# Working paper 1

## Introduction to research methods and considerations of their relative strengths

| | |
|---|---|
| Project | Standards of evidence in housing with care, support and health |
| Identifier | Working paper 1 |
| Title | Introduction to research methods and considerations of their relative strengths |
| Author | Jim Vine, Director of Evidence, Data and Insight, HACT |
| This revision | 1.1 |
| Date | 25 March 2015 |
| Revision history | 1.0: first release, 30 January 2015<br>1.1: Minor amendment for clarity |

# Contents

# 1. Introduction

The National Housing Federation's recent *Prescription For Success* guide states that:

> *"From a health commissioner's perspective, a perceived weakness in [most housing evaluation research] is that it seldom includes robust analysis of health economic impact"* [1]

The point generalises beyond the health economic impacts that *Prescription For Success* was focused on. The report refers to small scale studies and descriptions of single site interventions relying on descriptive case studies and qualitative data; those types of studies would struggle to generate a robust causal analysis of any kind of impact.

When considering the options around what types of evidence are needed, one of the first things to consider is its intended purpose. Different types of evidence serve different purposes, and noting that a form of evidence has little use in one context does not imply that it has no value when addressing other issues.

This paper has been prepared as part of a project to consider what standards the housing sector should set for the creation of evidence in the context of care, support and health activities. Consequently, it takes as its focus the sort of robust evidence of impact that will be convincing to commissioners, and indeed convincing to housing associations that they are maximising the impact that they achieve through their programmes of work. The purpose of the evidence discussed is, therefore, to improve our understanding of 'what works'.

To keep the paper as general as possible, it will tend to refer to the things that we are interested in assessing the impact of as "interventions". The term is meant to be interpreted broadly and could include a package of physical measures applied to the home, a model of support offered to individuals, or a way of working within a housing provider's business.

[1] McDaid, D., Park, A., Eliot, J., Livsey, L. and Swan, A. (2014) Prescription for Success - a guide to the health economy. National Housing Federation, London. (p19) http://www.housing.org.uk/publications/browse/prescription-for-success-a-guide-to-the-health-economy/

# 2. Types of evidence and evaluation

## Quantitative, qualitative, and mixed methods

The question of 'what works' is essentially a quantitative one. How much does X change on average if we do Y? When a number of people receive Z, what proportion of them achieve some desirable outcome? By careful study of the outcomes that are achieved in different situations, when different interventions are applied, a conclusion can be reached about whether something works: intervention A has this impact on average; intervention B seems to do nothing to the outcome we are interested in; intervention C also has an impact on the outcome we are interested in, but on average its impact is lower than that of intervention A.

The question of 'what works' can also be enhanced by expanding its horizons to consideration of 'what works, for whom, in which contexts, and at what costs'. These are also questions where definitive answers will rely on quantitative analysis. Identifying those subsets of the population that will achieve most benefit from an intervention requires study of the impacts on each (using relevant statistical techniques). These questions are particularly salient in health-related research as tackling health inequalities is a particular concern for government and its agencies; consequently, it is important to understand whether interventions are effective when targeting the most vulnerable and whether they address the gap between richest and poorest. Assessing costs by placing values on the various inputs to the intervention has the potential to facilitate comparison of interventions based on their cost effectiveness. Intervention C might be half as effective as intervention A, but if it is also far cheaper, then the benefits achieved per pound might be higher; for the same fixed budget you might be able to deliver intervention C across a whole city, but only deliver intervention A to a single street.

Whilst attempts to use qualitative evidence to investigate whether a particular intervention achieves the desired outcomes will, at best, give only an indicative answer, it can play a far stronger role in contributing valuable insight into related evaluative questions, including insight into the processes at play in an intervention. The approach of embedding qualitative research methods alongside a robust quantitative evaluation – a mixed-methods approach – can permit the generation of information that neither type would deliver on its own.

