HM TREASURY

Department of Energy & Climate Change

Department for Environment Food & Rural Affairs

Quality in policy impact evaluation:

understanding the effects of policy from other influences (supplementary Magenta Book guidance)

December 2012

HM TREASURY

Department of Energy & Climate Change

Department for Environment Food & Rural Affairs

Quality in policy impact evaluation:

understanding the effects of policy from other influences (supplementary Magenta Book guidance)

Authors: Siobhan Campbell
Gemma Harper

December 2012

Official versions of this document are printed on 100% recycled paper. When you have finished with it please recycle it again.

If using an electronic version of the document, please consider the environment and only print the pages which you need and recycle them when you have finished.

Any queries regarding this publication should be sent to us at: public.enquiries@hm-treasury.gov.uk.

ISBN 978-1-909096-49-3
PU1422

Contents

		Page
Executive summary		3
Chapter 1	Introduction	5
Chapter 2	Quality in policy impact evaluation	9
Chapter 3	Strong research designs in the measurement of attribution	11
Chapter 4	Weaker/riskier research designs in the measurement of attribution	17
Annex A	Acknowledgements	23
Annex B	References	25

Executive summary

Quality in policy impact evaluation (QPIE) is a supplement to the Magenta Book and provides a guide to the quality of impact evaluation designs. It has been developed to aid policy makers and analysts understand and make choices about the main impact evaluation designs by understanding their pros and cons and how well each design can allow for any measured change to be attributed to the policy intervention being investigated (see table overleaf).

	Brief Description	Ability to establish attribution
Strong research designs in the measurement of attribution	Random allocation/experimental design. Individuals or groups are randomly assigned to either the policy intervention or non-intervention (control) group and the outcomes of interest are compared. There are many methods of randomisation from field experiments to randomised control trials.	Random allocation design means that systematic differences between groups are unlikely and so any differences and changes in outcomes between the two groups can be confidently attributed to the policy intervention.
	Quasi-experimental designs *Intervention group vs well matched counterfactual*. Outcomes of interest are compared between the intervention group and a comparison group directly matched to the intervention group on factors known to be relevant to the outcome. Done well, the matched comparison group can be treated as though it was created randomly. *Strong difference-in-difference design*. In this quasi-experimental design there is no direct matching. Instead it involves a before and after study comparing two groups where there is strong evidence that outcomes for the two groups have historically moved in parallel over time.	Quasi-experimental designs match the groups on relevant factors, i.e. factors which could have an impact on the measured outcomes. If the matching is done well, any differences between the two groups can be concluded to be the result of the policy intervention (as there are no other observable differences between the two groups). A strong dif-in-dif design can provide good evidence on what would have happened in the absence of a policy intervention, and therefore allows a strong assessment of the impact of the policy.
Weaker/riskier research designs in the measurement of attribution	Intervention group vs unmatched comparison group. Outcomes of interest are compared between the intervention group and a comparison group. Here the comparison group has not been well matched, or there is no strong evidence that the two groups have historically moved in parallel to allow a strong dif-in-diff design, and so there is a risk that it may not provide an accurate comparison.	If a comparison group is not well matched there is a risk that measured differences/a lack of measured difference between the two groups might not be due to the policy, but rather inherent differences between the groups or 'noise'.
	Predicted vs actual. Outcomes of interest are compared to expected or predicted outcomes (often constructed/modelled at the appraisal stage) of what would be expected if no action was taken (i.e. in the absence of the policy). Outcomes are only monitored for those experiencing the policy.	Such designs only "predict" a counterfactual, rather than directly measure it, so might provide an indication of whether there has been an effect, but may not be able to provide a precise statistical estimate of its size. A long time series before and after can help improve reliability.
	No comparison group. A relationship is identified between intervention and outcome measures in the intervention group alone. This frequently takes the form of a before and after design, in which outcomes of interest are compared to baseline measures taken before the implementation of the policy.	These designs provide a weak estimate of the counterfactual, particularly if there is only a single data point before and after the intervention: any number of factors could have influenced the measured change in the 'after' data. This typically results in the lowest level of confidence in attributing any measured change to the intervention, except in the rare cases where this is the only plausible explanation.

1 Introduction

Purpose

1.1 *Quality in policy impact evaluation* (QPIE) provides a guide for the design and assessment of impact evaluations of government policy.

1.2 This guidance focuses on empirical impact evaluations "which provide a quantitative measure of the extent to which any observed changes in an outcome of interest were caused by the policy".[1] The guidance illustrates how higher quality research designs can help meet the challenge of attributing measured outcomes to the policy in question (as opposed to other influences), whereas lower quality designs reduce confidence in whether it was the policy that achieved those outcomes.

1.3 QPIE will help both policy makers and analysts select the most appropriate and proportionate approach to evaluate the impacts of a policy by explaining various types of approaches and setting out their pros and cons. It aims to improve the quality of evidence by providing guidance on research designs that will produce more rigorous and reliable findings.

