What do we mean by standards of evidence?

Introduction

"Standards of evidence" refers to **how confident we can be that findings from service evaluations are accurate?**

This is a complex issue on which – ultimately – there is no consensus (academics will continue to debate the issue long into the future). But, to put it succinctly, higher standards of evidence are achieved through investing in more sophisticated research designs and adopting good practice from research methodology.

For Voluntary, Community and Social Enterprise (VCSE) organisations, the key question is not about how to achieve the highest standard of evidence possible, but to determine what is; a) realistic given time and resource constraints ; and b) appropriate for the stakeholders you want to engage.

- If your evaluation aim is to improve your services then your stakeholders are internal, so the standard of evidence needed is whatever is sufficient to persuade yourselves that findings are accurate. However, if you take this position it's important to challenge yourself that the information you collect is reliable. It's nearly always likely that better research designs will give you more insights.
- If your evaluation aim is to persuade funders or potential funders about the quality/impact of your services, then the standard of evidence needed is likely to be higher (though it will still depend on the views of individual funders).
- If your evaluation aim is to assess value for money or contribute to the wider evidence-base, then the standard of evidence needed is likely to be the highest possible.

Hence, our first recommendation **is to think through *who* you want to engage in your evaluation and what standard of evidence *they* think is appropriate.**

This guidance aims to help you further with this by briefly describing what people mean when they say "higher standard of evidence" and to signpost you to more information. The information presented may also help you to assess the standards of other research you come across during your work.

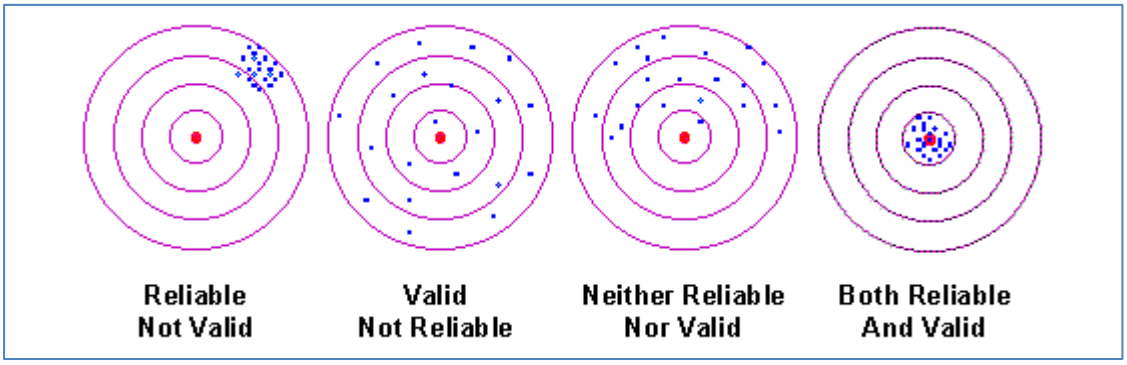
Validity and reliability

If you read text books on research methodology you will quickly encounter the concepts of reliability and validity, and it's useful to appreciate what they mean.

Validity refers to whether your research method actually measures or examines the issue you want it to. For example, 'the amount of crime' is better measured through victim surveys than police statistics because the former includes non-recorded crime.

Reliability refers to how well your chosen method measures or examines the issue it is intended to measure. For example, a better quality victim survey will interview a representative sample of people from all parts of the community.

The useful metaphor of archery is shown below[1]:



Following this example:

