unicef
Office of Research-Innocenti

Methodological Briefs
Impact Evaluation No. 8

Quasi-Experimental Design and Methods

Howard White and Shagun Sabarwal

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., & S. Sabarwal (2014). Quasi-experimental Design and Methods, *Methodological Briefs: Impact Evaluation 8*, 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: Greet Peersman
Reviewers: Nikola Balvin, Sarah Hague, Debra Jackson

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. QUASI-EXPERIMENTAL DESIGN AND METHODS: A BRIEF DESCRIPTION

Quasi-experimental research designs, like experimental designs, test causal hypotheses. In both experimental (i.e., randomized controlled trials or RCTs) and quasi-experimental designs, the programme or policy is viewed as an 'intervention' in which a treatment – comprising the elements of the programme/policy being evaluated – is tested for how well it achieves its objectives, as measured by a pre-specified set of indicators (see Brief No. 7, Randomized Controlled Trials). A quasi-experimental design by definition lacks random assignment, however. Assignment to conditions (treatment versus no treatment or comparison) is by means of self-selection (by which participants choose treatment for themselves) or administrator selection (e.g., by officials, teachers, policymakers and so on) or both of these routes.[1]

Quasi-experimental designs identify a comparison group that is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics. The comparison group captures what would have been the outcomes if the programme/policy had not been implemented (i.e., the counterfactual). Hence, the programme or policy can be said to have caused any difference in outcomes between the treatment and comparison groups.

There are different techniques for creating a valid comparison group, for example, regression discontinuity design (RDD) and propensity score matching (PSM), both discussed below, which reduces the risk of bias. The bias potentially of concern here is 'selection' bias – the possibility that those who are eligible or choose to participate in the intervention are systematically different from those who cannot or do not participate. Observed differences between the two groups in the indicators of interest may therefore be due – in full or in part – to an imperfect match rather than caused by the intervention.

There are also regression-based, non-experimental methods such as instrumental variable estimation and sample selection models (also known as Heckman models). These regression approaches take account of selection bias, whereas simple regression models such as ordinary least squares (OLS), generally do not. There may also be natural experiments based on the implementation of a programme or policy that can be deemed equivalent to random assignment, or to interrupted time series analysis, which analyses changes in outcome trends before and after an intervention. These approaches are rarely used and are not discussed in this brief.

Methods of data analysis used in quasi-experimental designs may be ex-post single difference or double difference (also known as difference-in-differences or DID).

Main points

1. Quasi-experimental research designs, like experimental designs, test causal hypotheses.
2. A quasi-experimental design by definition lacks random assignment.
3. Quasi-experimental designs identify a comparison group that is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics.
4. There are different techniques for creating a valid comparison group such as regression discontinuity design (RDD) and propensity score matching (PSM).

[1] Shadish, William R., et al., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, 2002, p. 14.

2. WHEN IS IT APPROPRIATE TO USE QUASI-EXPERIMENTAL METHODS?

Quasi-experimental methods that involve the creation of a comparison group are most often used when it is not possible to randomize individuals or groups to treatment and control groups. This is always the case for ex-post impact evaluation designs. It may also be necessary to use quasi-experimental designs for ex-ante impact evaluations, for example, where ethical, political or logistical constraints, like the need for a phased geographical roll-out, rule out randomization.

Quasi-experimental methods can be used retrospectively, i.e., after the intervention has taken place (at time t+1, in table 1). In some cases, especially for interventions that are spread over a longer duration, preliminary impact estimates may be made at mid-term (time t, in table 1). It is always highly recommended that evaluation planning begins in advance of an intervention, however. This is especially important as baseline data should be collected before the intended recipients are exposed to the programme/policy activities (time t-1, in table 1).

Table 1. Timing of intervention and data collection for impact evaluations with a large sample size

Pre-intervention	Intervention	Post-intervention
t-1 Baseline	t (Mid-term survey)	t+1 Endline

t = a specific time period

3. QUASI-EXPERIMENTAL METHODS FOR CONSTRUCTING COMPARISON GROUPS

Propensity score matching (PSM)

What is matching?

Matching methods rely on observed characteristics to construct a comparison group using statistical techniques. Different types of matching techniques exist, including judgemental matching, matched comparisons and sequential allocation, some of which are covered in Brief No. 6, Overview: Strategies for Causal Attribution. This section focuses on propensity score matching (PSM) techniques.