1.4 This paper builds on previous work on impact evaluation[2] and has been designed to apply more generally with applicability across government policies. It provides supplementary guidance to the cross-government guidance on evaluation, HM Treasury's *The Magenta Book: Guidance for evaluation* and should be used to inform the design of empirical policy impact evaluations, and to quality assure existing evaluation evidence.

Background

1.5 HM Treasury's Magenta Book provides in-depth guidance on how evaluation – process, impact and economic – should be designed and undertaken. It presents standards of good practice in conducting evaluations, and seeks to provide an understanding of the issues faced when undertaking evaluations of projects, policies, programmes and the delivery of services.

1.6 Chapter 5 of the Magenta Book describes the stages of evaluation, including the role of logic models, key research questions and types of evaluation and should be referred to for guidance on when to conduct any type of evaluation, including an impact evaluation. It sets out clearly the primary purpose of the main types of evaluation:[3]

- process evaluations: aim to "assess whether a policy is being implemented as intended";

[1] HMT Magenta Book, 2011:97
[2] For example, Sherman et al., 1997; Harper & Chitty, 2005
[3] HMT Magenta Book, 2011:17

- impact evaluations: aim to "provide an objective test of what changes have occurred, and the extent to which these changes can be attributed to a policy"; and
- economic evaluations: aim to "compare the benefits of a policy with its costs".

1.7 A good quantitative impact evaluation should provide evidence of not only if a difference is observed but also if this change can be attributed to the intervention in question. This finding often needs to be further unpacked using robust qualitative methods designed to explain why the intervention did/did not deliver the expected change – i.e. a process evaluation. This is especially important when evaluating complex policy interventions or where analysis of the quantitative data concludes that the intervention worked to different degrees with different groups.[4]

1.8 The evaluation method required will depend on the evaluation questions that need to be answered, although typically a mix of techniques associated with process, impact and economic evaluation will be required when developing evidence to inform the policy cycle[5] and to understand what happened and why. The guidance presented here focuses solely on the quality of the design of impact evaluations.

1.9 Impact evaluations are essential to answering the questions "did the policy work?" and "by how much?" As stated in the Magenta Book: "Answering the question of what difference a policy has made involves a focus on the outcomes of the policy. Outcomes are those measurable achievements which either are themselves the objectives of the policy – or at least contribute to them – and the benefits they generate. Questions under this heading might ask:

- "What were the policy outcomes, were there any observed changes, and if so by how much of a change was there from what was already in place, and how much could be said to have been caused by the policy as opposed to other factors?
- "Did the policy achieve its stated objectives?
- "How did any changes vary across different individuals, stakeholders, sections of society and so on, and how did they compare with what was anticipated?
- "Did any outcomes occur which were not originally intended, and if so, what and how significant were they?"[6]

1.10 Good quality impact evaluation evidence will be both theoretically driven[7] and provide confidence that the measured outcomes can be attributed to the policy and provide an estimate of the size of that impact. It can also provide evidence of any unintended impacts. Impact evaluation evidence can be used to make decisions about the continued implementation of a policy, and to inform the development of policy in the future.[8]

1.11 More detailed guidance on impact evaluation can be found in the Magenta Book. For example, Chapter 3 discusses the role of impact evaluation in policy design, including the importance of the counterfactual and adjustments that can be made to the public policy process to improve evaluation chances. Chapter 9 provides detailed conceptual guidance on impact evaluation, including establishing the counterfactual, risk, selection bias, validity, power of design, 'identification strategy' and reporting.

[4] For detailed guidance on qualitative evaluation see Spencer et al., (2003), *Quality in qualitative evaluation: a framework for assessing research evidence*.
[5] For more information on the policy cycle, see HM Treasury Green Book.
[6] HMT Magenta Book, 2011:18-19
[7] See HMT Magenta Book for the importance of logic models, which should be established ex ante.
[8] HMT Magenta Book, 2011:19

1.12 It is worth reproducing Box 9.B from the Magenta Book to illustrate that impact evaluations are not always practical or feasible, and the availability and quality of data will always be a key consideration: the best design will fail if the data quality is poor. Across many different impact evaluation designs, the quality of the measurement will be key.

1.13 This guide describes the QPIE and the considerations that should be made in striving for the highest quality impact evaluation design.

Box 1.A: Circumstances affecting whether empirical impact evaluation is feasible

	MORE FEASIBLE IF…	LESS FEASIBLE IF…
Scale of impact	Direct relationship between outcome of interest and driver whose effect it is desired to assess	Complex ("distant") relationship between outcome of interest and driver of interest, with many potential confounding factors
	Large effect relative to other changes taking place is expected	Small effect is expected
	Effect is realised within a short time period (and does not vanish immediately thereafter)	Effect builds up gradually over an extended time period
Data availability: what was done where,when, to whom outcomes	Policy involves a distinctive change in practice with respect to identifiable subjects (individuals, institutions or areas)	Policy involves a consolidation of existing best practice, or is poorly differentiated between subjects
	Data available on individual subjects	Only coarsely aggregated totals available
	Data available on precise time periods	Uncertainty over timing of implementation (requires aggregation over time)
	Data to support evaluation collected before and during policy	Data to support evaluation not sought until policy already established
Potential comparison groups	Pilot undertaken at the start including data collection in non-policy areas	No pilot, or data available only for the pilot areas themselves
	Phased start across areas	Simultaneous launch nationwide
	Objective allocation, for example using a cut-off score or random allocation	Subjective allocation
	Accidental factors influencing allocation	Optimal targeting: a "perfect" allocation can frustrate impact evaluation by leaving no equivalent comparison group