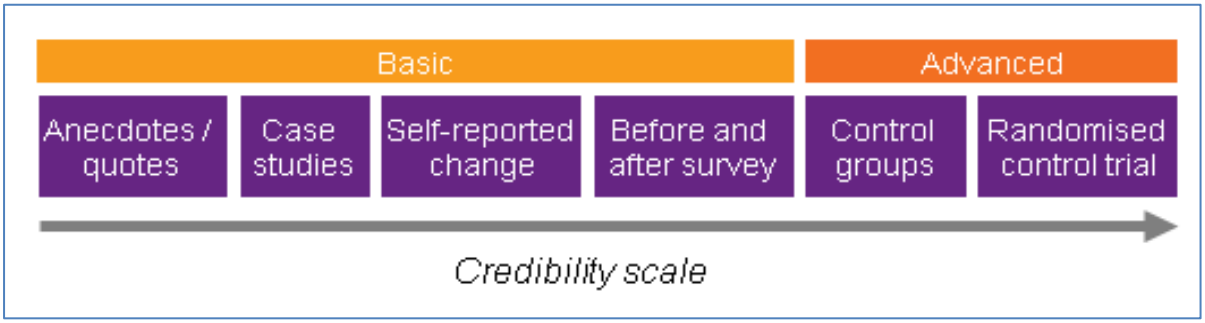
- **Police statistics are reliable but not valid.** They deliver consistent findings but do not include all crime.
- **A badly designed victim survey is valid but not reliable.** For example a victim survey on a website has the potential to record the level of crime but the results will depend on who can be bothered to fill it in
- **Gauging the level of crime from reports in a newspaper is neither valid nor reliable.** Newspapers tend to exaggerate crime and are unreliable because the level of coverage depends on what else is going on.
- **A well designed victim survey is both valid and reliable.** For example the Crime Survey for England and Wales

What validity means in practice

When social scientists and professional evaluators talk about standards of evidence they are generally talking about validity (i.e. whether the research design measures the thing it is intended to). Accordingly some researchers have created hierarchies of evidence from the least to the most valid approaches. In criminal justice evaluation the most notable of these is The Maryland Scale of Scientific Methods developed by Professor Larry Sherman and colleagues at the University of Maryland. This is widely referred to across Government and elsewhere (see Nutley et al: What counts as good evidence? (2012) for a fuller description of some other hierarchies).

[1] Taken from http://www.socialresearchmethods.net/kb/relandval.php

We have paraphrased these hierarchies in the chart below, dividing the scale into "basic" and "advanced".



The three most important things to note are;

1) Many stakeholders including policy makers and commissioners will regard qualitative research as less valid than quantitative. This is for a number of reasons including the fact that fewer service users are consulted through the research, less attention is given to how people are selected and questions are not asked or analysed in a consistent way. This is one of the most enduring debates in social research. Keen proponents of qualitative research claim that it is in fact *more* valid because it provides a better description and understanding of the social world. On a day-to-day basis we can usually ignore this debate. Most people agree that both have their uses; qualitative research gives insights into why and how things happen, while quantitative research provides more confident estimates of the extent of change, causes and effects and differences across contexts and subgroups. Hence all good evaluation will combine both approaches. Nonetheless it's useful to remember that many stakeholders will regard quantitative as inherently superior - especially when it comes to reporting impact.

2) "Experimental methods" are seen as the most valid approach. These are methods which mimic models of research in natural sciences. Results from 'treatment groups' are *compared* to those from 'control groups' - with the difference representing the 'effect' or 'impact' of a service.

3) Just because an approach or study is lower down the hierarchy doesn't mean it is useless. Indeed a wealth of useful information comes from qualitative research, observations, talking to stakeholders etc. (but it doesn't hurt to make this as high quality as possible – see following section).

A more detailed description of the different types of evidence on the scale above is below.

- **Anecdotes / quotes:** Observations from staff and users may be compelling but are not regarded as a high standard of evidence because they are not selected or analysed in systematic way.

- **Case studies:** A record of research into the development of a particular person, group, or situation over a period of time. These can provide strong evidence and are useful in illustrating and describing a service and its impact. But they are also susceptible to bias, most commonly cherry-picking when the "best" cases are implied as being typical.

- **Self-reported change:** Beneficiaries or service users read a question and select a response by themselves, usually giving a view on whether the service has made a difference to them or not. These responses are regarded as subjective and less reliable, and it is often difficult for beneficiaries to judge whether any change is attributable to a particular intervention or programme. Data from a practitioner or a family member can sometimes be used as an additional point of view to improve the approach.

- **Before and after measures:** These may be surveys filled in by a participant before (or at the beginning of) a programme and then again at the end (or shortly after it). They can measure change in areas, such as health, employment, housing status, and attitudinal change. However, this change does not necessarily demonstrate the impact of the programme because there may be other external factors that have caused it. There is also a tendency for respondents to over-report change and overstate the significance of a service because they think it's what you want to hear.