Perfect matching would require each individual in the treatment group to be matched with an individual in the comparison group who is identical on all relevant observable characteristics such as age, education, religion, occupation, wealth, attitude to risk and so on. Clearly, this would be impossible. Finding a good match for each programme participant usually involves estimating as closely as possible the variables or determinants that explain the individual's decision to enrol in the programme. If the list of these observable characteristics is very large, then it becomes challenging to match directly. In such cases, it is more suitable to use PSM instead.

What is PSM?

In PSM, an individual is not matched on every single observable characteristic, but on their propensity score – that is, the likelihood that the individual will participate in the intervention (predicted likelihood of participation) given their observable characteristics. PSM thus matches treatment individuals/households with similar comparison individuals/households, and subsequently calculates the average difference in the indicators of interest. In other words, PSM ensures that the average characteristics of the treatment and comparison groups are similar, and this is deemed sufficient to obtain an unbiased impact estimate.

How to apply PSM

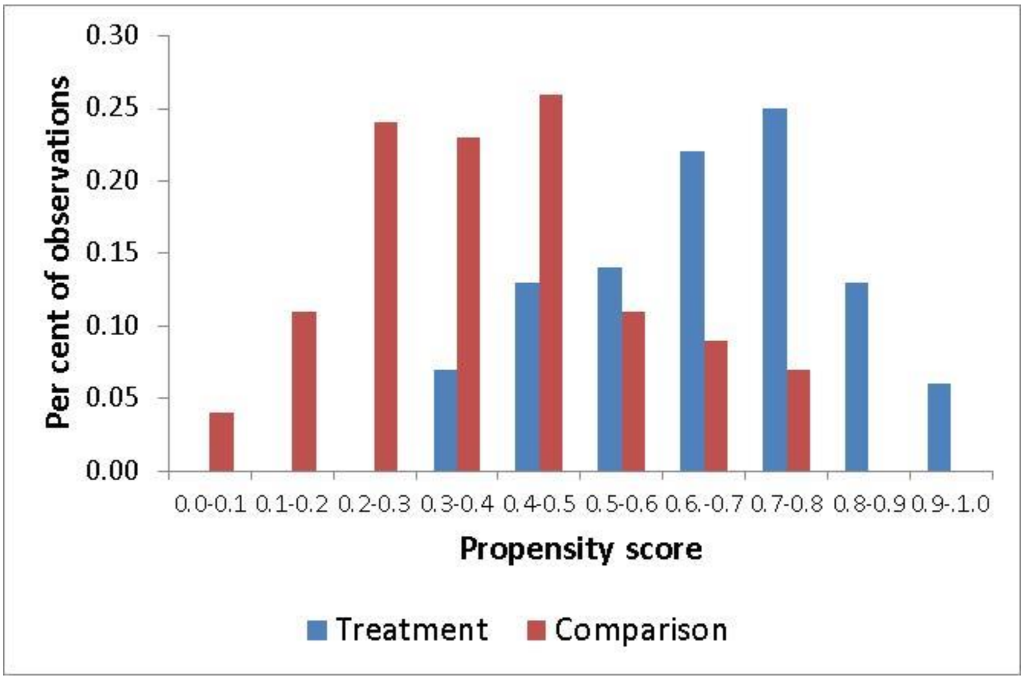
PSM involves the following five steps:

1. **Ensure representativeness** – Ensure that there is a representative sample survey of eligible participants and non-participants in the intervention. Baseline data are preferred for calculating propensity scores. This technique can, however, also be used with endline data: the matching variables must be variables that are unaffected by the intervention.
2. **Estimate propensity scores** – The propensity scores are constructed using the 'participation equation', which is either a logit or probit regression with programme participation as the dependent variable (in the programme = 1, not in the programme = 0). The characteristics deemed to affect participation should be well considered and as exhaustive as possible, but should exclude characteristics that may have been affected by the intervention. For this reason, it is best to use baseline data, where available, to estimate the propensity scores.
3. **Select a matching algorithm** – Each member of the treatment group is then matched to one or more members of the comparison group. There are different ways of doing this such as matching each participant to their 'nearest neighbour' non-participant. The mean of the closest five neighbours is most commonly used.

   A single individual in the comparison group may be matched to several different individuals in the treatment group.

