About our Reviews

How to use these reviews

The Centre's reviews consider a specific type of evidence – **impact evaluation** – that seeks to understand the causal effect of policy interventions and to establish their cost-effectiveness. In the longer term, the Centre will produce a range of evidence reviews that will help local decision makers decide the broad policy areas on which to spend limited resources. Figure 1 illustrates how the reviews relate to the other work streams of the Centre.

Using the findings to inform programme design

Individual evidence reviews outline what tends to work in a given policy area – 'Best Bets' – based on the best available impact evaluations. For example, the Employment Training review suggests that:

- If an employment training programme aims to improve the employment prospects for an individual, it's probably a good idea to involve employers in the design of the programme and providing on-the-job training.
- The greater the length of a training programme, the greater the need for extra support to participants to counteract the 'lock-in' effects of being taken away from job searching.
- Extending shorter length training programmes to a wider group of recipients is likely to produce a better success rate than providing longer training to a smaller group.

It is important to note that the evidence from these impact evaluations is **a complement, not a substitute, for local, on-the-ground practitioner knowledge**.

These policy reviews outline what tends to work – based on the best available impact evidence – but will not address 'what works where' or 'what will work for a particular individual'. Programmes must be tailored and targeted and an accurate diagnosis of the specific challenges a policy seeks to address is the first step to understanding how the evidence applies in any given situation.

Using the source evaluations to design specific programmes

The evidence base provides a rich source of detail that many policy makers will find useful in thinking through specific issues.

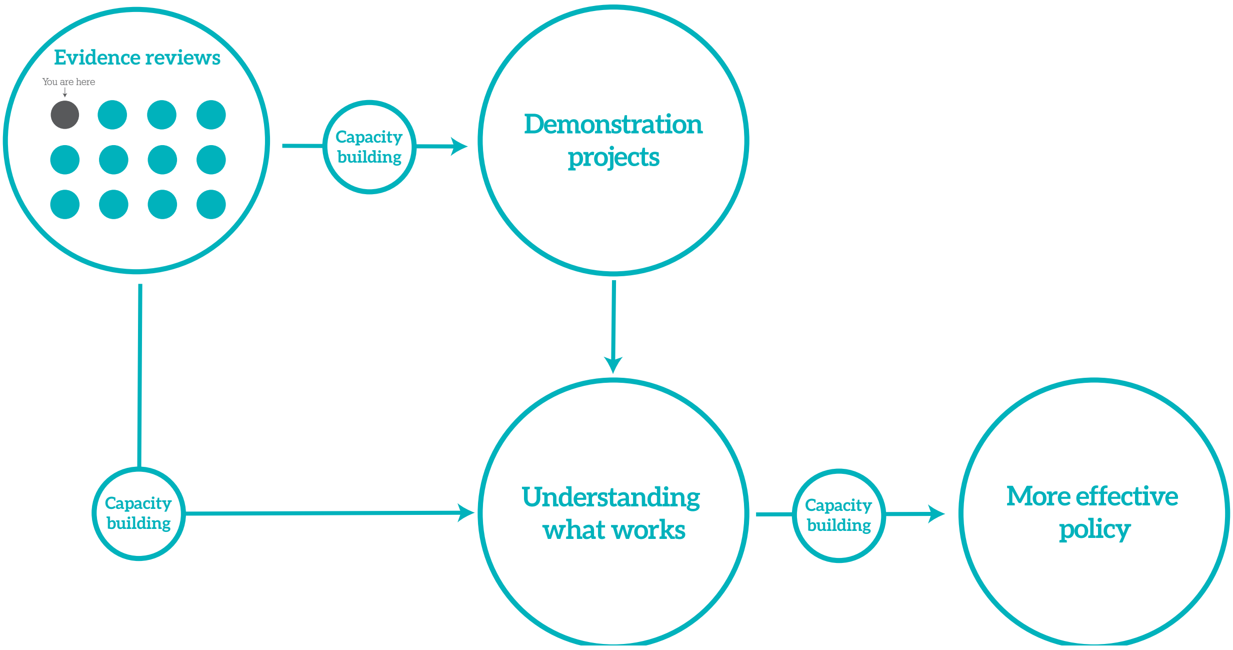
For example, referring again to the employment training review: if policy makers are facing the closure of a major employer, looking at some of the analyses of why retraining does not have a good track record can help to craft a more successful response.

