Number 04 July 2013

Centre for Development Impact
PRACTICE PAPER
Innovation and learning in impact evaluation

Impact Evaluation when N=1

Abstract A common presumption holds that when there is only one unit of observation, such as in the case of a national-level policy or a small scale intervention, causality cannot be established and impact evaluation methods do not apply. Yet many development interventions have single communities or organisations as their target, just as in other cases we are interested in the impact of a programme in a particular community, not for the *average* community. The experimental observation of single subjects or groups is normally practiced in sciences such as psychology and biology. Social sciences do not enjoy the controlled environments that allow conducting single-case experiments. However, this CDI Practice Paper by Edoardo Masset shows that, under some circumstances, methods to assess impact when there is only one unit of observation *are* possible, and that we should try to create the right conditions for these 'experiments' to take place rather than neglect them.

Introduction

As time passes we forget events. Yet forgetting itself follows a curve: events are easily forgotten right after occurring but the rate of forgetting slows down over time. The first experiment in the study of forgetting was conducted by Hermann Ebbinghaus (Schacter 1996). In 1885 Ebbinghaus became his own and sole subject, by starting to memorise long lists of three letter syllables. Ebbinghaus then tested his ability to remember the syllables at different times, and discovered that he forgot a large amount between a one-hour and a nine-hour delay, but much less between a one-day and a two-day delay. Nearly one hundred years later, Waagenar (1986) began to record a salient event per day in a diary, including a number of cues: what had happened, when it happened, who was involved, where and when. He collected 2,400 events in this way over a period of four years. In the fifth year he started experimenting with his own memory. Waagenar chose one of the cues at random and tried to remember the event. He found that he would remember pleasant events more easily compared to unpleasant ones; that the *when* cue helped very little compared to the *what* and *where* cues; and that no event was completely forgotten. These studies have some practical implications on how we should design surveys. But what is truly remarkable about these studies is that both Ebbinghaus and Waagenar explained how memory works, and employed *statistical* methods to test their findings by observing just one subject, without employing a control group.

This paper discusses what can be learned about the effectiveness of development projects by observing a single unit over time without a control group. There are several cases in which there is only one unit of intervention and no control groups. There are also cases in which there are multiple units of observation, but we are interested in the impact of just one of them. This paper offers examples of possible methods of investigation in such cases, drawing from both old and recent literature on experimental methods.

N=1 projects

Standard impact evaluation methods are based on counterfactual analysis using data collected from a large number of observations both with and without the intervention. Normally, a survey is conducted of households, schools or clinics before and after the intervention. The size of the sample of units surveyed is set in such a way as to allow statistical testing of the expected impact of the intervention. If conditions allow, selection bias is prevented by randomly assigning units to the intervention. A good impact evaluation study relies on data from a large number of units and on the establishment of a valid control group. The result of the evaluation is the estimation of an average effect of the intervention that can be used to predict the impact of the same, or similar, interventions in other areas or contexts.

There are two cases in which this approach cannot be adopted. In the first case, the unit of intervention and of

observation is either too small or too large to allow statistical testing, or even the construction of a control group. In the second case, the *average* impact is simply not interesting, since the evaluator is concerned with the impact of the intervention on a *specific* unit. These two cases will be discussed in turn.

The author recently conducted a scoping study to assess the feasibility of adopting experimental methods for the assessment of a number of projects in the area of governance and accountability in Malawi (the 'Tilitonse' project). In the majority of cases implementing a randomised controlled trial would not be possible. In some cases the project promotes social accountability within a single community, town or city, whilst in others, the project strengthens the institutional capacity of just one civil society organisation or of a local government unit. In some other cases the project reaches a small population with special needs, such as sex workers or prisoners. In all these cases the size of the intervention was too small (often equal to one) to allow for large samples (including within the unit), let alone control groups and statistical testing.