   In order for the matching to be valid, it is essential to compare 'observed values' for participants and non-participants with the same range of characteristics. Observations in the comparison group with a propensity score lower than the lowest observed value in the treatment group are discarded. Similarly, observations in the treatment group with a propensity score higher than the highest observed value in the comparison group are also discarded. What remains is known as 'the region of common support' – an example is detailed in figure 1.

Figure 1. Example of a distribution of propensity scores – region of common support is 0.31 to 0.80



Source: Data created by authors for illustrative purposes only.

Figure 1 shows a typical distribution of propensity scores. The distribution for the treatment group is to the right of that of the comparison group – that is, treatment group individuals tend to have higher propensity scores than those in the comparison group. No member of the treatment group has a propensity score of less than 0.3, and no member of the comparison group has a propensity score of more than 0.8. So, in establishing the region of common support, the 39 per cent of comparison group observations with a propensity score of 0 to 0.3 are ignored, along with the 19 per cent of treatment group observations with a propensity score of 0.8 to 1. (In practice, a more precise cut-off would be used than that shown by the categorical classification of the data.)

Table 2 shows the matching for selected variables from a PSM analysis for a study of the impact of access to clean water in Nepal.[2] The 'before matching' column compares the average characteristics of households with access to clean water in the treatment group compared to all those households without access to clean water in the comparison group. These two groups of households are very different: those with access to clean water are more likely to be urban, better educated and better off than those without access to clean water. Any difference in child diarrhoea between the two groups cannot be readily attributed to access to clean water, however, since there are many other differences that may explain why the incidence of child diarrhoea varies between the groups.

[2] Bose, Ron, 'The impact of Water Supply and Sanitation interventions on child health: evidence from DHS surveys', conference paper, Bi-annual Conference on Impact Evaluation, Colombo, Sri Lanka, 22 to 23 April 2009.

Table 2. Observable characteristics before and after matching (percentage of group displaying the characteristic)

Variable	Before matching		After matching	
	Treatment (%)	Comparison (%)	Treatment (%)	Comparison (%)
Rural resident	29	78	33	38
Richest wealth quintile	46	2	39	36
Household head higher education	21	4	17	17

Source: Bose, Ron, 'The impact of Water Supply and Sanitation interventions on child health: evidence from DHS surveys', conference paper, Bi-annual Conference on Impact Evaluation, Colombo, Sri Lanka, 22 to 23 April 2009.

Following matching, the differences between the two groups are substantially reduced. Establishing the region of common support discards those households without access to clean water who are very dissimilar to those with access to clean water, so households in the matched comparison group are more urban, better educated and better off than households without access to clean water as a whole. Similarly, the least similar members of the treatment group have also been discarded from the evaluation.

4. **Check for balance** – The characteristics of the treatment and comparison groups are compared to test for balance. Ideally, there will be no significant differences in average observable characteristics between the two groups. Now that the treatment and comparison groups are similar on observable characteristics, variance in the incidence of child diarrhoea between the treatment and comparison groups can be attributed to differences such as access to clean water.

5. **Estimate programme effects and interpret results** – Finally, the impact estimate, either single or double difference, is calculated by firstly calculating the difference between the indicator for the treatment individual and the average value for the matched comparison individuals, and secondly averaging out all of these differences.

   Table 3 shows an example (with nearest neighbour matching) using data on learning outcomes for grade (or year) six children on a standardized test. Column 1 shows the test score for individuals from the treatment group, and columns 4 to 8 show the test score for the nearest five neighbours of each from the comparison group. The average score for the five neighbours is shown in column 2, and the difference between the treatment individual's test score and this average is shown in column 3. The single difference impact estimate is the average of the values in column 4.

Table 3. Calculation of the propensity score impact estimate: Example using test score data

Observed (i)	Y_{1i}	$Y_{0i(ave)}$	Y_{1i}-Y_{0i}	$Y_{0i(1)}$	$Y_{0i(2)}$	$Y_{0i(3)}$	$Y_{0i(4)}$	$Y_{0i(5)}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	48.2	42.4	5.8	44.1	45.1	43.8	43.2	35.8
2	50.2	42.6	7.6	42.1	45.2	48.1	38.4	39.3
3	50.6	43.1	7.5	40.8	43.7	45.3	44.1	41.8
4	48.1	38.9	9.1	43.6	35.6	36.9	41.4	37.2
5	69.0	59.7	9.3	55.6	57.6	57.1	62.4	65.8
…	…	…	…	…	…	…	…	…
199	58.6	52.2	6.4	55.5	48.2	54.7	53.4	49.1
200	45.4	39.3	6.1	41.2	39.1	38.7	40.1	37.5
Average	52.9	45.5	7.4					

In practice, these calculations do not need to be done manually as statistical packages (e.g., Stata, SAS or R) are available to conduct the analysis.