Filling the Evidence Gaps

Reviews may not find answers to some of the questions which will be foremost in policy makers' minds. These gaps highlight the need for greater experimentation in policy making, with better evaluation to be embedded in policy design, and thinking differently about the policy cycle as a whole.

More guidance on how to improve impact evaluation will be forthcoming from the Centre.

Figure 1: What Works Centre work programme



Impact evaluation

Governments around the world increasingly have strong systems to monitor policy inputs (such as spending on a training programme) and outputs (such as the number of people who have gone through the programme). However, they are less good at identifying policy outcomes (such as the effect of a training programme on employment or wages). In particular, many government sponsored evaluations that look at outcomes do not use credible strategies to assess the **causal impact** of policy interventions.

By causal impact, the evaluation literature means an estimate of the difference that can be expected between the outcome for individuals 'treated' in a programme, and the average outcome they would have experienced without it. Pinning down causality is a crucially important part of impact evaluation. **Estimates of the benefits of a programme are of limited use to policy makers unless those benefits can be attributed, with a reasonable degree of certainty, to that programme.** The credibility with which evaluations establish causality is the criterion on which our reviews assess the literature.

Using Counterfactuals

Establishing causality requires the construction of a valid counterfactual – i.e. what would have happened to programme participants had they not been treated under the programme. That outcome is fundamentally unobservable, so researchers spend a great deal of time trying to rebuild it. The way in which this counterfactual is (re)constructed is the key element of impact evaluation design.

A standard approach is to create a counterfactual group of similar individuals not participating in the programme being evaluated. Changes in outcomes can then be compared between the 'treatment group' (those affected by the policy) and the 'control group' (similar individuals not exposed to the policy).

A key issue in creating the counterfactual group is dealing with the 'selection into treatment' problem. Selection into treatment occurs when individuals participating in the programme differ from those who do not participate in the programme.

An example of this problem in training programmes is skimming. Where providers choose participants with the greatest chance of getting a job the estimates of impact may be biased upwards.

Selection problems may also lead to downward bias. For example, people may participate to extend benefit entitlement and such people may be less likely to get a job independent of any training they receive. These factors are often unobservable to researchers.

So the challenge for good programme evaluation is to deal with these issues, and to demonstrate that the control group is plausible. If the construction of plausible counterfactuals is central to good policy evaluation, then the crucial question becomes: **how do we design counterfactuals?** Box 1 provides some examples.

Box 1: Impact evaluation techniques

One way to identify causal impacts of a programme is to randomly assign participants to treatment and control groups. For researchers, such Randomised Control Trials (RCTs) are often considered the 'gold standard' of evaluation. Properly implemented, randomisation ensures that treatment and control groups are comparable, thus identifying the causal impact of policy. However, implementation of these 'real world' experiments is challenging and can be problematic. RCTs may not always be feasible for local economic growth policies – for example, policy makers may be unwilling to randomise. And small-scale trials may have limited wider applicability.

Where randomised control trials are not an option, 'quasi-experimental' approaches of randomisation can help. These strategies can deal with selection on unobservables, by (say) exploiting institutional rules and processes that result in some people randomly receiving treatment.

Even using these strategies, though, the treatment and control groups may not be comparable. Statistical techniques such as Ordinary Least Squares (OLS) and matching can be used to address this problem.

Note, however, that higher quality impact evaluation uses identification strategies to construct a control group and then tries to control for remaining differences in observable characteristics. It is the combination that is particularly powerful: OLS or matching alone raise concerns about the extent to which unobservable characteristics determine both treatment and outcomes and thus bias the evaluation.