## Evidence of 'what works' – methods that test causation

As noted above, identifying what works requires quantitative methods. But not all quantitative methods are equally suited to addressing this question. Beyond being quantitative, for a method to accurately identify whether something works or not, it must also test whether the intervention actually caused the outcome.

Two important concepts are useful to understand in avoiding mistakenly attributing a causal relationship between intervention and outcome where it is not proven:

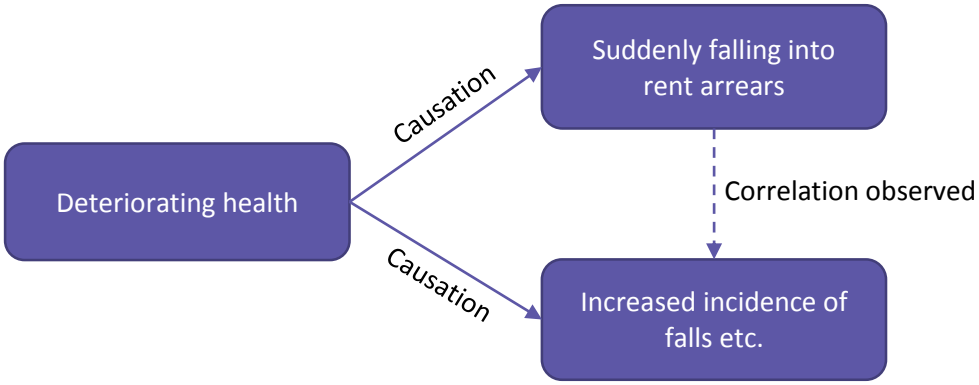
- Consider the counterfactual
- Correlation does not imply causation

Being able to compare against a **counterfactual** is an essential element of establishing causation. In broad terms, the counterfactual is the assessment of what would have happened anyway, in the absence of the intervention being tested. Imagine a group of 100 people that have a particular health problem and an intervention given to them all that is intended to ease the symptoms; after a period of 6 months the people are surveyed and only 50 of them now say they are suffering the symptoms of the condition. A simplistic assumption might be that the intervention has a 50% effectiveness rate, but without knowledge of the counterfactual this conclusion cannot be supported. Perhaps the condition tends to resolve of its own accord with time, and 50% of the people would be expected to be fine at 6 months anyway. Or perhaps the natural rate of recovery is only 25%, in which case the intervention is still effective, but not as effective as a naïve assessment would have found.

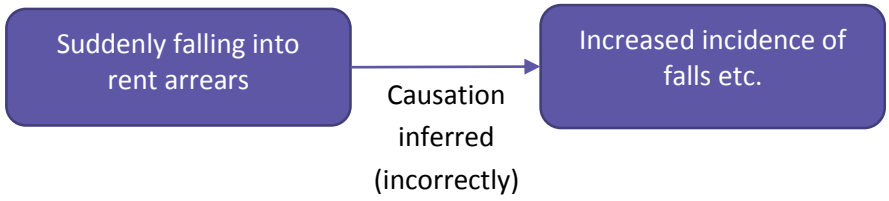
The comparison with the counterfactual can also be thought of as the difference between 'outcomes' and 'impacts'. When an intervention is made and the results are observed (but without allowance for the counterfactual) then a set of **outcomes** can be found for the programme participants. But the **impact** of the intervention cannot be discerned without the consideration of the counterfactual: the impact the intervention makes is the difference between what happens with it and what would have happened without it.

If a study identifies that two variables are **correlated**, that does not mean that one necessarily causes the other. One of the reasons why a correlation might be observed between the variables without causation is that both variable can be caused by some third factor (which might not be measured or observed in the study). Imagine that some data analysis showed that tenants over a certain age suddenly falling into rent arrears had a high probability of suffering trips and falls in the home in the subsequent months. That would not necessarily imply that the rent arrears were the cause of the falls; perhaps a more likely explanation is that deterioration in health (for example decreased mobility or the onset of dementia) might be the root-cause of both.