What is needed to conduct PSM?

PSM requires data from both the treatment group and a potential comparison group. Both samples must be larger than the sample size suggested by power calculations (i.e., calculations that indicate the sample size required to detect the impact of an intervention) since observations outside the region of common support are discarded. Generally, oversampling must be greater for the potential comparison group than for the treatment group.

PSM can be conducted using data from surveys, administrative records, etc. The data for the treatment and comparison groups may come from different data sets provided that: (1) they contain data on the same variables (i.e., defined in the same way); and (2) the data were collected during the same time frame. The latter requirement is particularly important for seasonal variables – that is, variables that are sensitive to the different seasons such as weight for age.

Advantages and disadvantages of PSM

The two main advantages of PSM are that it is always feasible if data are available, and it can be done after an intervention has finished, including in the absence of baseline data (although this is not ideal). If baseline data are unavailable, 'recall' can be used to reconstruct pre-intervention characteristics. This can be imprecise, however, and common sense should prevail when deciding which variables can be recalled accurately.

The main drawback is that PSM relies on matching individuals on the basis of observable characteristics linked to predicted likelihood of participation. So, if there are any 'unobserved' characteristics that affect participation and which change over time, the estimates will be biased and thus affect the observed results.

An additional practical limitation of using PSM is the need for the assistance of a statistician or someone with skills in using different statistical packages.

Regression discontinuity design (RDD)

What is RDD?

This approach can be used when there is some kind of criterion that must be met before people can participate in the intervention being evaluated. This is known as a threshold. A threshold rule determines eligibility for participation in the programme/policy and is usually based on a continuous variable assessed for all potentially eligible individuals. For example, students below a certain test score are enrolled in a remedial programme, or women above or below a certain age are eligible for participation in a health programme (e.g., women over 50 years old are eligible for free breast cancer screening).

Clearly, those above and below the threshold are different, and the threshold criterion (or criteria) may well be correlated with the outcome, resulting in selection bias. Remedial education is provided to improve learning outcomes, and therefore those with poorer learning outcomes are picked to be in the programme. Older women are more likely to get breast cancer, and it is older women who are selected for screening. So, simply comparing those in the programme with those not in the programme will bias the results.

Those just either side of the threshold are not very different, however. If the threshold for being enrolled in a remedial study programme is a test score of 60, students enrolled in the programme who get a score of 58 to 59.9 are not very different from those who get a score of 60 to 60.9 and are not enrolled. Regression discontinuity is based on a comparison of the difference in average outcomes for these two groups.