Other times the scale of the intervention was too large. These projects were, for example, promoting policy change at the national level by advocacy or capacity building initiatives. When an intervention is implemented at the national level the unit of observation is again equal to one. Some national level interventions can be tested at a local level before being scaled up, which can be accomplished in imaginative ways. For example, the 'last mile project' employs an ingenious system of vouchers to test the impact of electrification infrastructure in Ethiopia (Bernard and Torero 2011). Yet in other cases, such as, for example, with the impact of trade reforms, this is not really possible. In other cases this is possible but not desirable. When there are sizable scale economies, that is when the impact of the intervention varies with the size of the area or population covered, the average treatment effect obtained through RCTs is no longer a fair prediction of the effect of the same intervention at the national level (Manski 2013).

In some other cases experimental approaches such as RCTs can be applied, but we are not at all interested in the average effect of the intervention. On the contrary, we are interested in the impact of the intervention on a specific unit. RCTs produce 'average' results. They are an efficient way of dealing with selection bias but they are designed with the goal of improving the living conditions of a general population, not of any specific individuals. Yet under any intervention, while some individuals benefit, others do not, and can even be harmed – a fact which is hidden by the use of averages. The average effect may have no application to individual cases and may not explain changes occurring in individuals. We may not be interested in what action would make our society better off, but what would make a specific person, community or organisation better off. In these cases the evidence produced by RCTs will be of little use.

Examples of N=1 methods

The 'one-shot case study' in which a single individual or group is studied once is given little scientific credence due to the lack of comparison or control group (Campbell and Stanley 1963). In the single case design the possible alternative explanations to the impact of the intervention are so many and the specificity of the case so high that it provides very little useful information.

Much can be achieved, however, by expanding the observation of a single individual or group *over time*. In the interrupted time-series design the observed group is, at different times, both the project and the control group. This design has a number of limitations but does have value in some applications. A considerable improvement to the interrupted time series design is the alternating treatment design. In alternating treatments the group is intervention and control at different times but in a random fashion, thus generating a comparison not too different from standard intervention-control design.

The one-shot design cannot produce evidence of impact but can be a valid method for disproving untested theories or project designs. This section will elaborate on the use of three N=1 methodologies: (1) the study of anomalies; (2) interrupted time series; and (3) alternating treatment designs.

Anomalies

Important discoveries can be made by spotting anomalies and trying to understand them. Freedman (2008) showed the relevance of anomalies in scientific discovery with examples from the history of medicine. For example, in 1941 the ophthalmologist Norman Gregg noticed in his practice an unusually large number of infants with cataract and birth defects which could not be attributed to a genetic cause – the conventional explanation at the time for birth defects. Further investigations led to the discovery of German measles. In another example, the epidemiologist Joseph Goldberger noticed that in hospitals and asylums the patients developed pellagra, but the attendants almost never did. This clashed with the traditional explanation at the time that pellagra was transmitted from person to person by insects. Further research showed that pellagra was caused by diets lacking vitamin B.

These examples show that a theory can be conclusively rejected by observation of a single anomaly. Rogowsky (2004) provides a number of examples from comparative politics in which single-case studies disconfirmed theories that were previously highly regarded. First, a study of

politics and society in just one country (the Netherlands) showed that the theory of 'cross-cutting cleavages' – predicting that mutual reinforcing cleavages (the overlapping of religion, social class and language) would increase conflict – was not necessarily true. The Netherlands is a peaceful country with virtually no overlap among social groups. Second, the study of the vibrant associational life of a single German town where Nazism prospered, disconfirmed the theory that Fascism would flourish in the absence of social life and organisations. Third, the development of the 'peripheral' Prussian state, disconfirmed Wallerstein's theory that core states of the world economy were the more likely to become strong.

Anomalies follow the principle of falsification. A single observation cannot lead to any universal theory, but any universal theory can be disconfirmed by a single observation. Blaug (1992) illustrates the point with a Popperian example: no amount of observations of white swans will prove that swans are white, but the observation of a single black swan is enough to refute that all swans are white. In other words, it can never be conclusively shown that a theory is true, but a single observation can show a theory is false.