**Plausible causal links:**




**Correlation mistaken for causation:**




As previously noted, different types of evidence can serve different purposes and methods that identify non-causal correlations have their place. In the case of a link between arrears and falls it would be wrong to assume that dealing with the rent arrears problem would necessarily prevent incidents of falls later. But if, instead, the sudden onset of arrears in the at-risk population were treated as a warning flag to seek to attend to the actual cause, triggering assessments of properties, offers of support to the tenant, or the installation of grab rails, the negative outcomes might plausibly be lessened.

## The contribution of qualitative research to questions around what works

Although the question of whether a particular intervention achieves a particular outcome or not is a quantitative one, there are important related questions that robust qualitative methods are well-suited to investigating. These can include studying:

- "How" and "why" something works (or does not work);
- Implementation of an intervention in practice (fidelity);
- Whether the intervention addresses a problem that matters; and
- Whether the intervention is acceptable to service users.

As with quantitative methods, qualitative research methods are not all equally good at serving every purpose, and appropriate methods need to be adopted accordingly. Simply collecting service user

case studies, for example, may be useful in providing narrative content to help to support a more evidence-focused piece of research, but, in isolation, will not provide answers to the types of questions detailed above. Methods such as interviews, focus groups, observation and documentary analysis are among those most relevant for evaluations of interventions, if delivered to a high quality.[2]

Mixed-methods approaches, combining appropriate quantitative and qualitative methods, have the potential to generate answers to the full suite of relevant questions.

[2] Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. (2012) Quality in qualitative evaluation: a framework for assessing research evidence (supplementary Magenta Book guidance). HM Treasury, London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190986/Magenta_Book_quality_in_qualitative_evaluation__QQE_.pdf. p9

# 3. Overview of methods

This section introduces some methods of evaluation, with a particular focus on those that attempt to contribute to the 'what works' question. The terminology has generally been selected to echo that used in health spheres, although alternative classifications exist (for example, in economics the term "field experiment" is sometimes used, specifically distinguishing those trials undertaken in the field from laboratory-based experiments).[3]

## Randomised Control Trials (RCTs)

RCTs are often held up as the 'gold standard' of evidence of what works. When properly designed, they provide a robust assessment of causation requiring few assumptions or caveats and relatively simple statistical analysis. Whilst they are not suitable in all instances, where they are feasible, the 'gold standard' label is well-deserved.

The basic design of an RCT is simple: the population being tested is split into two (or more) groups by random assignment; one of these groups acts as a control group, receiving no intervention (or perhaps the 'business as usual' intervention), whilst the others receive one or more interventions that are being tested for effectiveness. Because the population has been split up randomly, there is no inherent difference between the groups, so any difference in outcomes detected can safely be declared as being a result of the intervention(s), assuming the numbers in each of the groups are sufficiently large to rule out chance fluctuations.

The simplicity of analysing the results of an RCT is most evident when compared to a research design where the two groups have been formed by some process other than randomisation, such as participants self-selecting to receive an intervention. In the non-random case, the researcher inherently faces a challenge of trying to understand whether any difference observed between the groups is actually due to the intervention or is because of some difference between the people who ended up in each group; perhaps the people who were motivated to apply to receive an intervention might also be those who were more likely to have achieved better outcomes anyway.

The randomisation in an RCT can be either at the individual level or at a 'cluster' level; for example whole estates can be randomly allocated to the intervention or the control group. In either event, a key concern is ensuring a sufficient sample size to be able to detect the size of effect that the intervention is hoped to achieve. In the case of cluster randomisation, the key consideration is that the number of clusters should be sufficiently large – not the number of individuals.