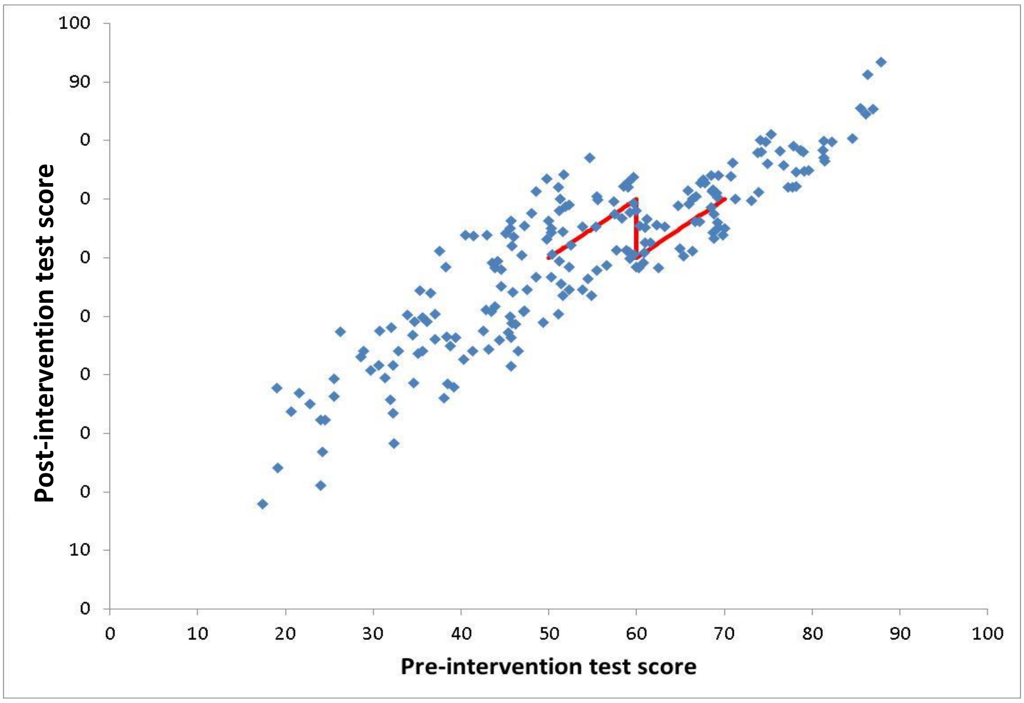
How to apply RDD

The first step is to determine the margin around the threshold and this is done using an iterative approach. At first, a small margin can be set up, and the resulting treatment and comparison groups can be tested for their balance or similarity. If the match is good, the margin can be widened a little and the balance checked again. This process must be repeated until the samples start to become dissimilar. Although balancing is based on observable characteristics, there is no reason to expect imbalance among non-observable characteristics (this is different in the case of PSM, as explained above).

Once the sample is established, a regression line is fitted. This is a line drawn through the data points that represents the 'best fit' between the variables being studied or that summarizes the 'relationship' between the selected variables – that is, when the line slopes down (from top left to bottom right) it indicates a negative or inverse relationship; when it slopes up (from bottom left to top right) a positive or direct relationship is indicated. In this case, the regression line is fitted on the selected outcome of interest (e.g., test scores). The sample for the regression is restricted to observations just either side of the threshold. Often a major challenge for RDD is the need for sufficient observations on either side of the threshold to be able to fit the regression line.

An example of a remedial education programme is shown in figure 2. The selection criterion for eligibility to participate in the programme is a pre-intervention test score, with a threshold of 60. The outcome variable is a post-intervention test score. The scatter plot shows that these two variables are, unsurprisingly, related. There is a positive relationship between pre- and post-intervention test scores. Children with a pre-intervention test score of below 60 received the remedial classes. The sample used for the analysis is taken from just either side of the threshold – those included have pre-intervention test scores in the range of 50 to 70, i.e., 10 units either side of the threshold. The fitted regression line has a 'jump'; this is the discontinuity. The size of this jump (which is 10) is the impact of the programme – that is, the remedial education programme increases test scores by 10 points on average.

Figure 2. Regression discontinuity design



What is needed for RDD?

Data are required on the selection variable and the outcome indicator for all those considered for an intervention, whether accepted or not. Many programmes do not keep information on individuals refused entry to the programme, however, which can make RDD more difficult.

Advantages and disadvantages of RDD

RDD deals with non-observable characteristics more convincingly than other quasi-experimental matching methods. It can also utilize administrative data to a large extent, thus reducing the need for data collection – although the outcome data for those not accepted into the programme often need to be collected.

The limits of the technique are that the selection criteria and/or threshold are not always clear and the sample may be insufficiently large for the analysis (as noted above). In addition, RDD yields a 'local area treatment effect'. That is, the impact estimate is valid for those close to the threshold, but the impact on those further from the threshold may be different (it could be more or less). In practice, however, where it has been possible to compare this 'local' effect with the 'average' effect, the differences have not been great. This indicates that RDD is an acceptable method for estimating the effects of a programme or policy.

Epidemiological approaches

Epidemiologists apply a range of statistical data collected from treated and untreated populations, including ordinary least squares and logistic regressions in the case of dichotomous outcomes (having the condition = 1, not having the condition = 0). When using these methods, it is preferable to: (1) use data from well matched treatment and comparison groups, and (2) restrict the regression analysis to observations from the region of common support. (These steps are not usually taken at present, however.)

Some epidemiological studies present the difference in means between treated and untreated observations, but this approach does not take account of possible selection bias.

Survival analysis can be an appropriate approach when data are censored, meaning that the period of exposure is incomplete because of the timing of data collection or the death of the study participant. The Cox proportional hazards model is commonly used in such circumstances.