2 Quality in policy impact evaluation

2.1 QPIE (summarised on page four) provides a guide to the quality of impact evaluation designs. As stated in the Magenta Book, "...evaluating policy impact involves:

- determining whether something has happened (outcome); and
- determining whether the policy was responsible (attribution)."[1]

QPIE is primarily concerned with the issue of attribution. The higher quality the evaluation design, the more confidence there will be in concluding that the intervention *caused* the measured outcome/s and to what extent. Simply monitoring outcomes will provide information as to whether there has been a change, but it will not say whether the policy intervention caused some or all of that change.

2.2 Key to being able to demonstrate that a particular policy was responsible for an outcome is to identify what would have occurred if the policy had not been implemented and compare this to the measured outcomes after the intervention. This alternate reality is called the 'counterfactual'. A good counterfactual will be subject to the same influences as the policy intervention group, except for the effects of the policy. The counterfactual is typically measured by way of a comparison or, in the case of randomisation, a control group. Causality can rarely be confidently attributed to the policy if the impact evaluation design does not include a robust counterfactual in the shape of a comparison/control group, which has been matched on relevant variables or randomly allocated. If the policy evaluation design falls short of these levels of quality, it cannot be regarded as a robust impact evaluation and reduces confidence that the policy has directly contributed to the outcomes.

Application of QPIE

2.3 As discussed in the Magenta Book, it is essential to ensure that the impact evaluation design used is appropriate and proportionate to the research questions being asked. To identify the most appropriate design to test the impact of a policy, the risks, scale and profile of the policy should be considered (see Table 4.C of the Magenta Book for more details). Affordability of the desired impact evaluation will also be an important consideration, which should be considered carefully in light of requirements. However, it is also important to ask "can you afford not to conduct a high quality impact evaluation of the policy?" In most cases, without a robust impact evaluation it will not be possible to identify whether or not the policy caused the desired outcomes, with the risk that false or misleading conclusions are drawn. This could have wide-ranging cost and efficacy implications.

2.4 However, although causality can only be validly attributed to a policy through research designs which establish attribution, it is important to consider the impact evaluation requirements of each policy on a case by case basis. For example, there are some infrequent

[1] HMT Magenta Book, 2011:98

occasions when a before and after design (with no comparison group) may be sufficient. On these occasions, "the system being studied would be so simple that the policy is the only thing that could reasonably be expected to influence the result. Unfortunately, real social systems are seldom that simple. Unless there is a strong justification for ruling out influences other than the policy (not simply a lack of obvious alternative explanations), this design should not be reported as an impact evaluation".[2]

2.5 There may also be some situations when it is not possible to conduct more robust impact evaluation. For example, if European Union (EU) directives state that a policy must be rolled out at a national level, uniformly and all at the same time, it might not be possible to establish a valid comparison group. There may also be occasions where the EU defines the evaluation framework to be used for statutory evaluations, which may constrain the chosen evaluation designs. There are also challenges to robust evaluation from localism initiatives which may have small sample sizes which would constrain the ability to implement more robust designs. Nonetheless, the quality issues remain and these kinds of constraints do not alter what conclusions can be drawn about "what works" from particular impact evaluation designs.

2.6 The remainder of this paper provides further details on types of research design, setting out examples of the designs, a list of pros and cons, and an illustration of what this might mean in practice.