There are some circumstances when an RCT approach is not possible. Ethical concerns would prevent RCTs being undertaken when one of the groups would be exposed to probable harm. There may also be operational considerations that constrains the ability to randomise, such as the need to

[3] List, J. and Metcalfe, R. (forthcoming) Field Experiments in the Developed World: An Introduction (to be published as introduction to a special edition of the Oxford Review of Economic Policy on field experiments)

undertake a programme of major works in a systematic fashion to maximise efficiency, or impracticality in having different business processes in place for different customers.

*Further reading:*

For a short accessible introduction to running RCTs, see *Test, Learn, Adapt*.[4]

For a more detailed guide see *Designing Randomised Trials in Health, Education and the Social Sciences* or *Successful Randomized Trials*.[5,6]

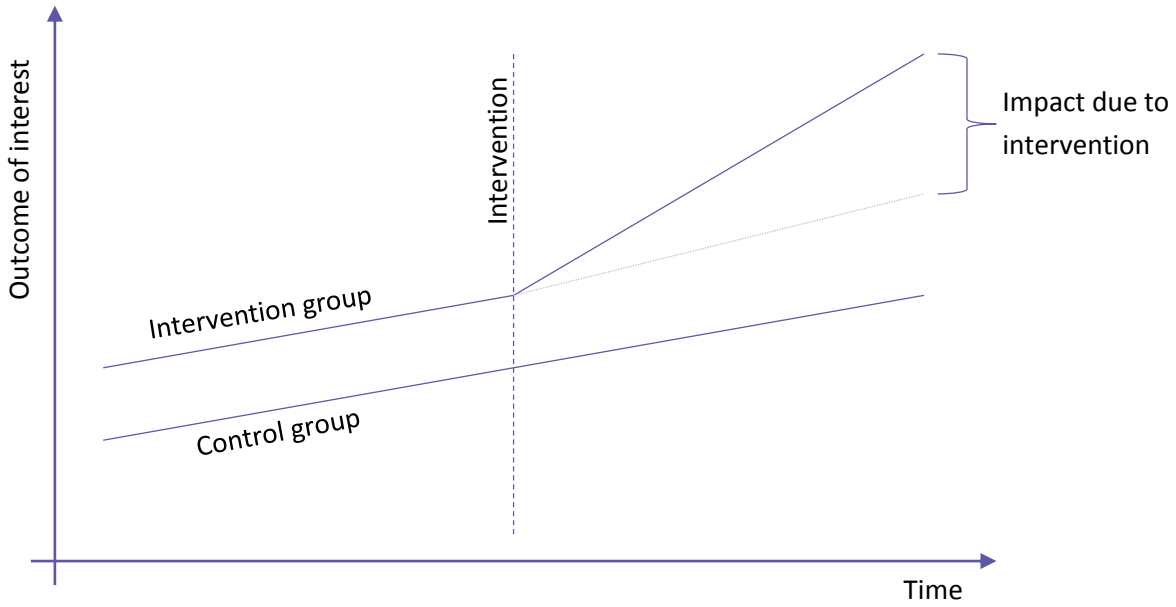
## Quasi-experimental designs

Quasi-experimental designs seek to create a causal link to assess the effectiveness of an intervention when randomisation is not possible. They include Instrumental Variables, Difference-in-Difference, Regression Discontinuity and Propensity Score Matching. These designs require the application of more sophisticated statistical techniques in their attempts to account for the effects of potential underlying differences between the intervention and control group. One such approach is to identify trends in some outcome for both the intervention group and the control group, and observe whether there is a change in the trend of intervention group, whilst the control group's stays on target.

[4] Haynes, L. Service, O. Goldacre, B. and Torgerson, D. (2012) Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team, London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf

[5] Torgerson, D. and Torgerson, C. (2008) Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction. Palgrave Macmillan, Basingstoke.

[6] Domanski, M. and McKinlay, S. (eds) (2009) Successful Randomized Trials: A Handbook for the 21st Century. Lippincott Williams & Wilkins, Philadelphia.