4. QUASI-EXPERIMENTAL METHODS FOR DATA ANALYSIS

Single difference impact estimates

Single difference impact estimates compare the outcomes in the treatment group with the outcomes in the comparison group at a single point in time following the intervention (t+1, in table 1).

Difference-in-differences

What is 'difference-in-differences'?

Difference-in-differences (DID), also known as the 'double difference' method, compares the changes in outcome *over time* between treatment and comparison groups to estimate impact.

DID gives a stronger impact estimate than single difference, which only compares the difference in outcomes between treatment and comparison groups following the intervention (at t+1). Applying the DID method removes the difference in the outcome between treatment and comparison groups at the baseline. Nonetheless, this method is best used in conjunction with other matching methods such as PSM or RDD. If DID is used without matching, the researchers should test the 'parallel trends assumption', i.e., that the trend in outcomes in treatment and comparison areas was similar before the intervention.

Below is a hypothetical example of the DID method. Table 4 shows data for nutritional status, as measured by weight-for-age z-scores (WAZ), for treatment and comparison groups before and after a programme of nutritional supplementation.

Table 4. Child nutritional status (WAZ) for treatment and comparison groups at baseline and endline

	Baseline	Endline	Change
Treatment (Y_1)	-0.66	-0.48	+0.18
Comparison (Y_0)	-0.62	-0.58	+0.04
Difference		+0.10	+0.14

The magnitude of impact estimated by the single and double difference methods is very different. The single difference (SD) estimate is difference in WAZ between treatment and comparison groups following the intervention, that is, SD = -0.48 – (-0.58) = 0.10. The DID estimate is the difference in WAZ of the treatment group at the baseline and following the intervention minus the difference in WAZ of the comparison group at the baseline and following the intervention, that is, DID = [-0.48 – (-0.66)] – [-0.58 – (-0.62)] = 0.18 – 0.04 = 0.14.

The double difference estimate is greater than the single difference estimate since the comparison group had better WAZ than the treatment group at the baseline. DID allows the initial difference in WAZ between treatment and comparison groups to be removed; single difference does not do this, and so in this example resulted in an underestimate of programme impact.

How to apply the DID method

The first step involves identifying the indicators of interest (outcomes and impacts) to be measured relevant to the intervention being evaluated. Following this, the differences in indicator values from before and after the intervention for the treatment group are compared with the differences in the same values for the comparison group. For example, in order to identify the effects of a free food scheme on the nutritional status of children, the mean difference for both the treatment group and the comparison group would be calculated and then the difference between the two examined, i.e., by looking at the difference between changes in the nutrition status of children who participated in the intervention compared to those who did not. Ideally, the intervention and comparison groups will have been matched on key characteristics using PSM, as described above, to ensure that they are otherwise as similar as possible.

Advantages and disadvantages of the DID method

The major limitation of the DID method is that it is based on the assumption that the indicators of interest follow the same trajectory over time in treatment and comparison groups. This assumption is known as the 'parallel trends assumption'. Where this assumption is correct, a programme impact estimate made using this method would be unbiased. If there are differences between the groups that change over time, however, then this method will not help to eliminate these differences.

In the example above, if the comparison states experienced some changes that affect the nutritional status of children – following the start of the free food scheme in other states – then the use of DID alone would not provide an accurate assessment of the impact. (Such changes might occur, for example, because of development programmes that raise the income levels of residents, meaning they can afford to give their children a more nutritious diet.)

In summary, DID is a good approach to calculating a quantitative impact estimate, but this method alone is not usually enough to address selection bias. Taking care of selection bias requires matching to ensure that treatment and comparison groups are as alike as possible.

Regression-based methods of estimating single and double difference impact estimate

Single and double difference impact estimates may also be estimated using ordinary least squares regression. This approach is applied to the same matched data, including a programme or policy dummy variable on the right-hand side of the regression equation. Variables that capture other confounding factors can also be included on the right-hand side to eliminate the remaining effect of any discrepancies in these variables between treatment and comparison areas on the outcomes after matching.

5. ETHICAL ISSUES AND PRACTICAL LIMITATIONS

Ethical issues

Quasi-experimental methods offer practical options for conducting impact evaluations in real world settings. By using pre-existing or self-selected groups such as individuals who are already participating in a programme, these methods avoid the ethical concerns that are associated with random assignment – for example, the withholding or delaying of a potentially effective treatment or the provision of a less effective treatment for one group of study participants (see Brief No. 7, Randomized Controlled Trials).