[2] HMT Magenta Book, 2011:122

3 Strong research designs in the measurement of attribution

Random allocation/experimental design

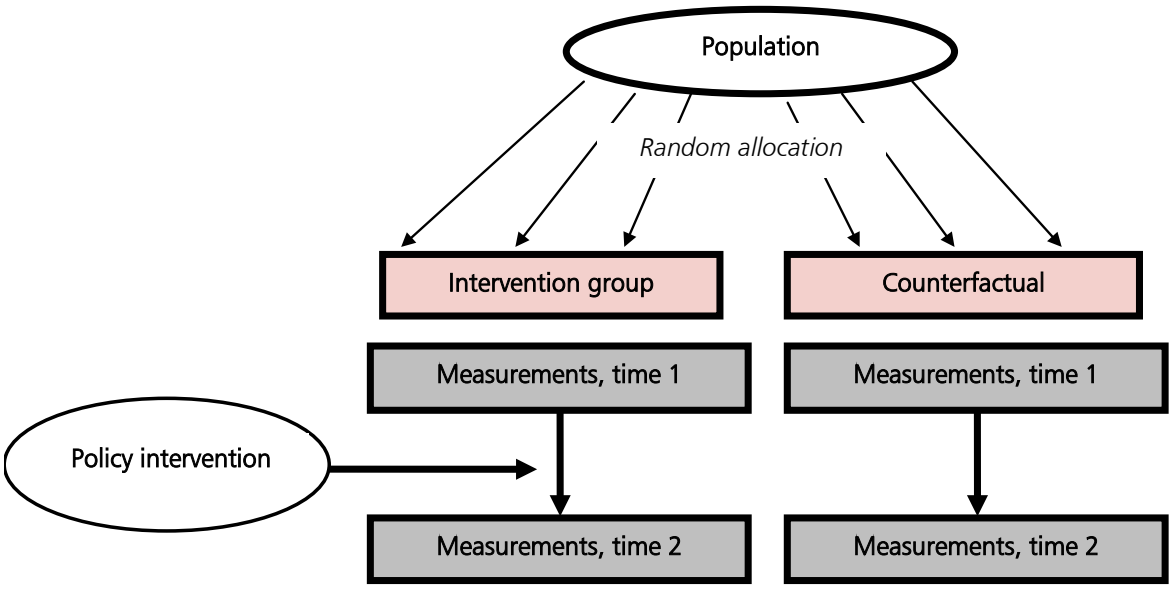
An experimental design, conducted properly, will establish whether an intervention caused an outcome. Such evaluation designs use random allocation to assign units of assessment (individuals/groups) to either the intervention or counterfactual group (often called 'control group' in experimental design). Given appropriate sample sizes and appropriate allocation to experimental or control groups, this is the strongest form of design for an impact evaluation, as the random allocation minimises the likelihood of any systematic differences – either known or unknown – between the groups. It therefore allows for an attribution of cause and effect.

There are many different ways to randomise in an experiment, including randomised field experiments and randomised control trials (RCTs) either at the individual or group level.[1] In medical trials there is typically a 'double blind' allocation to conditions (where both participants and those implementing the intervention are unaware of which group each participant is in). This is rarely possible with social interventions, but nonetheless the random allocation to intervention or control group does allow for a reasonable degree of confidence in attributing any measured outcome to the policy intervention.

Typically, random allocation takes place as part of a pilot to test a policy intervention prior to larger-scale roll-out, or else as part of a phased roll-out, where randomly assigned individuals or groups experience the policy intervention before the control individual or groups. The strength of these studies will depend on whether the two groups are representative of the population of interest, whether the sample size is sufficiently large and whether randomisation has been conducted appropriately.

[1] See http://www.cabinetoffice.gov.uk/sites/default/files/resources/TLA-1906126.pdf for discussion of RCTs.

Chart 3.A: Random allocation design



Examples of random allocation designs

- Randomisation at an individual level (e.g. individuals are randomly assigned to either the intervention or control group).
- Randomisation at a group level (e.g. groups of interest are randomly assigned, often referred to as a 'cluster randomised trial').
- 'Waiting list' designs, where individuals are randomly allocated to the waiting list or intervention, and short-term outcomes are compared.

Pros	Cons
The most robust, reliable findings which give confidence that any measured difference between the groups are the result of the intervention.	Can be difficult to conduct at a population level, especially for national programmes, which require pilot or trial to be valid for that population.
Random allocation should overcome any systematic difference between groups, even in unknown or unobservable variables. This should be checked to ensure comparability.	Field/quasi-experiments often have greater external validity: when randomisation occurs as a pilot or a trial, there is a risk that the findings are not relevant /scalable to a national/population level. This is a risk for all evaluation designs, but should not be neglected in favour of internal validity of random allocation.
Internationally recognised as a 'gold standard' method, so it is harder to argue with the findings.	Can be difficult to persuade others of the benefit of this design (i.e. withholding an intervention based on chance), and it can therefore involve substantial effort to gain agreement from policy leads, Minsters, stakeholders and the community being studied (although this is not a disadvantage to actual attribution).

Confidence in the effect size, and the relationship between the intervention and the outcome.	Can present some ethical issues (i.e. withholding an intervention based on chance, although if the efficacy of the intervention is not known, or if demand exceeds supply these are easily overcome). Other ethical issues can relate to deception (if people are not aware they are in a trial) and informed consent. The GSR Ethics Guidance[2] has more information on these issues. Can take longer to set up than quasi-experiments, can take more management to ensure it is conducted properly and as a result can be more expensive. For an RCT to be robust, good experimental conditions must be defined and maintained throughout. Without this, the quality of the findings are severely compromised.

Illustration of random allocation

The policy intervention is to distribute leaflets to household on energy efficiency with the aim of reducing domestic energy consumption (to help meet government Greenhouse Gas reduction targets).

For random allocation, the population of interest would be identified (e.g. a specific Local Authority area). Individuals in this population would then be randomly assigned either to receive a leaflet or not receive a leaflet – ensuring both groups are of sufficient size relevant to the expected effect to allow significant statistical differences in outcomes to be detected. A comparison between the before and after measures for each of these two groups would be made, ideally by someone who is not aware which group is which.