These approaches make attempts to control for potential sources of differences between the intervention and control group (known as confounding bias) but may not always be able to do so with certainty; if potential confounding factors remain unmeasured and unaccounted for, this can hinder the ability to make definitive statements about effectiveness. Because the techniques also rely on relatively sophisticated statistical handling they require particular expertise to be conducted rigorously, and the logic behind the findings may be less transparent to some audiences. Nonetheless, they present useful options as alternatives where randomisation is not possible.

## Observational studies

Observational studies differ from trials in that the assignment of the intervention and control groups is not controlled by the researcher. This presents the challenges associated with identifying whether the effects observed are caused by the intervention or by some underlying difference between the groups. Two common observational study designs are described below.

**Cohort studies** work by identifying a group of individuals and following them over a period of time. They were originally developed in epidemiology to examine the effects of exposures that are thought might be risk factors for disease.

**Case-control studies** are designed for identifying the causes of some outcome (typically a negative outcome such as a medical condition) by identifying a group of people with the condition and another group that do not have it (but that are, hopefully, otherwise similar). These two groups are then compared to identify how their characteristics differ.

Observational study designs might be useful in instances where randomisation would be unethical, such as testing whether a particular situation or exposure is harmful, but the same ethical

considerations do not apply to investigating exposures that have already happened. They will not, however, provide the same evidence of effect as a well-designed RCT.

## Combining findings from multiple sources

Approaches that synthesise the findings of other research, such as **systematic reviews**, have an important role in themselves. They can combine the results of several studies that only have tentative findings alone to produce a compelling understanding of whether an approach works. They can also compare a number of interventions that aim to achieve the same outcome and discern their relative effectiveness. This class of approaches includes **meta-analysis**, which focuses on employing statistical methods to compare and combine the findings of studies to achieve a greater degree of certainty than they deliver alone. Another approach, **realist synthesis**, aims to take a more context-sensitive approach, building up a theory of what works, for whom and in what circumstances, through the analysis of previous research.

Whichever approach is used, these methods rely on multiple studies having been conducted in related fields. This is one reason why it is useful to conduct replication research – testing something that has previously been tested elsewhere. Not only can it contribute to greater certainty overall, but it can also identify differences in the effectiveness across contexts; if a previous study has found that an intervention was not successful in one area for one population, it may still be worth re-testing in a different place and time to see whether it is more effective there. Conversely, an intervention that has previously been shown to be effective somewhere should not automatically be assumed to be effective always and everywhere.

*Further reading:*

An example of synthesis in housing and health see *Housing and health inequalities: a synthesis…*.[7]

[7] Gibson, M., Petticrew, M., Bambra, C., Sowden, A., Wright, K., Whitehead, M. (2011) Housing and health inequalities: a synthesis of systematic reviews of interventions aimed at different pathways linking housing and health. Health and Place, 17: 175-184.

# 4. Existing standards of evidence

## Hierarchy of (medical) evidence

One approach that is used to assess the standard of evidence is to create hierarchies based on study design types.

This hierarchy is typical of hierarchies of evidence found in the medical literature:

1. Systematic reviews and meta-analyses;

2. Randomised controlled trials with definitive results (confidence intervals that do not overlap the threshold clinically significant effect);

3. Randomised controlled trials with non-definitive results (a point estimate that suggests a clinically significant effect but with confidence intervals overlapping the threshold for this effect);

4. Cohort studies;

5. Case-control studies;

6. Cross sectional surveys; and

7. Case reports.[8]

Several alternatives exist, with some variations (for example, not including meta-analyses or grouping some of the design types together). Often, as here, the hierarchy will make some effort to handle RCTs with different levels of statistical power differently, but the broader assumption – that the studies at any given level have been well designed and well implemented – remains implicit.

## Maryland Scientific Method Scale

Academics at the University of Maryland developed a "scientific methods scale" as part of a study looking at crime prevention interventions.[9] The scale ranks evidence from 1 to 5, with 5 being the strongest scientific evidence.