Practical limitations

The lack of good quality data is often a key barrier to using quasi-experimental methods. Any method that is applied after a programme or policy has already finished may suffer substantially from the lack of baseline data.

Because quasi-experimental methods are based on certain assumptions (see box 1), conclusions made about causality on the basis of such studies are less definitive than those elicited by a well conducted randomized controlled trial (RCT). In most cases, however, if done well and presented clearly (i.e., making explicit the limitations and how they affect the results), quasi-experimental methods are generally well accepted by decision makers.

6. WHICH OTHER METHODS WORK WELL WITH THIS ONE?

As highlighted above, it is advisable to use different quasi-experimental methods together – for example, DID can be combined with PSM. It is recommended that qualitative methods are used in conjunction with quasi-experimental methods to gain better insights into ‘why’ a programme or policy has worked or not.

7. PRESENTATION OF RESULTS AND ANALYSIS

When writing up results based on a quasi-experimental evaluation, it is important to provide details about the specific methodology, including data collection. Since the success of these methods depends greatly on the quality of data collected (or already available), some sort of assurance of quality should be provided. It is also important to provide information about the tenability of the assumptions on which these methods are based. Although some of the assumptions cannot be tested directly (e.g., parallel trends assumptions) authors should provide clear reasoning as to why they believe these assumptions hold.

It is recommended that the description of the methodology includes details of the sampling method as well as the approach to the construction of treatment and comparison groups (including the number of individuals, households or clusters involved). The results can be analysed and reported for the entire sample as well as for important (predefined) subgroups (e.g., by age or by sex) to identify and discuss any differential effects. The findings then need to be linked to the theory of change (see Brief No. 2, Theory of Change) and used to answer the key evaluation questions (KEQs) – for example, do the findings support the theory of change? If not, which assumption behind the theory of change was not fulfilled?

These types of analyses can help evaluators to identify concrete programme or policy recommendations, which should make up the conclusion of the report. In most cases, it would also be useful to include a discussion around whether and to what extent the results can be extrapolated to different settings. Conclusions drawn from quasi-experimental designs are causally valid as long as the assumptions regarding the particular matching method are met. The quality of the match should also be tested and reported.

8. EXAMPLE OF GOOD PRACTICES

A UNICEF example using DID and PSM

UNICEF undertook an impact evaluation of Chile Solidario,[3] a conditional cash transfer (CCT) programme in Chile that sought to improve several socio-economic outcomes of families living in poverty. To evaluate the impact of this programme, the authors used data from two rounds of a national panel survey (CASEN survey).

The authors used DID and PSM to obtain an unbiased estimate of the impact. Individuals who received the conditional cash transfer were different in certain key respects to those who did not receive the transfer, so the impact evaluation design had to address the issue of selection bias. For example, programme recipients were poorer, less educated and had worse living conditions than non-recipients.

The use of PSM enabled the construction of a comparison group comprising individuals similar to the treatment group individuals on most observable characteristics. The authors estimated the participation equation by including pre-intervention values of household market income, the greatest number of years of education among individuals within the household and the number of children (under 14 years old) within the household, plus three characteristics of the domestic environment (i.e., water supply, roof condition and number of people per room) and region of residence to account for area-specific factors that may have affected participation rates. Treatment households were matched with the nearest four non-programme ‘neighbours’ (i.e., those with the closest propensity scores). The authors did not provide any tables or figures showing the quality of the match, however, which represents a shortcoming in transparency of the presentation.

A range of single and double difference impact estimates were calculated, which showed that the conditional cash transfer had a significant impact in lifting families out of extreme poverty. In addition, this study also found that the programme contributed to increased participation in school for children aged 6 to 15 years old as well as to this subgroup’s increased enrolment with public health services.

9. EXAMPLES OF CHALLENGES

The largest potential pitfall in quasi-experimental methods is the risk of obtaining a poor quality match. The comparison group needs to be as similar as possible to the treatment group before the intervention. Checking the quality of the match, by reporting balance tables for determinants of the outcomes of interest and the outcomes themselves, is thus very important.

Another potential pitfall lies in the tendency to focus on statistically significant findings to the detriment of statistically insignificant results. All results should be reported and the discussion should not focus unduly only on those that are statistically significant.