It can be argued that impact evaluations do not test theories, but they just collect evidence on a particular issue. Yet there is a large literature on theory-based evaluation arguing that evaluations are testing, or should test, the theory behind each intervention (Funnell and Rogers 2011; Weiss 1972; White 2009). However, failures in implementation often imply that theory-based evaluations can explain why an intervention did not work but cannot test the theory informing the intervention. One example of this is the evaluation of the Bangladeshi Integrated Nutrition Project (White and Masset 2007). The programme was based on growth monitoring and promotion (GMP), which consists of weight monitoring of child growth and the provision of counselling and supplementary feeding. GMP is based on the theory that physical growth in the first two years of age follows a standardised curve, and that deviations from the pattern of growth can be corrected with appropriate supplementation of medications and food. GMP is widely used. The World Bank and other organisations have implemented projects based on GMP in many poor countries. The effectiveness of GMP intervention, however, is unknown – as shown by a Cochrane review (Panpanich and Garner 2009). There are theories alternative to GMP maintaining that child growth is non-linear and occurs in spurts (see for example Lample, Veldbuis, and Johnson 1992), in which case GMP would be ineffective if not harmful.

The study of anomalies could be used to disprove GMP. This could be conducted, for example, by daily monitoring of the application of GMP over a two month period on one child, or a group of children, in one community. The focus on one community allows the study to be implemented in controlled conditions so that any impact can be attributed to GMP rather than something else. This approach cannot prove the validity of GMP, but can *disprove* its validity when implemented in ideal conditions. It is a useful tool for those cases in which untested theory dominates, and when that theory cannot be evaluated through field trials due to the complexity and interference of contextual factors.

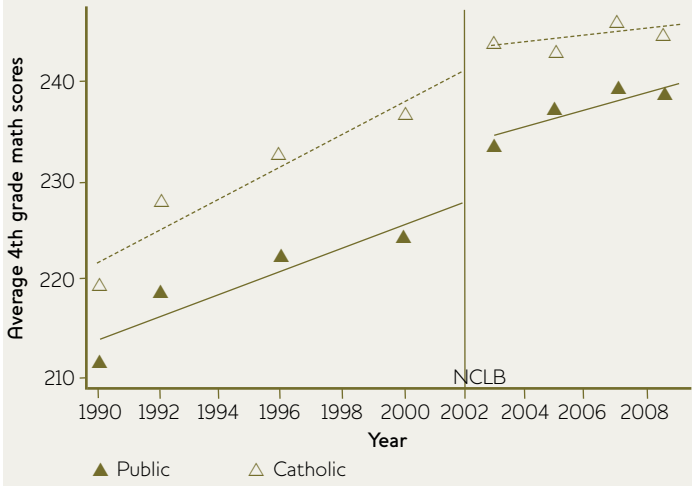
Interrupted time series

The interrupted time series design finds application when we have data of an outcome variable for an individual or for a group over a long period before and after an intervention. The impact is observed as an interruption or a discontinuity at the point when the intervention begins. This design typified much of classical nineteenth-century physical science and is considered 'experimental' by these disciplines (Campbell and Stanley 1963). The interrupted time series design is not widely used since it is very difficult to control any other time-varying factors, and so to uniquely attribute the observed change to the intervention.

The analysis of the impact of the No Child Left Behind programme (NCLB) is an example of the application of this method. The NCLB programme was officially enacted in the United States in January 2002 and holds elementary and middle schools accountable by monitoring progress towards meeting state-specific standards in reading and maths. Schools failing to meet the targets are required to take corrective actions that become increasingly onerous. Failing the first year requires the school to inform the parents. Failing the second year requires giving parents the option to transfer the child to a better school. Failing the third year requires the provision of supplementary services. Failing the fourth year requires the replacement of staff and hiring consultants. Failing the fifth year requires the school to close down.

The programme applies to all schools accepting federal funds and is therefore national in scope. Wong *et al.* (2009) used National Assessment Progress Data for the period 1990–2009 for fourth grade reading, as well as fourth and eighth grade maths tests to assess the impact of the programme. They found that the programme had only a moderate impact on the reading scores of fourth – grade students, but the impact on maths scores was substantial, and visible, as shown in Figure 1 for the fourth grade students. In comparison with Catholic schools (which, not being recipient of federal funds served in this case as a control group), Wong *et al.* found an increase of 0.30 standard deviations in maths scores corresponding to a gain of six/seven months of school. The differences are statistically significant.