[8] Hierarchy found in in Greenhalgh, T. (1997) How to read a paper. BMJ 1997;315:246. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2127173. Referenced to: Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. JAMA 1995;274:1800-4.

[9] Sherman, L.W. (1997) Chapter Two: Thinking About Crime Prevention. https://www.ncjrs.gov/works/chapter2.htm

1. Correlation between a crime prevention program and a measure of crime or crime risk factors

2. Temporal sequence between the program and the crime or risk outcome clearly observed, or a comparison group present without demonstrated comparability to the treatment group

3. A comparison between two or more units of analysis, one with and one without the program

4. Comparison between multiple units with and without the program, controlling for other factors, or a nonequivalent comparison group has only minor differences evident

5. Random assignment and analysis of comparable units to program and comparison groups

Levels 3 and above are those that include some form of control group in an attempt to demonstrate that the same trend would not have been observed without the intervention, with level 4 requiring more sophistication in the selection of the control and level 5 effectively an RCT.

## Nesta's Standards of Evidence

The standards of evidence developed by Nesta for impact investing contain some additional elements to the scales described above. Levels 2 and 3 reflect much of the hierarchy from demonstrating correlation but not causation through to an RCT approach (at the top of Level 3), that would be common ground with other scales of evidence of effectiveness. At the low end, Nesta's scale creates an entry point of being able to provide a coherent description of the intervention. At the top end, the specific mention of independent evaluation as necessary to reach a higher level (4) is not a common feature in other hierarchies, and the focus on replication research (level 5) takes on a different tenor, as it specifically addresses the question of whether the intervention can be scaled up.

| Level | Our expectation | How the evidence can be generated |
|---|---|---|
| At Level 1 | You can give an account of impact. By this we mean providing a logical reason, or set of reasons, for why your intervention could have an impact and why that would be an improvement on the current situation. | You should be able to do this yourself, and draw upon existing data and research from other sources. |
| At Level 2 | You are gathering data that shows some change amongst those receiving or using your intervention. | At this stage, data can begin to show effect but it will not evidence direct causality.<br>You could consider such methods as: pre and post-survey evaluation; cohort/panel study, regular interval surveying. |
| At Level 3 | You can demonstrate that your intervention is causing the impact, by showing less impact amongst those who don't receive the product/service. | We will consider robust methods using a control group (or another well justified method) that begin to isolate the impact of the product/service. Random selection of participants strengthens your evidence at this Level, you need to have a sufficiently large sample at hand (scale is important in this case). |
| At Level 4 | You are able to explain why and how your intervention is having the impact you have observed and evidenced so far. An independent evaluation validates the impact.<br>In addition, the intervention can deliver impact at a reasonable cost, suggesting that it could be replicated and purchased in multiple locations. | At this stage, we are looking for a robust independent evaluation that investigates and validates the nature of the impact. This might include endorsement via commercial standards, industry Kitemarks etc.<br>You will need documented standardisation of delivery and processes. You will need data on costs of production and acceptable price points for your (potential) customers. |
| At Level 5 | You can show that your intervention could be operated up by someone else, somewhere else and scaled up, whilst continuing to have positive and direct impact on the outcome, and whilst remaining a financially viable proposition. | We expect to see use of methods like multiple replication evaluations; future scenario analysis; fidelity evaluation. |

Source: [10]

## GRADE (Grading of Recommendations Assessment, Development and Evaluation) system

*What Counts as Good Evidence* describes the development of the GRADE system to respond to particular challenges to traditional hierarchies that are based on study design; the system has been adopted by NICE and other health bodies.[11] The criticism of traditional hierarchies centred on the view that a focus on study design is too narrow to assess the quality of evidence on a topic. Quality

[10] Puttick, R. and Ludlow, J. (2013) Standards of Evidence: An approach that balances the need for evidence with innovation. http://www.nesta.org.uk/sites/default/files/standards_of_evidence.pdf

[11] Nutley, S., Powell, A. and Davies, H. (2013) What Counts as Good Evidence? Provocation paper for the Alliance for Useful Evidence. Alliance for Useful Evidence, London. http://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf. pp11-12

of evidence is defined as "the amount of confidence that a clinician may have that an estimate of effect from research evidence is in fact correct" and is graded from high to very low, where high reflects a judgement that further research is not likely to change our confidence in the effect estimate.

