment. The study found that at 12 months of age, the children in all three of the intervention groups had significantly greater cognitive, language, motor and social-emotional scores compared to those in the control group. The integrated early childhood development and enhanced nutrition groups had significantly better cognitive and language scores than the enhanced nutrition group alone. At 24 months of age, all three of the intervention groups had significantly greater cognitive, language and motor scores compared to the control group, but the two groups with exposure to early childhood development did better than those exposed to enhanced nutrition alone.

These are examples of good practices in using an RCT, both in terms of its appropriateness for the situation and the way in which it was implemented and used.

The programme was well defined and had clear objectives. A list of eligible communities was identified and the communities randomly assigned to treatment and control groups. Data collection and analysis were based on a strong theory of change, allowing the evaluators to assess the pathways through which the intervention has been successful in achieving its objectives.

8. EXAMPLES OF CHALLENGES

Maintaining the integrity of the design: Even if random assignment is put in place, there are several potential challenges. These are: (1) low take-up of the intervention; (2) lack of compliance with intended procedures; (3) contamination of the control group by other interventions affecting similar outcomes or through self-contamination; and (4) change in the design or location of the programme being evaluated. Most of these problems can be dealt with at the analysis stage, but the evaluators need to collect the necessary data in order to be aware of the issues and able to address them.

Low take-up affects schemes that are not of interest, or poorly understood, by the intended recipients. For example, insurance schemes often suffer from low take-up. Qualitative data are usually required to understand the reasons for low take-up.

[4] Yousafzai, A. K., et al., 'The Pakistan Early Child Development Scale Up (PEDS) Trial: Outcomes on child development, growth and health', PEDS Trial outcome data report, UNICEF Pakistan, 2012.

As an example of lack of compliance, an RCT in China[5] providing eyeglasses to high school students found that glasses usage had also risen in some of the control group. Further enquiry revealed that the doctors performing the eye tests had glasses left over from the treatment group and so had given them to students in the control areas – an example of self-contamination. The study used a matched pair design and so was able to drop the pairs in which the control had been contaminated.

Failing to adjust standard errors when a cluster design has been used: This is a common technical error, which artificially increases power and may incorrectly conclude that a programme is working when in fact it is not. For example, suppose schools are randomized to different treatment arms but during the analysis of the results – in which the learning outcomes are compared using the test results of children in the treatment and control arms – there is a failure to control for school-level clustering, then the estimate of the impact would be an overestimate. All statistical software used for impact evaluations allows for this adjustment to be made, so researchers have no reason not to do so.

Excessive focus on the average treatment effect: An RCT provides an unbiased estimate of the mean effect of a programme. This is, however, rarely the finding of most interest to policymakers, who are often particularly interested in how effective a programme is for particular subgroups, especially those programmes that address equity issues. For example, an evaluation of Early Head Start,[6] an early childhood intervention programme in the USA, found that the programme, which was on average effective, was actually harmful for the most vulnerable families. A simple reporting that focused on its average effect would provide misleading guidance to policymakers and service providers. A way to counter this problem is to report results for different subgroups that might be impacted differently by the intervention. This type of design must be identified during the study design phase, however, and power calculations done accordingly.

Opposition to random assignment: There is often opposition to random assignment from the staff of the implementing agency. Having management agreement may be insufficient to gain the cooperation of fieldworkers. The study of eyeglasses provision in China is an example of this, where doctors gave away glasses to the control group.

9. KEY READINGS AND LINKS

Ambroz, Angela, and Marc Shotland, 'Randomized Control Trial (RCT)', web page, BetterEvaluation, 2013. See http://betterevaluation.org/plan/approach/rct.

Bloom, Howard, 'The Core Analytics of Randomized Experiments for Social Research', MDRC Working Papers on Research Methodology, MDRC, New York, 2006. Open access: http://www.mdrc.org/sites/default/files/full_533.pdf.

Duflo, Esther, et al., 'Using Randomization in Development Economics Research: A Toolkit', Department of Economics, Massachusetts Institute of Technology and Abdul Latif Jameel Poverty Action Lab, Cambridge, 2006. See http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf.