Quasi-experimental methods

Quasi-experimental methods attempt to mimic the conditions of randomisation so any measured difference can be attributed to the intervention. This is typically done through matching or through a comparison of two groups where the outcome/s of interest have historically moved in parallel.

Intervention group vs well matched counterfactual: Outcomes of interest are compared between the intervention group and a comparison group directly matched to the intervention group on factors known to be relevant to the intervention outcome. If this is done well, then in principle the matched comparison group can be treated as though it was created randomly – hence the description as 'quasi-experimental'. The statistical techniques to achieve this matching are covered in detail in the Magenta Book.

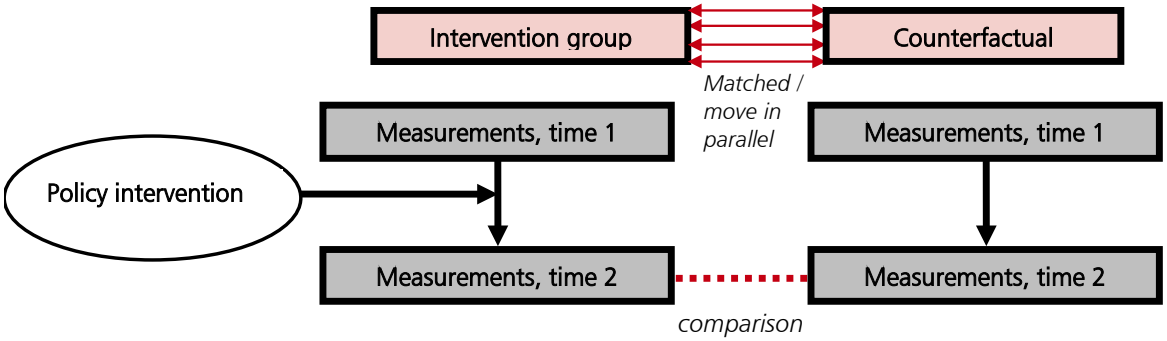
Strong difference-in-difference methodology. In this quasi-experimental design there is no direct matching; instead it involves a before and after study comparing two groups where there is strong evidence that outcomes for the two groups have historically moved in parallel

[2] http://www.civilservice.gov.uk/wp-content/uploads/2011/09/ethics_guidance_tcm6-5782.pdf

over time. This is similar to the matched-comparison group design, but there is no literal matching. Instead, the trends over time for the two groups are compared to provide an estimate of the overall impact of a policy.

The quality of these types of evaluations can vary, and is strongly tied to the quality of the data used with a substantial amount of data often being needed to do this well, and new primary data sometimes required. As with experimental designs, sufficient sample size is also important.

Chart 3.B: Quasi-experimental design



Examples of quasi-experimental designs

- Direct matching (e.g. Propensity Score Matching) (participants in the treatment group are matched with non-intervention participants to create a counterfactual that does not differ significantly from the intervention group on all known significant variables).
- Regression Discontinuity Design (participants are assigned to intervention or comparison groups based on a cut-off in a pre-intervention measure – e.g. those only just eligible are compared to those not quite eligible).
- Multiple regression (e.g. using longitudinal data).
- Difference-in-difference (the trends over time for the two groups are compared to provide an estimate of the overall impact of a policy).
- Instrumental Variables (groups are allocated on the basis of an external factor which influences the likelihood of policy exposure but which does not affect outcomes).

Pros	Cons
Provides a strong, 'quasi-experimental' design which can provide reasonably strong evidence of the relationship between the intervention and the measured outcomes.	Matching techniques tend to require a lot of data in both the intervention and comparison groups, which can often be difficult and/or expensive to acquire.
Can be used in situations when random allocation is not possible.	A good understanding is required of the factors that need to be matched. Without this, it remains possible that there are systematic differences between the two groups that are not being controlled for.

Ex-ante (pre-intervention) randomisation is not required, which avoids design and ethical issues typically associated with randomisation.	Even when matching on all theoretically relevant factors have been controlled for, it remains possible that there are other relevant but unmeasurable or unknown differences between the groups that will bias the measure of effect size. If new data are required, collecting data from the counterfactual group can be difficult, as the individuals, their gatekeepers or other organisational unit might have fewer incentives to collect or provide the required data, or to ensure the data are of a high quality. Such designs can require detailed, complex analytical work, and specialist knowledge will be required to conduct the analysis. Matching is only a successful means of reducing systematic differences between groups if all the factors that influence allocation to the control/treatment group are observed in the data.

Illustration of intervention vs matched comparison group

The policy intervention is to distribute leaflets to household on energy efficiency with the aim of reducing domestic energy consumption (to help meet government Greenhouse Gas reduction targets).

For this type of design, the outcome of interest (energy consumption) would be measured before the leaflets were dropped and again afterwards compared with the same measurement in a matched comparison group. This group would be matched at either the individual level, if sufficient data was available on factors associated with energy consumption (e.g. previous energy consumption, house type, area deprivation, population density, household size, household composition etc.), or else at a group level. Where matching is conducted at a group level, it would be necessary for there to be sufficient number of groups tested and compared to provide statistical rigour.