To address this question, the GRADE system starts from an assessment based on study design (with an RCT design still initially rated more highly), but also factors in:

- Study limitations;
- Inconsistency of results;
- Indirectness of evidence;
- Imprecision; and
- Reporting bias.

There still remain arguments that even further aspects of evidence quality should be included in a broader assessment (in relation to medical evidence) such as:

- **Biological plausibility** – based on current biological knowledge of the mechanisms of disease, do the findings make sense?
- **Consistency in evidence across studies** – finding reproducibility in the effect of an intervention in numerous studies and across diverse populations and settings over time should add confidence.[12]

## Assessing the quality of qualitative evidence

The HM Treasury Magenta Book supplement on quality of qualitative evidence provides a framework for the assessment of the quality of qualitative research based around four central principles. It advises that the research should be:

- **Contributory** in advancing wider knowledge or understanding about policy, practice, theory or a particular substantive field;
- **Defensible in design** by providing a research strategy that can address the evaluative questions posed;
- **Rigorous in conduct** through the systematic and transparent collection, analysis and interpretation of qualitative data; and
- **Credible in claim** through offering well-founded and plausible arguments about the significance of the evidence generated.[13]

[12] Bagshaw and Bellomo (2008) referenced in *What Counts as Good Evidence?*

[13] Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. (2012) Quality in qualitative evaluation: a framework for assessing research evidence (supplementary Magenta Book guidance). HM Treasury, London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190986/Magenta_Book_quality_in_qualitative_evaluation__QQE_.pdf. p11

## MRC guidance on the evaluation of complex interventions

The Medical Research Council has published guidance specifically on the subject of complex interventions.[14] Complex interventions are defined as those with several interacting components (either in the package of intervention, in the range of possible outcomes, or in their variability in the target population), and are noted for the additional challenges they present to evaluators, over and above those that any evaluation will need to address. Many of the extra problems relate to:

- The difficulty of **standardising** the design and delivery of the interventions;
- Their sensitivity to features of the local **context**;
- The **organisational and logistical** difficulty of applying experimental methods to service or policy change; and
- The length and **complexity of the causal chains** linking intervention with outcome.

The guidance covers all stages of the process: developing an intervention, piloting and feasibility, evaluating the intervention, reporting, and implementation. It notes that all of these are important, stating "too strong a focus on the main evaluation, to the neglect of adequate development and piloting work, or proper consideration of the practical issues of implementation, will result in weaker interventions, that are harder to evaluate, less likely to be implemented and less likely to be worth implementing".

The guidance recommends that randomisation should always be considered for assessing effectiveness "because it is the most robust method of preventing the selection bias that occurs whenever those who receive the intervention differ systematically from those who do not, in ways likely to affect outcomes". It provides options for experimental designs that should be considered where randomisation is not possible at the individual level, including cluster randomisation (described above) and stepped wedge designs (where everyone eventually gets the intervention but the order in which they receive it is randomised).

## Matrix approaches to addressing various aspects of evidence need

A typical hierarchy seeks, usually implicitly, to capture the relative merits of different types of evidence *for some purpose*. Well-run RCTs might be the gold standard for determining whether particular interventions generate particular impacts, but they will not, on their own, necessarily reveal much about other questions like 'how does it work?' and 'is this an outcome that matters?'.

One response to the constraints inherent in a single ordered list is to develop matrices that allow an assessment of various study designs for a variety of purposes.