In reporting the findings it is important to discuss the size of the effect as well as its significance. This is usually overlooked, even though statistical significance alone is not necessarily enough for an intervention to be of interest to policymakers, or for it to be a cost-effective option. There must also be evidence that a sufficiently large effect has occurred.

All quantitative studies rely on the data being of a sufficiently high quality, so data quality checks should be performed.

[3] Martorano, Bruno and Marco Sanfilippo, ‘Innovative Features in Conditional Cash Transfers: An impact evaluation of Chile Solidario on households and children’, *Innocenti Working Paper* No. 2012-03, UNICEF Innocenti Research Centre, Florence, 2012. See http://www.unicef-irc.org/publications/pdf/iwp_2012_03.pdf.

10. KEY READINGS AND LINKS

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, New Jersey, 2009, pp. 227–259.

Caliendo, Marco and Sabine Kopeinig. 'Some Practical Guidance for the Implementation of Propensity Score Matching', *IZA Discussion Paper* No. 1588, Institute for the Study of Labor (IZA), Bonn, 2005. See http://ftp.iza.org/dp1588.pdf.

Gertler, Paul J., et al., *Impact Evaluation in Practice*, World Bank, Washington, D.C., 2010, pp. 81–116. See http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf.

Kaplan, Josiah, 'Propensity Scores', web page, BetterEvaluation, 2011, http://betterevaluation.org/evaluation-options/propensity_scores.

Khandker, Shahidur R., et al., *Handbook on Impact Evaluation: Quantitative Methods and Practices,* World Bank, Washington, D.C., 2010, pp. 53–103. See http://bit.ly/1d2Ve8m.

Lee, David S. and Thomas Lemieux, 'Regression discontinuity designs in economics', *NBER Working Paper* No. 14723, National Bureau of Economic Research, Cambridge, MA, 2009. See http://www.nber.org/papers/w14723.pdf?new_window=1.

Ravallion, Martin, 'Assessing the Poverty Impact of an Assigned Program', in F. Bourguignon and L.A. Pereira da Silva (eds.), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, *Volume 1*, Oxford University Press, New York, 2003. See http://origin-www.unicef.org/socialpolicy/files/Assessing_the_Poverty_Impact_of_an_Assigned_Programme.pdf.

Martorano, Bruno and Marco Sanfilippo, 'Innovative Features in Conditional Cash Transfers: An impact evaluation of Chile Solidario on households and children', *Innocenti Working Paper* No. 2012-03, UNICEF Innocenti Research Centre, Florence, 2012. See http://www.unicef-irc.org/publications/pdf/iwp_2012_03.pdf.

Qasim, Qursum and 3ie, 'Regression Discontinuity', web page, BetterEvaluation, 2011, http://betterevaluation.org/evaluation-options/regressiondiscontinuity.

Shadish, William R., et al., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, 2002, pp. 103–243.

GLOSSARY

Baseline study	*An analysis describing the situation prior to an intervention, against which progress can be assessed or comparisons made. (OECD-DAC definition, 2010)*
Comparison group	*In a quasi-experimental research design, this is the group of research participants/subjects that, for the sake of comparison, does not receive the treatment/intervention given to the treatment/intervention group. Comparison group subjects are typically not randomly assigned to their condition, as would be true of control group subjects in an experimental design study. See: control group, treatment group.*
Cox proportional hazards model	*A statistical/modelling technique for exploring the survival of a patient and a number of exploratory variables. (Definition from www.whatisseries.co.uk)*
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.*
Instrumental variable estimation	*A statistical technique for estimating causal relationships when an RCT is not feasible or when an intervention does not reach every participant/unit in an RCT.*
Logistic regression	*A regression technique which estimates the probability of an event occurring. (University of Strathclyde definition) See: regression.*
Ordinary least squares (OLS) regression	*A generalized linear modelling technique that may be used to model a single response variable which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been appropriately coded. (Hutcheson's definition, 2011)*
Randomized controlled trials (RCTs)	*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.*
Regression	*A statistical procedure for predicting values of a dependent variable based on the values of one or more independent variables. Ordinary regression uses ordinary least squares to find a best fitting line and comes up with coefficients that predict the change in the dependent variable for one unit change in the independent variable. (University of Strathclyde definition)*
Treatment group	*Subjects/participants exposed to the independent variable; also called the experimental or intervention group.*