Limitations

- It can be difficult to get the level of data required to match at an individual level.
- Ideally matching would be done on all known predictors of energy consumption, but factors such as household income, attitudes and family make-up and circumstances can be difficult and expensive to collect.
- The effect size from this intervention is likely to be small, and there is likely to be a lot of 'noise' in the data: this could mask any real effects.

4 Weaker/riskier research designs in the measurement of attribution

Intervention group vs unmatched comparison group

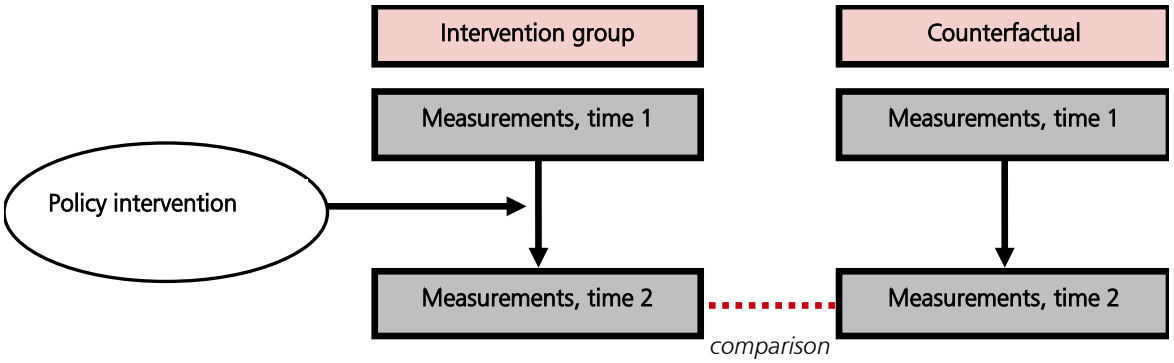
This design involves the collection of data from a comparison group who did not receive the policy intervention and where there is no attempt to match the two groups, so it is likely that the two groups are systematically different. The impact evaluation compares the outcomes from the two groups – those who were and were not subject to the policy intervention – to assess the impact of the policy against a no intervention or business as usual 'counterfactual'.

There are weaker and stronger designs in this type of evaluation. An example of a weaker design would be a straightforward comparison of outcomes for those volunteering for a particular opportunity to those offered but not taking up the same opportunity. In this situation, it is highly unlikely that the two groups (volunteers and non-volunteers) are at all comparable. The reason for any measured difference between the two groups might be a result of whatever made them volunteer or not, rather than the policy intervention being studied.

A slightly stronger design would be to use the general population as a counterfactual. Here you would measure what happened in the intervention group before and after the policy with what happened in the general population over the same period. There is less likelihood of a mis-match between the intervention and comparison group in this instance, but there will be a lot more 'noise' which could mask the true effect.

In some circumstances, this design will be the strongest design that can be adopted, but it will be important to take the limitations into account.

Chart 4.A: Unmatched comparison design



Examples of intervention vs unmatched comparison group

- Volunteer vs non-volunteer comparison (usually a weak design).
- Intervention group (however selected) vs population-level comparison (e.g. administrative data for all those in social housing, or a representative survey sample of GB adults etc.). Can be a stronger design but this approach is problematic if the intervention was targeted at specific groups, which are not representative of the population or the effect sizes are likely to be small.
- Simple comparison of one population who received a particular intervention with another population (e.g. comparing England to Scotland; Manchester to Birmingham). (Although this is stronger if done as a robust diff-in-diff design).
- Early adopters vs late adopters (weak)/fast starters vs slow starters (weak)/those suffering accidental delays vs rests of the population (potentially stronger).

Pros	Cons
There is a 'no intervention' counterfactual or comparison which allows an assessment of what would happen in the absence of the policy.	Unmatched intervention and comparison groups could be systematically different on important variables (for example, if the intervention group are volunteers with an interest in the policy area and the comparison group non-volunteers with no interest). In this case there is a risk that any measured difference is due to the difference between the two groups, and not the policy intervention. Comparing unmatched intervention and comparison groups can result in a lot of 'noise' and variability in the data, masking any real effects. If new data are required, collecting data from the counterfactual group can be difficult, as the individuals, their gatekeepers or other organisational unit might have fewer incentives to collect or provide the required data, or to ensure the data are of a high quality. The quality of the comparison depends heavily on how similar the intervention group is to the population they are being compared to.

Illustration of intervention vs unmatched comparison group

The policy intervention is to distribute leaflets to household on energy efficiency with the aim of reducing domestic energy consumption.

For this design, the outcome of interest (energy consumption) would be measured before the leaflets were dropped and again afterwards compared with the same measurement in a different, comparison group, for example a neighbouring Local Authority area or town without reference to the historical trends of the two groups as part of a robust difference-in-difference design.

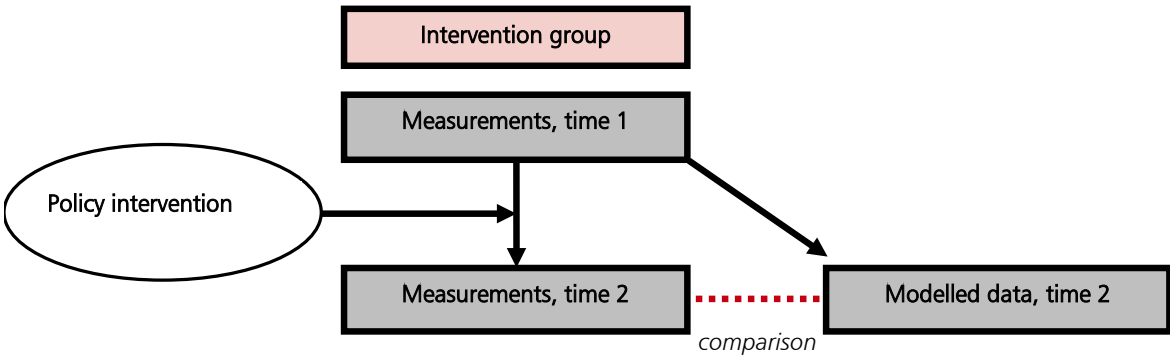
Limitations

- There is always a risk there is a systematic difference between the two groups (due to demographic, area or other effects).
- Especially when the effect size is likely to be small, there will be a lot of 'noise' in the data, and this could mask any real effects.

Predicted vs actual

These studies do not use separate comparison or control groups, but use time trends and modelling of the intervention group alone. The analysis compares real post intervention data with modelled/predicted data to assess the impact of the intervention. This has links with the appraisal/impact assessment process as, by necessity, this process uses modelled/predicted data about the future state of affairs to assess the likely impact of the policy. This approach is heavily dependent on the quality of the model that is being used to generate the prediction, and also has difficulty in accounting for the effects of other, unexpected, contemporaneous factors.

Chart 4.B: Predicted/modelled comparison design



Examples of predicted vs actual designs

- Comparing impact assessment predictions/modelled data with actual data at the individual or group level (e.g. re-running the analysis conducted to generate the

estimates used in the impact assessment using actual monitoring data for the intervention group and comparing the two outputs).

- Interrupted Time Series designs (a data series, with numerous data points, is collected both before and after a time-marked intervention). Any 'interruption' in the time series after the introduction of the intervention is then attributed to the intervention.

Pros	Cons
The pros are similar to those of pre- post-test monitoring: • At core it is basic monitoring. • It provides important information on what is being put in, and what comes out as a result of an intervention. • It can often use existing administrative or performance management data, which can be timely and cost-effective and involve good quality data.	The cons are also similar: • It is rarely possible to confidently attribute any measured change to the policy intervention. • It is rarely possible to predict exactly what would have happened in the absence of the policy. • The conclusions are open to challenge and interpretation.
Modelling is typically conducted as a standard part of appraisals, and so does not need to be produced specifically for an assessment of impact.	Measured differences could be the result of the quality of the model, rather than the result of the intervention.
The comparison can be undertaken without specialist training, although the modelling, and assessing the quality of the modelled data, does require specialist skills.	The usefulness of the model is heavily dependent on the quality of the data that has been used to inform the model.
With an interrupted time-series design, extensive time-series data before and after the intervention can provide useful evidence about the effect of the intervention in the underlying trend.	Effect size needs to be significant to overcome 'noise' in the data. Multiple data points are required before and after the intervention. Conclusions are still open to challenge/ misinterpretation.

Illustration of predicted vs actual design

The policy intervention is to distribute leaflets to household on energy efficiency with the aim of reducing domestic energy consumption (to help meet government Greenhouse Gas reduction targets).

For this design, the outcome of interest (energy consumption) would be measured before the leaflets were dropped, and again afterwards and this would be compared to the predicted level of change based on the previous time trends or other predicted trends presented in the policy appraisal.

Limitations

- Seasonality, weather conditions, recession, media campaigns all might have had an influence on any measured difference between before and after and could make the prediction invalid: there is no way of being sure that any difference was *caused* by the leaflets or of understanding what would have happened anyway.
- The predicted impact of the leaflets was based on previous weak evidence, so it cannot confidently be said whether it is a conservative or optimistic target.
- This does not capture what *actually* happened in the wider context.

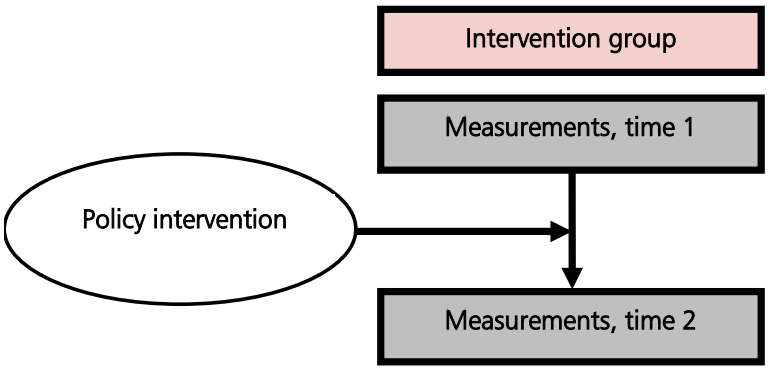
No comparison group

Studies without a comparison group frequently take the form of a before and after design. These designs simply take a measure of the situation before a policy intervention has been introduced (typically through a number of different measurements), and compare it to the situation afterwards. The initial measurements are called the baseline, against which subsequent measures (at various time points) will assess the change over time. Measurements typically cover what the policy is designed to change and the resources going in and coming out, if relevant.

No comparison group studies identify a relationship between the intervention and outcome measures, but it is not possible to say what would have happened anyway (the counterfactual).

In this type of study, measurement is only taken for the group experiencing the intervention.

Chart 4.C: No comparison design



Examples of no comparison group designs

- Comparing administrative data (e.g. levels of resource used) collected before and after the intervention.
- Comparing performance management data collected before and after the intervention.
- Collecting new data (e.g. about awareness, stated behaviour etc.), through surveys, census or other means, and comparing findings before and after the intervention.

Pros	Cons
This is simply monitoring, and is an essential part of any programme implementation.	Although monitoring is important, it is limited in what it can tell us about a policy intervention.
Monitoring provides important information on what is being put into the policy intervention (resources, materials, regulations) and whether the desired outputs are being achieved.	It is rarely possible to attribute any measured change to the policy intervention, or to understand what would have happened in the absence of the policy. Change could have been the result of factors other than the policy intervention, for example broader changes in public opinion, economic conditions, media campaigns, etc.

Monitoring can be conducted as part of ongoing performance management, and so is both timely (real time data can be obtained) and offers good value for money (often with little extra expenditure). Collecting data for operational/performance management purposes can often mean the quality of the data is high, as those delivering the intervention directly use the data and so have a vested interest in its quality.	The policy intervention will often be open to challenge that there was no additionality (i.e. that it did not achieve anything above what would have happened anyway); no value for money; or no robust evidence base.

Illustration of no comparison group

The policy intervention is to distribute leaflets to household on energy efficiency with the aim of reducing domestic energy consumption (to help meet government Greenhouse Gas reduction targets).

For this study, the outcome of interest (energy consumption) would be measured (via meter readings) before the leaflets were dropped, and again afterwards.

Limitations

- Seasonality, weather conditions, recession, media campaigns all might have had an influence on any measured difference between before and after: there is no way of being sure that any difference was *caused* by the leaflets or of understanding what would have happened anyway.

Acknowledgements

The authors would like to thank the peer reviews of an earlier draft of this paper: Susan Purdon, Anna Vignoles, Patricia Broadfoot, Simin Davoudi and Liz Dowler.

They would also like to thank the Cross-Government Evaluation Group for their comments and input. In particular, thanks go to:

Mike Daly

Jenny Dibden

Rachel Dubourg

Daniel Fujiwara

Hiroko Plant

Thanos Alifantis

The authors also gratefully acknowledge the work of DECC and Defra colleagues on this guidance: Kate Viner, Emily Hancock, Rachel McCloy, Rachel Muckle and the Defra Social Research Group.

B References

GSR Professional Guidance. *Ethical Assurance for Social Research in Government* http://resources.civilservice.gov.uk/wp-content/uploads/2011/09/ethics_guidance_tcm6-5782.pdf

Harper, G. & Chitty, C. (2005). The impact of corrections on re-offending: A review of 'What works'. *Home Office Research Study 291.* London: Home Office http://webarchive.nationalarchives.gov.uk/20110218135832/rds.homeoffice.gov.uk/rds/pdfs04/hors291.pdf

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012). *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials.* London: Cabinet Office http://www.cabinetoffice.gov.uk/sites/default/files/resources/TLA-1906126.pdf

HM Treasury *The Green Book: Appraisal and Evaluation in Central Government* http://www.hm-treasury.gov.uk/d/green_book_complete.pdf

HM Treasury *The Magenta Book: Guidance for evaluation* http://www.hm-treasury.gov.uk/data_magentabook_index.htm

Sherman, L.W., Gottfreson, D.C., MacKenzie, D.L., Eck, J., Reuter P. and Bushway, S.D (1998). Preventing Crime: What Works, What Doesn't, What's Promising. *Research in Brief*. National Institute of Justice. U.S. Department of Justice http://www.ncjrs.gov/pdffiles/171676.pdf

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. (2003). *Quality in qualitative evaluation: a framework for assessing research evidence* http://www.hm-treasury.gov.uk/quality_qualitative_evaluation.htm

HM Treasury contacts

This document can be found in full on our website: http://www.hm-treasury.gov.uk

If you require this information in another language, format or have general enquiries about HM Treasury and its work, contact:

Correspondence Team
HM Treasury
1 Horse Guards Road
London
SW1A 2HQ

Tel: 020 7270 5000

E-mail: public.enquiries@hm-treasury.gov.uk

ISBN 978-1-909096-49-3