[14] Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008) Developing and evaluating complex interventions: new guidance. Medical Research Council. http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/

| Research question | Qualitative research | Survey | Case–control studies | Cohort studies | RCTs | Quasi-experimental studies | Non-experimental studies | Systematic reviews |
|---|---|---|---|---|---|---|---|---|
| Does doing this work better than doing that? | | | | + | ++ | + | | +++ |
| How does it work? | ++ | + | | | | | + | +++ |
| Does it matter? | ++ | ++ | | | | | | +++ |
| Will it do more good than harm? | + | | + | + | ++ | + | + | +++ |
| Will service users be willing to or want to take up the service offered? | ++ | + | | | + | + | + | +++ |
| Is it worth buying this service? | | | | | ++ | | | +++ |
| Is it the right service for these people? | ++ | ++ | | | | | | ++ |
| Are users, providers, and other stakeholders satisfied with the service? | ++ | ++ | + | + | | | | + |

Source: Adapted from Petticrew and Roberts 2003, Table 1, p.528.

Source: [15]

## Validity of evidence

An additional lens through which the quality of evidence can be considered is validity, in various forms. Internal validity – sometimes referred to as simply 'validity' is the assessment of whether the findings are an accurate, i.e. valid, representation of the real situation in the context where the study was made. External validity – or generalisability – is the question of the extent to which they are an accurate representation of the situation in wider contexts.

Generalisability might be limited where, for example, the population involved in the study comes predominantly or exclusively from a particular demographic or socio-economic group: an intervention that has substantial impact when delivered to a group of older social housing tenants might have a different impact profile for owner occupiers of a similar age and might not have any impact if offered to a younger group.

Study design can sometimes involve a trade-off between internal and external validity; internal validity might be increased by constraining the context in which the intervention is delivered, but at a cost of creating an artificial setting that is unlikely to be replicated if the intervention is delivered elsewhere (or even in the same location after the end of the trial). Economists sometimes frame a

[15] Nutley, S., Powell, A. and Davies, H. (2013) What Counts as Good Evidence? Provocation paper for the Alliance for Useful Evidence. Alliance for Useful Evidence, London. http://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf. pp16

distinction between laboratory experiments and field experiments, with field experiments ideally being delivered in a context that is as close to the ultimate delivery context as possible.

Beyond these primary types of validity, typologies have also been developed that refer to 'statistical conclusion validity' and 'construct validity'.[16]

[16] For further information see: Shadish, Cook and Campbell (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. p38

## 5. Evidence and innovation

If moving towards more evidence-based practice meant only ever using interventions that had been rigorously tested and evaluated, it would quickly put a halt to innovation. But done properly, evidence and innovation can be perfectly complementary: new ideas can and should continue to be implemented, but done so in a way that tests their effectiveness to establish whether the intervention should become established practice, whether it should be shelved, or whether it needs enhancements / amendments to make the maximum possible contribution.

One option to enable innovations to be made and tested in a way that is proportionate is to consciously adopt a tiered approach. A proof of concept pilot programme would be accompanied by relatively lightweight evaluation, but there would be an expectation that if the programme were rolled out further it would be re-evaluated more robustly.

In the USA, several government agencies have adopted a "tiered evidence" or "innovation fund" designs for grant making.[17] A three tier design might be categorised as:

- Proof of concept;
- Validation; and
- Scale up.

The lowest level (proof of concept) would be eligible for the least funding but would require the lowest level of evidence. Those applying for grants know that to be considered for funding, they must provide demonstrated evidence behind their approach and/or be ready to subject their models to evaluation. There is an explicit goal that interventions move up the tiers as their evidence becomes stronger.

[17] Burwell, S. M., Muñoz, C., Holdren, J. and Krueger, A. (2013) Next Steps in the Evidence and Innovation Agenda. Executive Office of the President, Washington DC. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-17.pdf. p8