Figure 1 **Main NAEP 4th grade math scores by year: Public and Catholic schools**



Source: Wong *et al.* (2009)

This example shows the main advantage of an interrupted time series design. The methodology does not require sophisticated statistical methods. Impact can be displayed in a chart in a way that is credible and easily understood. This design, however, has a number of limitations. First, and most obviously, the observed change may be the result of other factors changing over time. The study is more valid when the environment in which it takes place is more controlled. Second, the design can be affected by changes in measurement. For example, the introduction of a new (yet ineffective) vaccine might lead doctors to incorrectly classify disease occurrences since they believe patients have been vaccinated. As such, people affected by the disease are misclassified under other diseases. In this case, the observed reduction in the incidence of the disease is the result of a change in reporting, not a change in the occurrence of the disease. Third, there are difficulties in designing statistical tests – particularly when the number of observations is small. Fourth, the observed change has to be large in order to be detected, which implies that small but important effects may go unnoticed. Finally, a much discussed selection problem arises if there is a change in the composition of the observed group prior to the treatment and because of the treatment. In this instance the treatment effects may be confused with the effect of characteristics correlated with the treatment. This is also known as the Ashenfelter's dip problem, after a study on the impact of labour training programmes on earnings (Ashenfelter 1978).

Despite all these limitations, the main difficulty in the application of the interrupted time series design is the availability of data. Any researcher would like to track the aggregate trends of outcome indicators before and after an intervention, even if only for illustrative purposes. Unfortunately this type of data are rarely available (Cook and Campbell 1979). Monitoring data normally cover only the period of the intervention, sometimes with one baseline observation, but rarely allow observing the trends prior to the intervention.

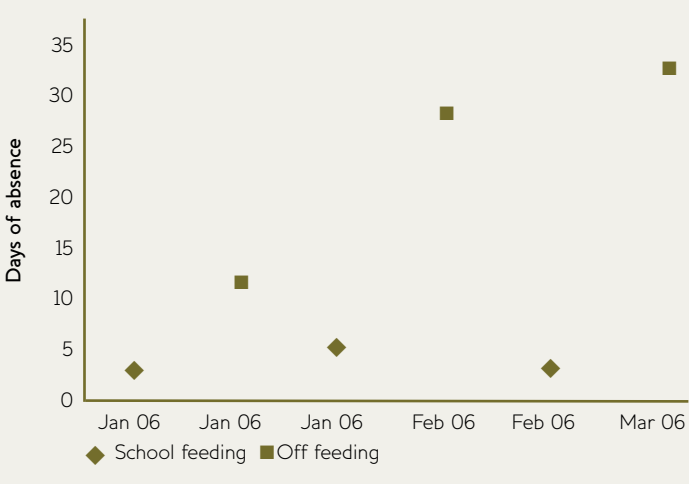
Alternating treatments

The interrupted time series design can be considerably improved by including a comparison group. This can be accomplished in the single-case design if the unit observed acts as its own control group. This type of design has been widely used in mental health studies (Barlow and Hersen 1984). One popular approach is the withdrawal design, in which a treatment is withdrawn after follow up. The subject is observed at the baseline (no treatment) at the follow-up (after treatment) and at the endline (after withdrawal of the treatment). If the unit observed reverts to its original condition after removing the treatment, any change observed between baseline and follow-up can be safely be attributed to the treatment. Problems with this type of design include: (1) the experiment is unethical unless the treatment is eventually given after the first withdrawal; (2) impact of the treatment and of the withdrawal may be confounded by other factors varying over time and affecting the subject; (3) there might be learning by the subject over time that changes the response to the treatment.