- **Control groups:** In this approach the level of change in a group of service users (treatment group) is compared to change in a similar group who do not receive the service (treatment group). It's seen as more valid because the results from the control group indicate what may have happened without the service (as shown in the chart). However, it is important to ensure that treatment and control groups are the same. For example it's invalid to compare volunteers for a service with non-volunteers because the fact that they volunteered makes them different. Similarly, treatment and control groups need to be matched by all the characteristics you think might be associated with service outcomes (e.g. age, gender, and history of offending). It's also important to know what – if any – alternative services the control group are receiving,

- **Randomised Control Trials (RCTs):** These are seen as the most rigorous way to select a control group. It involves randomly choosing those who receive the service and those who do not. Any differences in the outcomes of the control group can be attributed to the service or to chance (which can be statistically calculated). This approach helps to isolate a programme's impact from some of the other external factors that are affecting the same outcome[2]. This can sometimes present ethical issues for organisations in the VCSE Sector as it can mean denying a service to a group of potential beneficiaries. If you are unable to randomly assign users to a service an alternative is what is known as a quasi-random approach, in which a control group is drawn from people you know have had no access to the service and can be matched to your treatment group. (A good example of this is the Justice Data Lab, which uses a statistical technique to draw a control group from the large Police National Computer dataset).

For most VCSE organisations conducting evaluations at the advanced end of the scale is not feasible because of the resources and time required. Nonetheless, it is useful to be aware of

[2] A good example can be seen here http://www.theppt.org.uk/documents/Bilderbeck_Farias_2013_J_Psych_Res.pdf

the hierarchy when planning your evaluation activities and when reviewing research conducted by others.

What reliability means in practice?

Reliability issues apply no matter how valid the research design. Another way to think about reliability is that it's about quality. There is good and bad practice whatever methods you are using, and a basic methodology done well may provide better evidence than an advanced method that is poorly designed and executed.

It's impossible in a short note like this to provide a summary of all reliability issues - because each method brings its own considerations. However, the list below provides some general advice, which may be useful - either when reviewing someone else's evaluation or planning your own.

- All **evaluation should be underpinned by a theory of change**[3] or some other description of the intended outcomes (and how activities are intended to achieve them).
- The **latest research and relevant academic literature on the subject should have been reviewed** and reflected in both service design and evaluation.
- Systematic efforts should be made **to engage a wide range of service users** in the evaluation (including – most importantly – those who do not normally volunteer themselves to be consulted).
- **Sample sizes should be sufficient** and – if a before/after approach is used – the level of drop-out[4] should not be too high. In addition, there should be a discussion of how representative the sample is (i.e. by comparing age and gender profiles against the service user population as a whole).
- **Good evaluations always combine quantitative and qualitative evidence.** Quantitative data is used to determine the extent of change while qualitative data helps you understand the underlying mechanisms and provides real life illustrations.
- Aside from service users, **the views of other stakeholders should be sought** - including partner agencies, staff and volunteers.
- **The description of the research methods should be complete and transparent.** Including information about study design, methods, procedures, sample sizes, response rates etc. This is often missing from evaluations and undermines perceptions of reliability.
- **Questionnaires and other 'instruments' used are available to review** and reflect good practice themselves.

[3] http://www.clinks.org/sites/default/files/TheoryofChangeGuide.pdf

[4] Also known as 'attrition'

- **Results are presented in a clear and impartial way**, with consideration of alternative explanations for the findings and an assessment of any possible sources of bias.