Gertler, Paul J., et al., Impact Evaluation in Practice, World Bank, Washington, DC, 2010. See http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf.

[5] Glewwe, Paul, et al., 'The Impact of Eyeglasses on the Academic Performance of Primary School Students: Evidence from a Randomized Trial in Rural China', conference paper, University of Minnesota and University of Michigan, 2006. See http://ageconsearch.umn.edu/bitstream/6644/2/cp06gl01.pdf.

[6] Westhorp, Gill, 'Using complexity-consistent theory for evaluating complex systems', *Evaluation*, 18(4), 2012, pp. 405–420.

Glennerster, Rachel, and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*, Princeton University Press, Princeton, 2013.

Raudenbush, S.W., et al., Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01), software, 2011. Available from www.wtgrantfoundation.org.

Spybrook, J., et al., Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0, software documentation, 2011. Available from www.wtgrantfoundation.org.

Urbaniak, Geoffrey C., and Scott Plous, Research Randomizer (Version 4.0), Resources for random assignment and random sampling, software, 2013. Available at http://www.randomizer.org/.

White, Howard, 'An introduction to the use of randomised control trials to evaluate development interventions', *Journal of Development Effectiveness*, 5(1), 2013, pp. 30–49. Open access: http://www.tandfonline.com/doi/pdf/10.1080/19439342.2013.764652.

GLOSSARY

Attrition	*When participants of the study drop out from the sample between one data collection round and another. Attrition can be a threat to the internal validity of a study, and it can change the composition of the study sample.*
Causal pathway analysis	*An analysis based on a causal pathway (AKA: analytical framework), which depicts direct and indirect relationships between the independent (interventions) and dependent (outputs/outcomes/impacts) variables.*
Cluster RCT design	*An experimental design in which the unit of assignment contains multiple treatment units rather than one subject. Examples include education interventions, which are usually assigned at the level of the school, even though the intervention takes place at the level of the teacher, classroom or individual child, and effects are measured at the level of the child See: Randomized Controlled Trial (RCT).*
Contamination	*The inclusion of an individual or group of respondents in a treatment (intervention) group who do not represent the population or who are not supposed receive the treatment. This can occur when participants/subjects in a control group inadvertently receive treatment, thereby reducing the effects of the treatment on outcome measures.*
Difference-in-differences (DID)	*Also known as the 'double difference' method, DID compares the changes in outcome over time between treatment and comparison groups to estimate impact.*
Encouragement design	*An experimental design that does not randomly assign participants to an intervention per se but to receiving promotional materials or additional information about the benefits of the available intervention to encourage participation. Encouragement designs can be used when a programme is universally available but has not been universally adopted. See: experimental design.*
Endline survey	*A survey undertaken at the conclusion of the intervention, usually for the purposes of comparing the results to the baseline survey. Related: baseline survey.*
Experimental design (RCT)	*A research or evaluation design with two or more randomly selected groups (an experimental group and control group) in which the researcher controls or introduces an intervention (such as a new programme or policy) and measures its impact on the dependent variable at least two times (pre- and post-test measurements). In particular RCTs – which originated in clinical settings and are known as the 'gold standard' of medical and health research – are often used for addressing evaluative research questions, which seek to assess the effectiveness of programmatic and policy interventions in developmental settings.*
Non-experimental design	*A type of research design that does not include a control or comparison group and/or does not include a baseline evaluation. Thus, several factors prevent the attribution of an observed effect to the intervention. See: experimental research design, quasi-experimental research design.*

Process tracing	*A case-based approach to causal inference which involves developing alternative hypotheses and then gathering evidence (clues) within a case to determine whether or not these are compatible with the available hypotheses.*
Propensity score matching	*A quasi-experimental method which matches treatment individuals/households with similar comparison individuals/households, and subsequently calculates the average difference in the indicators of interest.*
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.*
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.*
Treatment group	*Subjects/participants exposed to the levels of the independent variable; also called the experimental or intervention group.*