There are a number of extensions to the standard withdrawal design which include (Barlow and Hersen 1984): (1) repeating the introduction and removal of the treatment several times; (2) comparing many treatments for the same subject over time; (3) varying the intensity of the treatment; (4) evaluating interaction effects between treatments. Finally, a particularly powerful variation of this design is the alternating (or simultaneous) treatment design, whereby the same treatment is randomly assigned over time to the same unit of observation. For example, a clinic could implement a procedure following the random series 01001110101101 on a daily basis (where 1 is one treatment type and 0 is another treatment type or no treatment). An institution could run a service in two or more different ways on a weekly or monthly basis following a random series and then compare the outcomes of the treatments after one year by plotting the point of each treatment on a chart. Given the randomness of the assignment this design is immune to the influence of time-varying factors. This method should be implemented with caution because the same ethical problems related to the implementation of RCTs also apply in this case.

For illustrative purposes, Figure 2 uses fictional data to show the impact of school feeding and non school feeding (in periods of two weeks randomly assigned over three months) on days of absence. The treatment and no-

Figure 2 **School feeding and days of absence in one primary school**



treatment conditions are compared for the same school. The chart above provides evidence of impact. Note that the goal of this analysis is not generalising the findings to other schools, but assessing whether the programme is having an impact on this particular school. The great advantage of this design is that it can be conducted at a small scale and over a short period of time.

Conclusions

This paper has reviewed three methods (the study of anomalies, interrupted time series and alternating treatments) which could be employed in the impact assessment of development interventions when there is only one unit of observation – be it a community, an organisation or an entire country. Each of these methods rely not on observing 1,000 households for one hour or observing 100 clinics for 2 hours, but on observing one school for 100 days or one community for 1,000 hours. What is lost in variability observing only one unit can be compensated by observing that same unit many times. They are not interested in the *average*, but the *specific*.

The study of anomalies is particularly useful in disconfirming theories that underpin development interventions but are mostly untested. The observation of a single case, the neglected 'one-shot case study', is not able to provide evidence of impact, but is potentially able to disconfirm existing theories. There are several untested or poorly tested theories in development. Growth monitoring and promotion (GMP) of children is one example. The theory, originally proposed by Fisher (1930), that poor people are unable to save because they're 'impatient' – which underpins the reasoning behind many microfinance programmes – is another example. Testing these theories through field trials is rarely possible because the impact of the treatment is obfuscated by the complexity of the intervention. For instance, one application of this method would be testing the validity of incentive schemes, such as in the case of Tilitonse projects mentioned earlier. Social accountability interventions such as the use of scorecards are based on the assumption that a system of rewards and punishment can increase the efficiency of organisations. There is a long tradition, particularly in economics, in favour of this line of thought. However, recent research in psychology (Deci, Koestner, and Ryan 1999) and economics (Rebitzer and Taylor 2013) has shown that incentives may in fact have a *negative* impact on productivity. Testing the validity of the mainstream efficiency theory of incentives by closely observing staff behaviours under different conditions would make an interesting contribution to the study of social accountability programmes.

In those cases where a project is implemented at a national level or in a single community or organisation, the interrupted time series design is particularly appropriate. Elements and appropriate statistical tools of the interrupted time series design are well established (Cook and Campbell 1979). What is missing is the data. Many development interventions aim to improve indicators such as people awareness, institutional capacity of organisations and local governments, or quality of services provided by public authorities. These outcomes can only be observed at the national level or at the level of a single community or organisation. In all these cases, an impact assessment would be possible if we had several observations of the same unit before the intervention. This suggests that evaluators should start collecting data in project areas well before an intervention begins in order to allow a credible analysis of trends.

Finally, a promising and unutilised design is alternating treatment design. The alternating treatment design is a powerful tool for assessing impact with a good level of confidence, a limited number of observations and few resources. This seems a valuable tool for testing the effectiveness of alternative implementation strategies. For example, it could be used for testing the comparative advantages and effectiveness of different typologies of service delivery or capacity building, including the case of no-service (the 'control group'), within the same organisation or local authority. Again, this method has the advantage of being implemented at a small scale, relying on the observation of a single unit and being carried out over a short period of time.

References

Ashenfelter, O. (1978) 'Estimating the Effect of Training Programs on Earnings', *The Review of Economics and Statistics* 60.1: 47–57

Barlow, D.R. and Hersen, M. (1984) *Single-Case Experimental Design*, New York: Pergamon Press

Bernard, T. and Torero, M. (2011) *Randomizing the 'Last Mile'*, IFPRI Discussion Paper 01078

Blaug, M. (1992) *The Methodology of Economics: How Economists Explain*, Cambridge: Cambridge University Press

Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research*, Chicago: Rand McNally College Publishing Company

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis for Field Settings*, Boston: Houghton Mifflin Company

Deci, E.L.; Koestner, R. and Ryan, R.M. (1999) 'A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation', *Psychological Bulletin* 125.6: 627–68

Fisher, I. (1930) *The Theory of Interest, as Determined by the Impatience to Spend Income and Opportunity to Invest It*, New York: Macmillan

Freedman, D. (2008) 'On Types of Scientify Inquiry: The Role of Qualitative Reasoning', in J.M. Box-Steffensmeier, H.E. Brady and D. Collier (eds), *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press

Funnell, S.C. and Rogers, P.J. (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*, San Francisco CA: Wiley

Lample, M.; Veldbuis, J.D. and Johnson, M.L. (1992) 'Saltation and Stasis: A Model of Human Growth', *Science* 258: 801–03

Manski, C.F. (2013) 'Identification of Treatment Responses with Social Interactions', *The Econometrics Journal* 16.1: S1–23

Panpanich, R. and Garner, P. (2009) 'Growth Monitoring in Children', *Cochrane Systematic Review*

Rebitzer, J.B. and Taylor, L.J. (2013) 'Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets', in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics 4A*, San Diego CA: North Holland

Rogowsky, R. (2004) 'How Inference in the Social (but Not the Physical) Sciences Neglects Theoretical Anomaly', in H.E. Brady and D. Collier (eds), *Rethinking Social Inquiry*, Lanham MD: The Rowman and Littlefield Publishing Group, Inc.

Schacter, D.L. (1996) *Searching for Memory: The Brain, the Mind and the Past*, New York: Basic Books

Waagenar, W.A. (1986) 'Memory: A Study of Autobiographical Memory over Six Years', *Cognitive Psychology* 18: 225–52

Weiss, C.H. (1972) *Evaluation Research: Methods of Assessing Program Effectiveness*, Upper Saddle River NJ: Prentice Hall

White, H. (2009) 'Theory-based Impact Evaluation: Principles and Practice', *Journal of Development Effectiveness* 1.3: 271–84

White, H. and Masset, E. (2007) 'Assessing Interventions to Improve Child Nutrition: A Theory-Based Impact Evaluation of the Bangladesh Integrated Nutrition Project', *Journal of International Development* 19.5: 627–52

Wong, M.; Cook, T.D. and Steiner, P.M. (2009) *No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each With Its Own Non-Equivalent Comparison Series*, Institute for Policy Research Northwestern University Working Paper Series WP-09-11

> These methods rely not on observing 1,000 households for one hour or observing 100 clinics for two hours, but on observing one school for 100 days or one community for 1,000 hours. What is lost in variability observing only one unit can be compensated by observing that same unit many times. They are not interested in the *average*, but the *specific*.

Centre for Development Impact (CDI)

The Centre is a collaboration between IDS (www.ids.ac.uk) and ITAD (www.itad.com).

The Centre aims to contribute to innovation and excellence in the areas of impact assessment, evaluation and learning in development. The Centre's work is presently focused on:

(1) Exploring a broader range of evaluation designs and methods, and approaches to causal inference.
(2) Designing appropriate ways to assess the impact of complex interventions in challenging contexts.
(3) Better understanding the political dynamics and other factors in the evaluation process, including the use of evaluation evidence.

This CDI Practice Paper was written by **Edoardo Masset**.

ISSN: 2053-0536
AG Level 2 Output ID: 303

Institute of Development Studies, Brighton BN1 9RE, UK
T +44 (0) 1273 915637 **F** +44 (0) 1273 621202 **E** ids@ids.ac.uk **W** www.ids.ac.uk