Evidence included in the review

We include any evaluation that compares outcomes for people receiving treatment (the treated group) after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual.

This means we look at evaluations that do a reasonable job of estimating the impact of treatment using either randomised control trials, quasi-random variation or statistical techniques (such as OLS and matching) that help make treatment and control groups comparable. We view these evaluations as providing credible impact evaluation in the sense that they identify effects which can be attributed, with a reasonable degree of certainty, to the implementation of the programme in question.

Evidence excluded from the review

We exclude evaluations that provide a simple before and after comparison only for those receiving the treatment (because we cannot be reasonably sure that changes for the treated group can be attributed to the effect of the programme).

We also exclude case studies or evaluations that focus on process (how the policy is implemented) rather than impact (what was the effect of the policy). Such studies have a role to play in helping formulate better policy but they are not the focus of our evidence reviews.

Methodology

To identify robust evaluation evidence on the causal impact of a policy, we conduct a systematic review of the evidence from the UK and across the world. Our reviews follow a five-stage process: scope, search, sift, score and synthesise.

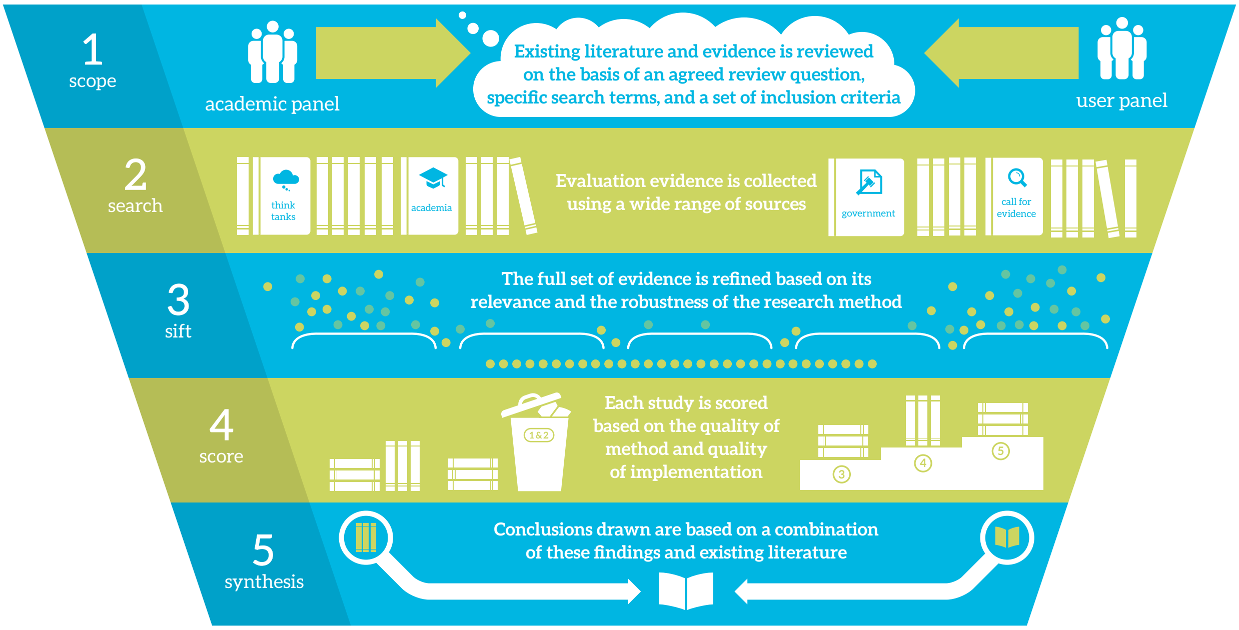
Scope of Review

Working with our User Panel and a member of our Academic Panel, we agree the review question, key terms and inclusion criteria. We also use existing literature reviews and meta-analyses to inform our thinking.