Choosing the right level of evidence for you

Given that stakeholders can have different opinions, it can be tricky for VCSE organisations to decide on what level of evidence you need. Unfortunately there is no clear answer. Your choices will depend on a combination of what is desired or needed and what is practical. But to help, you could ask yourself the following questions:

- **Who do you want to engage in your evaluation findings, and how will they view the evidence?**
    - If you hope to use the evidence to prove your impact externally you will want to be able to defend it against criticism. If you are trying to persuade government departments to make a big investment or change in policy, you will need strong evidence and you may want to get external advice.
    - If your audience is an internal audience, you may not need as strong evidence to convince them - but you will want good evidence so you know how to improve. It may be particularly important to get good qualitative data on when your service does and does not work.
- **What level of evidence is *accepted* within your sector?**
    - This is a useful check on the minimum level of evidence required. You should aim to achieve the minimum standards of your peers, but if you want to make your organisation stand out you will need to go further.
- **Is there a risk that your service might actually be doing harm, and if so, how much harm?**
    - There are examples from medical research of practices that were theoretically sound and appeared to benefit patients but were actually found to cause harm when tested through more robust methodologies[5].
    - You should take this possibility seriously. The robustness of your impact data should therefore be proportional to the likelihood that you could cause harm and the level of that harm. The higher the likelihood and the size of these impacts the more important it is to pursue a higher standard of evidence.
- **Are you demonstrating an impact that has not yet been proven?**
    - Before jumping into a high standard evaluation, check existing evidence and academic literature. If this is already strong, you can draw upon it to support your service and you only need to collect sufficient data to show your service should replicate the results.
    - If your service is untested you will need stronger evidence to persuade yourselves and others that it works. However, at the earliest stage of a new or innovative service it is better to do smaller-scale evaluations to pilot or test

[5] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/193913/Building_evidence_into_education.pdf

concepts and approach. Only when a service has been developed/refined to the level where it is being implemented consistently should it be subject to an evaluation with a higher standard of evidence. A useful thing to remember is that if a service is going to fail, it is better to fail on a small scale and fail quickly. You generally don't need a high standard of evidence to find this out, trial and error and continuous learning from the beginning will achieve this.

- **What resources (time and budget) do you have available?**
    - Higher standards of evidence will (typically, but not always) take more time and money spent in preparation, data collection and analysis.
    - Hiring an external consultant will help you design an appropriate monitoring and evaluation programme and maintain the rigour of your analysis. But this comes at a price.
    - The highest standards of evidence might be becoming easier to achieve for VCSE organisations through initiatives such as Nesta's Randomise Me , the Education Endowment Fund's DIY Evaluation Guide, and the Justice Data Lab.
    - However, if you are taking your own route and teaching yourself or your staff how to do a rigorous evaluation, it is worth investing in training[6].
- **Continue to review whether your standards of evidence are right for you**
    - Ultimately the standards you adopt will depend on the weighting you give to each of the above questions. But this is not fixed and will differ between organisations and sectors and at different times within your own organisation. For instance, a change amongst your trustees, an expansion of your programme to a new locality or a new cohort should prompt you to review the standard of evidence you feel is appropriate.

[6] As part of the Improving your evidence project we will be providing free training in early 2014

Appendix: Further reading and other useful evidence hierarchies

A comprehensive and accessible introduction to thinking on standards of evidence has been prepared by Nutley et al (November 2012), for the Alliance for Useful Evidence. We recommend this as a good place to start if you are more interested in the subject.

"Transforming Rehabilitation: A summary of evidence on reducing reoffending" (MOJ:2013). **Annex D** articulates the official view of different standards of evidence.

Project ORACLE is a London based project looking to improve the use of evidence by organisations working with young people. They have defined five levels of evidence which relate to a programme's evaluation plans. Level 1 (entry level) requires a sound theory of change or logic model with clear plans for evaluation and level 5 is the highest, requiring a 'system-ready' intervention that has been subject to multiple independent evaluations.

MoJs Correctional Services Accreditation Panel has a specific system for accrediting offender behaviour programmes. It is more of a checklist in the design and delivery of services and is summarised on pages 26-35 the report.

Nesta's standards of evidence look to provide a more holistic hierarchy of evidence, combining both methodological confidence and other aspects of the maturity of the service.

Social Research Unit standards of evidence. These take a slightly different approach by having four dimensions; combining methodological rigour, with achieved/projected impact, "intervention specificity" and "public service readiness". It has also been used for the Big Lottery Fund's Realising Ambition programme and adapted by NESTA for use in social investment.