Searching for Evaluations

We search for evaluation evidence across a wide range of sources, from peer-reviewed academic research to government evaluations and think tank reports. Specifically, we look at academic databases (such as EconLit, Web of Science and Google Scholar), specialist research institutes (such as CEPR and IZA), UK central and local government departments, and work done by think tanks (such as the OECD, ILO, ippr and Policy Exchange.) We also issue a call for evidence via our mailing list and social media.

Figure 2: Methodology


1
scope
academic panel
Existing literature and evidence is reviewed on the basis of an agreed review question, specific search terms, and a set of inclusion criteria
user panel
2
search
think tanks
academia
Evaluation evidence is collected using a wide range of sources
government
call for evidence
3
sift
The full set of evidence is refined based on its relevance and the robustness of the research method
4
score
1&2
Each study is scored based on the quality of method and quality of implementation
3
4
5
5
synthesis
Conclusions drawn are based on a combination of these findings and existing literature

Sifting Evaluations

We screen our long-list on relevance, geography, language and methods, keeping impact evaluations from the UK and other OECD countries, with no time restrictions on when the evaluation was done. We focus on English-language studies, but would consider key evidence if it was in other languages. We then screen the remaining evaluations on the robustness of their research methods, keeping only the more robust impact evaluations. We use the Scientific Maryland Scale (SMS) to do this.[1] The SMS is a five-point scale ranging from 1, for evaluations based on simple cross sectional correlations, to 5 for randomised control trials (see Box 2). We shortlist all those impact evaluations that could potentially score three or above on the SMS.

[1] Sherman, Gottfredson, MacKenzie, Eck,Reuter, and Bushway (1998).

Scoring Evaluations

We conduct a full appraisal of each evaluation on the shortlist, collecting key results and using the SMS to give a final score for evaluations that reflect both the quality of methods chosen and quality of implementation (which can be lower than claimed by some authors). Scoring and shortlisting decisions are cross-checked.

Synthesising Evaluations

We draw together our findings, combining material from our evaluations and the existing literature.

Box 2: The Scientific Maryland Scale

Level 1: **Correlation of outcomes with presence or intensity of treatment, cross-sectional comparisons of treated groups with untreated groups, or other cross-sectional methods in which there is no attempt to establish a counterfactual.** No use of control variables in statistical analysis to adjust for differences between treated and untreated groups.

Level 2: **Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention ('before and after' study).** No comparison group used to provide a counterfactual, or a comparator group is used but this is not chosen to be similar to the treatment group, nor demonstrated to be similar (e.g. national averages used as comparison for policy intervention in a specific area). No, or inappropriate, control variables used in statistical analysis to adjust for differences between treated and untreated groups.

Level 3: **Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference).** Some justification given to choice of comparator group that is potentially similar to the treatment group. Evidence presented on comparability of treatment and control groups but these groups are poorly balanced on pre-treatment characteristics. Control variables may be used to adjust for difference between treated and untreated groups, but there are likely to be important uncontrolled differences remaining.

Level 4: **Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (i.e. difference in difference).** Careful and credible justification provided for choice of a comparator group that is closely matched to the treatment group. Treatment and control groups are balanced on pre-treatment characteristics and extensive evidence presented on this comparability, with only minor or irrelevant differences remaining. Control variables (e.g. OLS or matching) or other statistical techniques (e.g. IV) may be used to adjust for potential differences between treated and untreated groups. Problems of attrition from sample and implications discussed but not necessarily corrected.

Level 5: **Reserved for research designs that involve randomisation into treatment and control groups.** Randomised control trials provide the definitive example, although other 'natural experiment' research designs that exploit plausibly random variation in treatment may fall in this category. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance.

This work is published by the What Works Centre for Local Economic Growth, which is funded by a grant from the Economic and Social Research Council, the Department for Business, Innovation and Skills and the Department of Communities and Local Government. The support of the Funders is acknowledged. The views expressed are those of the Centre and do not represent the views of the Funders.

HM Government

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL