WHAT COUNTS AS GOOD EVIDENCE?

PROVOCATION PAPER FOR THE ALLIANCE FOR USEFUL EVIDENCE

Sandra Nutley, Alison Powell and Huw Davies

Research Unit for Research Utilisation (RURU)
School of Management
University of St Andrews

February 2013

This is a paper for discussion.

The authors would welcome comments, which should be emailed to: **smn@st-andrews.ac.uk** or **Jonathan.Breckon@nesta.org.uk**

The paper presents the views of the authors and these do not necessarily reflect the views of the Alliance for Useful Evidence or its constituent partners.

The Alliance champions the use of evidence in social policy and practice. We are an open-access network of individuals from across government, universities, charities, business and local authorities in the UK and internationally. The Alliance provides a focal point for advancing the evidence agenda, developing a collective voice, whilst aiding collaboration and knowledge sharing, through debate and discussion. We are funded by the BIG Lottery Fund, the Economic and Social Research Council and Nesta. Membership is free. To sign up please visit: **www.alliance4usefulevidence.org**

Nesta

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

BIG LOTTERY FUND

LOTTERY FUNDED

CONTENTS

WHAT COUNTS AS GOOD EVIDENCE?

In brief

Making better use of evidence is essential if public services are to deliver more for less. Central to this challenge is the need for a clearer understanding about standards of evidence that can be applied to the research informing social policy. This paper reviews the extent to which it is possible to reach a workable consensus on ways of identifying and labelling evidence. It does this by exploring the efforts made to date and the debates that have ensued. Throughout, the focus is on evidence that is underpinned by research, rather than other sources of evidence such as expert opinion or stakeholder views.

After **setting the scene**, the review and arguments are presented in five main sections:

We begin by exploring **practice recommendations**: many bodies provide practice recommendations, but concerns remain as to what kinds of research evidence can or should underpin such labelling schemas.

This leads us to examine **hierarchies of evidence**: study design has long been used as a key marker for evidence quality, but such 'hierarchies of evidence' raise many issues and have remained contested. Extending the hierarchies so that they also consider the quality of study conduct or the use of underpinning theory have enhanced their usefulness but have also exposed new fault lines of debate.

More broadly, in **beyond hierarchies**, we recognise that hierarchies of evidence have seen most use in addressing the evidence for what works. As a consequence, several agencies and authors have developed more complex matrix approaches for identifying evidence quality in ways that are more closely linked to the wider range of policy or practice questions being addressed.

Strong evidence, or just good enough? A further pragmatic twist is seen by the recognition that evaluative evidence is always under development. Thus it may be more helpful to think of an 'evidence journey' from promising early findings to substantive bodies of knowledge.

Finally, we turn to the **uses and impacts of standards of evidence and endorsing practices.** In this section we raise many questions as to the use, uptake and impacts of evidence labelling schemes, but are able to provide few definitive answers as the research here is very patchy.

We **conclude** that there is no simple answer to the question of what counts as good evidence. It depends on what we want to know, for what purposes, and in what contexts we envisage that evidence being used. Thus while there is a need to debate standards of evidence we should be realistic about the extent to which such standard–setting will shape complex, politicised, decision making by policymakers, service managers and local practitioners.

1 SETTING THE SCENE

Background

Calls for the better use of rigorous evidence in developing and delivering public services in the UK are not new. They have, however, become more urgent in tone due to severe public expenditure cuts and the need to ensure that scarce funds are allocated in ever more cost-effective ways. The Civil Service Reform Plan (HM Government, 2012) suggests that there may be a need for an improved infrastructure to trial and assess what works in major social policy areas. The aim is to ensure that commissioners in central or local government have the evidence to support effective commissioning.

There may also be a need to improve commissioning processes. A recent study of social care commissioning guides (Huxley et al., 2010) found that they did not, in fact, pay much attention to research evidence (even when it was available) and relied instead on government documents or practice guidance. While there are no simple infrastructure changes that are likely to transform commissioning processes, clarity about evidence will be key.

Healthcare is often viewed as being ahead of other policy domains in setting standards of evidence on which to base clinical decisions. It is therefore not surprising that there has been a lot of interest in its evidence infrastructure. In January 2012, Sir Jeremy Heywood, Cabinet Secretary, held up the example of the National Institute for Health and Clinical Excellence (NICE) as a possible direction of travel for the social policy field. A social policy 'NICE', he said, could provide independent 'kitemarks' to vouch for the effectiveness of social policy schemes. We return to this proposition at the end of the paper, but first there are many prior issues to consider, starting with the nature of social policy evidence.

Debating evidence

The Alliance for Useful Evidence's November 2012 seminar on *What is good evidence? Standards, Kitemarks and Forms of Evidence* is timely and a good opportunity to take stock of developments, debates and options. The Alliance aims to champion the use of and demand for evidence that is rigorous, accessible and appropriate. This raises questions about whether it is possible to reach a workable consensus on the best ways of identifying and labelling such evidence. There are also questions about how to increase the likelihood that this evidence actually informs decision making. The Alliance's remit tends to assume that there is a stand-alone notion of evidence. Yet this raises a further crucial issue about whether evidence ever really exists in isolation: perhaps information only really becomes evidence in the social context of its application.

This briefing paper is concerned with what counts as good evidence. It acknowledges that what counts as high-quality evidence for social policy is a contentious and contested issue. It outlines various approaches, standards and criteria used by different 'kitemarking' bodies to assess strength of evidence. It considers debates surrounding the merits and limitations of different approaches and examines what we know more generally about the effects of schemes that seek to endorse evidence. Much of this debate assumes that evidence stands separate from the context in which it is used and we also discuss the implications of this.

Research as evidence

Throughout this paper our focus is on evidence that is underpinned by research rather than expert opinion or stakeholder views. Much of the debate about evidence quality is couched in these terms and there are several reasons why in this debate we (and others) privilege research as a way of knowing. The conduct and publication of research involves documentation of methods, peer review and external scrutiny. These features contribute to its systematic nature and they provide a means to judge trustworthiness of findings. They also offer the potential to assess the validity of one claim compared to another.

However, there are other ways of knowing things. One schema (Brechin and Siddell, 2000) highlights three different ways of knowing:

- **Empirical knowing** – the most explicit form of knowing, which is often based on quantitative or qualitative research study;
- **Theoretical knowing** – which uses different theoretical frameworks for thinking about a problem, sometimes informed by research, but often derived in intuitive and informal ways;
- **Experiential knowing** – craft or tacit knowledge built up over a number of years of practice experience.

It is not easy to maintain strict distinctions between these categories and there is a lot of interaction between them. For example, empirical research may underpin each of the other two categories. It may also be a means of gaining more systematic understanding of the experiences of practitioners and of those who use public services. The debate about evidence quality tends to focus on standards for judging empirical research studies. However, there is variation in the extent to which theoretical and experiential knowledge are also factored in, especially in schemes that seek to endorse particular practices or programmes.

Research for many applications

Our overall argument is that evidence quality depends on what we want to know, why we want to know it and how we envisage that evidence being used. In varying contexts, what counts as good evidence will also vary considerably.

Much of the time it is assumed that what policymakers, service commissioners and practitioners want to know is whether various practices and programmes are effective – the 'what works' question. This is indeed a key concern but it usually sits alongside other important and complementary questions (Petticrew and Roberts 2003; Cameron et al., 2011). Decision makers are interested in why, when and for whom something works, and whether there are any unintended side–effects that need to be taken into account. They are also concerned about costs and cost–effectiveness, and with the distributional effects of different policies. Public perceptions about the acceptability of a particular practice will also need to be considered. Moreover, decision makers will want to know about the risks and consequences of implementation failure. What will be the repercussions of trying something if it subsequently fails to deliver anticipated outcomes and impacts? A stronger case is likely to be needed for high–risk ventures.

More broadly, decision makers need descriptive evidence about the nature of social problems, why they occur, and which groups and individuals are most at risk. Additionally, those working in policy and practice domains can benefit from the 'enlightenment' effects of research – research findings and theoretical debates can shed light on alternative ways of framing policy issues with implications for how they might be addressed (Weiss, 1980). For example, should young carers be viewed as disadvantaged youth, social policy assets, part of a hidden and exploited workforce, and/or as a group requiring support in their own right?

It may be possible for sub–groups of stakeholders to reach agreement about what counts as good evidence in response to each of the questions and concerns raised above. However, overall consensus is likely to be an unreachable goal. There will always be dissenting voices and alternative perspectives. Quality judgements are contested because ultimately 'evidence' and 'good evidence' are value labels attached to particular types of knowledge by those able to assert such labelling (Foucault 1977). In any decision–making setting there will be people with greater power than others to assert what counts as good evidence, but this does not mean that the less powerful will agree.

2 PRACTICE RECOMMENDATIONS

The lay of the land

There are many bodies in the UK and around the world that provide practice recommendations variously labelled as good practices, best practices, promising practices, research-based practices, evidence-based practices and guidelines. In the UK these bodies include government agencies, independent public bodies, professional associations, public service providers from the public and charity sectors, audit and inspection bodies, academic research centres and collaborations (see Box 1 for examples). Their advice is often focused on particular policy domains (e.g. health, education, welfare, social care, etc.) and/or specific target groups (e.g. patients, children and families, older people, offenders, and substance misusers).

Box 1: Some examples of bodies that highlight practices from which others can learn

SCIE (Social Care Institute for Excellence) is an independent charity, funded by the UK, Wales and Northern Ireland governments. It identifies and disseminates the knowledge base for good practice in all aspects of social care throughout the United Kingdom. (www.scie.org.uk)

The EPPI-centre (Evidence for Policy and Practice Information and Co-ordinating Centre) is part of the Social Science Research Unit at the Institute of Education, University of London. It maintains an online evidence library that provides access to its systematic reviews of evidence relating to topics in education, social policy, health promotion and public health. (www.eppi.ioe.ac.uk)

Project Oracle is part of the Mayor of London's Time for Action programme. It offers a developing evidence hub that aims to understand and share 'what really works' in youth programmes in London (building up from provider experience). It seeks to offer an innovative space in which providers can interact and learn from each other. (www.project-oracle.com)

Ofsted (Office for Standards in Education, Children's Services and Skills) is an independent publicly-funded body that inspects a wide range of services in care and learning and shares examples of good practice through its website and conferences. (www.ofsted.gov.uk)

NPIA (National Police Improvement Agency) is funded primarily by the Home Office. It offers practice advice developed through research and consultation with stakeholders. The Agency aims to assist practitioners by promoting good practice. (www.npia.police.uk)

There is no shortage of advice therefore, but there is often some uncertainty about the provenance and supporting evidence for many of the recommendations. Moreover, there is no standard nomenclature that would immediately indicate the type of evidence underpinning the labels attached to particular practices. More detailed reading of the

recommendations may reveal something about the nature of the supporting evidence; however, the reader will still be faced with a dilemma about what weight and judgement to attach to different forms of evidence. Does the evidence need to come from multiple respected sources? What role does methodology play in providing reassurance? Does the evidence need to be compelling or just good enough?

An additional uncertainty is likely to arise about whether a practice that is said to work well in one context will work equally well if applied in another. Such doubts may be further compounded by confusion due to the availability of contradictory advice from different sources.

In the face of all this uncertainty, it is not surprising that many policymakers, service commissioners and practitioners often rely on personal experience and advice from people that they consider to be experts in the field. Is there a way of moving beyond this? Is it possible to introduce more clarity about the standing of various practice recommendations?

From possibly helpful to proven practices

There are suggestions about how we might clarify the standing of different forms of advice. For example, advice might be rated according to the degree of confidence it provides that a practice is effective and will improve outcomes for a specific group. In commenting on advice for children and family services, Perkins (2010) offers the following definitions:

- **Good practice** – 'we've done it, we like it, and it feels like we make an impact';
- **Promising approaches** – some positive findings but the evaluations are not consistent or rigorous enough to be sure;
- **Research–based** – the programme or practice is based on sound theory informed by a growing body of empirical research;
- **Evidence–based** – the programme or practice has been rigorously evaluated and has consistently been shown to work.

The intention of such a listing is to standardise the way in which we talk about recommended practices. In the US, the Washington State Legislature has gone so far as to produce legal definitions of such categories, although there is unease about the adequacy of these definitions (WSIPP 2012). The list from Perkins (above) indicates that, even with an 'evidence–based' label, there can be uncertainty about what evidence would count as 'good enough' to warrant such labelling.

3 HIERARCHIES OF EVIDENCE

Hierarchies of evidence based on study design

Classifications of the degree of evidential support for practices (such as the one provided by Perkins) raise questions about what criteria are used to make judgements about the rigour of the evidence base. In Annex 1 we provide some examples of standards of evidence schemes. As is clear from these examples, across several policy areas, study design has generally been used as the key marker of the strength of evidence. This is then moderated by critically appraising the quality with which a study was conducted.

When the research question is 'what works?', different designs are often placed in a hierarchy to determine the standard of evidence in support of a particular practice or programme (see Box 2). These hierarchies have much in common; randomised experiments with clearly defined controls (RCTs) are placed at or near the top of the hierarchy and case study reports are usually at the bottom.

Box 2: Two illustrations of simplified hierarchies of evidence based on study design

- **Level I:** Well conducted, suitably powered randomised control trial (RCT)
- **Level II:** Well conducted, but small and under powered RCT
- **Level III:** Non-randomised observational studies
- **Level IV:** Non-randomised study with historical controls
- **Level V:** Case series without controls

Source: Bagshaw and Bellomo 2008, p.2.

1. Systematic reviews and meta-analyses
2. RCTs with definitive results
3. RCTs with non-definitive results
4. Cohort studies
5. Case control studies
6. Cross-sectional surveys
7. Case reports

Source: Petticrew and Roberts 2003, p.527.

There are differences in the numbers of levels in such hierarchies, and the status accorded to systematic reviews and meta-analyses can also vary. In Box 2, cross-study synthesis methods are placed at the top of one hierarchy, but they are not mentioned in the other. Systematic reviews and meta-analyses are important when drawing together evidence from a range of studies that have studied a standard intervention (they are most commonly used in the assessment of medical treatments). In such instances, basing effectiveness judgements on the result of a single study, even if it is an RCT, can be dangerously

misleading. However, when the intervention being studied is complex, variable and context–dependent (such as new models of service delivery), there are dangers in using meta–analysis as a way of reading across diverse studies because of the important influence of study context on specific findings.

Versions of evidence hierarchies are used by many evidence review groups and endorsing bodies around the world, and they are particularly prevalent in healthcare internationally and in other policy areas in the US.

Challenges to hierarchies based on study design

The premise, structure and use of such hierarchies have been the source of much debate and here we touch upon five key issues:

- Hierarchies based on study design neglect too many important and relevant issues around evidence;
- Hierarchies based on study design tend to underrate the value of good observational studies;
- Using such hierarchies to exclude all but the highest–ranking studies from consideration can lead to the loss of useful evidence;
- Hierarchies based on study design pay insufficient attention to the need to understand what works, for whom, in what circumstances and why (programme theory);
- Hierarchies based on study design provide an insufficient basis for making recommendations about whether interventions should be adopted.

Each of these concerns is elaborated below.

Hierarchies based on study design are too narrow

Many traditional hierarchies tended to place more emphasis on study design than on a critical appraisal of how that design was implemented and how it fits with other studies on the same issue. They have subsequently been revised to respond, at least in part, to these criticisms.

In health, an informal working group was established to generate broad consensus for a revised classification system that addressed many of the shortcomings of traditional hierarchies based on study design. The result is the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system (Atkins et al., 2004). This defines ‘quality of evidence’ as the amount of confidence that a clinician may have that an estimate of effect from research evidence is in fact correct for both beneficial and harmful outcomes. Quality of evidence 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.

In reaching this judgement, the GRADE system starts by rating the available evidence on the basis of the study designs used. It then considers other factors that may affect the initial grading, including:

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

Although an RCT design is initially rated more highly than other designs, the final rating of the evidence emerging from a group of studies can change when these other factors have been taken into account.

The GRADE system has been adopted by many health bodies, including NICE, and it is seen as an improvement over traditional hierarchies based on study design. There are, however, concerns that GRADE's consideration of moderating factors does not go far enough. For example, Bagshaw and Bellomo (2008) argue for the inclusion of other aspects of evidence quality 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.

Another concern is that the GRADE system still focuses too narrowly on the question of what works, which means that large swathes of data are excluded. As we discuss later, these data are relevant to understanding whether an intervention addresses a problem that matters, whether it is acceptable to service users, and how its success might vary across groups or contexts (Petticrew and Roberts 2006).

Hierarchies based on study design underrate good observational studies

There is a growing body of literature which argues that hierarchies based on study design tend to undervalue the strength of evidence produced by well-conducted observation studies (e.g. Bagshaw and Bellomo 2008; Cook et al., 2008; Konnerup and Kongsted 2012). The argument is that certain observational study designs are capable of delivering unbiased estimates of the effects of interventions (that is they score highly on internal validity). At the same time, they tend to score highly on generalisability too (external validity) because they involve large, representative sample sizes. For these reasons, they can provide stronger evidence, and a more secure basis for practice recommendations, than single-centre RCTs.

Another advantage is that observational studies tend to be less expensive than RCTs. Good observational studies are particularly cost-effective when precise and unbiased measurements of a broad set of outcome variables are available from administrative data (as is said to be the case in Scandinavian countries – Konnerup and Kongsted 2012).

Using hierarchies to exclude all but the 'best' studies loses useful information

In systematic reviews of evidence about a particular practice or programme, evidence hierarchies are frequently used as a filtering device to ensure that review findings are based only on the strongest studies. This has the added advantage of reducing the number of studies to be considered in detail to a manageable level. For example, Konnerup and Kongsted (2012) found that the majority of Cochrane Collaboration reviews are limited to considering RCTs only. Critiques argue that this means that potentially important evidence is overlooked and this weakens the value of evidence syntheses (Ogilvie et al., 2005).

Pawson (2003) illustrates this point with regard to three studies relating to public disclosure of the identities of sex offenders. None of the three studies would have made it through hierarchical study design quality filters. However, he argues, together they provide a plausible account of why public disclosure is limited in its ability to reduce reoffending by sex offenders. He concludes that using evidence hierarchies as a technical filter prior to research synthesis is wasteful of useful evidence and can lead to misleading conclusions.

Hierarchies based on study design pay insufficient attention to programme theory

Many social programmes are complex and multifaceted. There is a need therefore to unpack the relevant components of the 'black box' in order to model multiple causal linkages and influences, and thus gain a better understanding of how a programme works (Chatterji 2008).

Some standards of evidence schemes do require verification of an underpinning theoretical rationale as well as evidence from an RCT study design to achieve the label of 'best' evidence. See, for example, the standards of evidence produced by the Social Research Unit at Dartington (Annex 2). However, this may not be considered necessary or even desirable by those with an overriding empiricist approach. The resulting debate reflects a deep epistemological divide about the necessity of experiments and the value of theory.

Advocates for more attention to be paid to programme theory are especially concerned about the use of experimental methods when intervention effects are heterogeneous. When an intervention works for some but not for others (perhaps even being harmful for some), looking at aggregate effects is misleading. This has led to calls for more attention to be paid to theory–driven evaluation and different forms of evidence synthesis. One response has been the development of realist synthesis methods which bring together theory, quantitative and qualitative evidence with the aim of shedding light on 'what works, for whom, in what circumstances, in what respects and why?' (Pawson et al., 2005).

Hierarchies based on study design provide an insufficient basis for recommendations

At the end of an evidence review process, there is the issue of whether allocating an evidence level to a practice or programme should be directly linked to a recommendation about its use. Many providers of systematic reviews (e.g. CRD, Cochrane Collaboration, and the EPPI–Centre) stop short of making recommendations about whether a practice or programme should be used. Organisations that provide registers of evidence–based practices and programmes sometimes signal their recommendations by labelling programmes as 'Proven', 'Model' or 'Promising' (e.g. RAND's Promising Practices Network,

and University of Colorado's Blueprints for Violence Prevention). Others, such as the National Registry of Evidence–based Programs and Practices (NREPP) do not make a recommendation, but they do provide information about whether an intervention or programme is ready for dissemination.

The GRADE working group, referred to earlier, tackled this issue and concluded that there was a need for a separate 'strength of recommendation' assessment. This assessment indicates the extent to which a practitioner can be confident that adherence to the recommendation will result in greater benefit than harm for the patient/service user. The 'strength of recommendation' builds on the 'quality of evidence' assessment by incorporating additional factors such as target patient population, baseline risk and individual patients' values and costs (Bagshaw and Bellomo 2008).

Of course, the leap from 'quality of evidence' to 'decision to apply' can never be a simple technocratic choice. It will necessarily involve judgement and political considerations.

4 BEYOND HIERARCHIES?

A matrix of evidence

In the social policy field, there are many aspects of knowledge development and evidence identification that raise questions about the appropriateness and feasibility of standards of evidence based primarily on a hierarchy of study designs. Central here are concerns about the availability, appropriateness and feasibility of RCT designs in social policy (see Box 3). In this section, we focus in particular on the need for standards of evidence to address more than the question of what works.

Box 3: Challenges to using hierarchies of evidence based on study design in social policy

- There is a paucity of studies based on experimental designs (particularly outside of the US) and this could lead to a misleading conclusion that we have very little evidence on which to base practice.
- Research funding levels in many social policy areas are inadequate to support rigorous multi–centre RCTs.
- It is often difficult or impossible to implement rigorous RCT designs that ensure that service recipients, practitioners and analysts are unaware of whether subjects are in experimental or control groups (i.e. completely blinded RCT designs).
- The breadth and complexity of practice can mean that even unblinded RCT designs for assessing effectiveness may be inappropriate.
- Such hierarchies focus too narrowly on the question of what works whereas those interested in evidence–based practice also want answers to other questions such as what matters and what is acceptable.
- A commitment to a participatory approach to service development (which involves service users, practitioners and evaluators working together) emphasises research designs that would typically score low on such scales.

Source: Bagshaw and Bellomo 2008, p.2.

Petticrew and Roberts (2003) argue that we need to think more in terms of a matrix rather than a hierarchy of evidence, even for seemingly straightforward questions about what works. Here types of research design are differentially rated according to the research question being addressed. They argue that policymakers and practitioners are interested in at least eight questions (see Box 4) and that RCT designs are inappropriate for answering half of these.

Box 4: A matrix of evidence to address various aspects of 'does this work?'

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.

We now discuss two bodies that have adopted a broad matrix of evidence approach: the EPPI–Centre and SCIE.

The EPPI–Centre

The ideas underpinning a matrix of evidence approach are used by the EPPI–Centre (Institute of Education, University of London) in its methods for producing systematic evidence reviews. The EPPI–Centre has developed a 'weights of evidence' framework, which weights each study identified as potentially relevant for a particular review question according to three dimensions: (a) internal validity; (b) appropriateness of study method; (c) appropriateness of samples, context and measures. There are standards of evidence guidelines for each main type of research question or design. After rating each of the three dimensions separately, they are combined into an overall weight of evidence judgement (high to low). Findings of lower–quality studies are either excluded or given less weight in syntheses of evidence.

Social Care Institute for Excellence (SCIE)

SCIE guidance for systematic research reviews uses a similar approach to appraising the quality of research studies (see Rutter et al., 2010). SCIE research reviews usually evaluate the effectiveness of interventions, but they also address other questions including how and why interventions work, and broader questions of policy and practice.

In theory the SCIE reviews aim to incorporate knowledge from the five sources identified by Pawson et al., (2003):

- Organisational knowledge;
- Practitioner knowledge;
- User knowledge;
- Research knowledge;
- Policy community knowledge.

(See Annex 3 for further details.)

In practice, SCIE reviews draw primarily on knowledge derived from empirical research. However, where the views and experiences of users and carers are not available through research other forms of user and carer testimony are taken into account. SCIE research reviews do not apply a formal set of evidence standards, but their review guidelines do provide a list of minimum criteria to be considered when assessing the quality of a wide variety of empirical studies. The resulting strength of evidence judgements are based on the same three dimensions used by the EPPI–Centre.

5 STRONG EVIDENCE OR GOOD ENOUGH EVIDENCE?

The 'evidence journey'

Evidence-endorsing schemes vary in the extent to which they focus on ensuring that recommended practices and programmes are underpinned by what they consider to be the strongest levels of evidence. For example, the Coalition for Evidence-Based Policy in the US focuses on 'top tier' or 'near top tier' interventions. Others have sought to recognise practices and programmes that may be helpful but are not yet fully evidence-based. For example, SCIE produces a range of knowledge products that are underpinned by different 'levels' of evidence (see Box 4). Recognition via the Good Practice Framework may be the first stage in a journey to becoming fully recognised as an evidence-based practice.

Box 5: Four of SCIE's knowledge products

The **Good Practice Framework** is an online facility to help social care professionals put forward good practice examples for others to see and use. In order that people can trust these practice examples, the submission and review process uses a combination of guided self-evaluation, external review and classification of submitted examples.

A **Practice Enquiry** is primary research conducted by SCIE – involving survey, qualitative and/or case study methods – to draw out knowledge about a topic from practice environments.

Research Briefings provide structured accounts of the research on a given topic, based on a systematic but limited search of the literature for key evidence. Because SCIE do not thoroughly assess the quality of the research identified, a research briefing acts as a signpost for further reading, rather than as a definitive account of what works.

Knowledge Reviews provide the strongest levels of evidence on a given topic. They combine knowledge from a systematic research review with knowledge from practice environments (often gained from a Practice Enquiry).

Source: Bagshaw and Bellomo 2008, p.2.

Some standards of evidence schemes enable endorsing bodies to document where a practice or programme is on this 'evidence journey'. A good example is the standards of evidence for assessing intervention effectiveness developed by the Social Research Unit at Dartington. These standards consider four factors or dimensions when evaluating the evidence in support of an intervention:

- Evaluation quality (study design and conduct quality).
- Intervention impact (sizable and significant effects with no adverse impacts).

- Intervention specificity (clear target population, intended outcomes and programme logic).

- System readiness (documented implementation processes and resources).

For each factor, a set of criteria are articulated for 'good enough' as well as 'best' evidence (see Annex 2).

What counts as good enough evidence depends on how we envisage it being used.

There is a reasonable consensus that the answer to what counts as good evidence depends on the type of research/policy/practice question to be answered. We also need to factor in what the evidence will be used for (e.g. option generation, decision making, ongoing learning and development, continuing to do something, stopping doing something, and developing innovative ways of working).

Much of the debate around standards of evidence has focused on an instrumentalist view of evidence use, which involves the direct application of research to policy and practice decisions. However, we know that research and other sources of evidence are often used in much more indirect and subtle ways. In these instances, use may be as much about shaping attitudes and ways of thinking as having a direct influence on decision making – often referred to as the enlightenment impact of research (Weiss 1980, Nutley et al., 2007).

If our interest is in reframing and re-problematising policy and practice concerns, this will in part be achieved through developing new concepts, models and theories. These necessitate different ways of thinking about standards of evidence. Quality appraisal is more complex for such studies, which may or may not have empirical underpinnings. Rutter et al., (2010) suggest that non-empirical studies should be assessed for topic relevance, methodological fitness for purpose, and the scope or selective nature of the material on which they are based. They also point to the need to be aware of potential conflicts of interest in such material (although conflicts of interest may exist in other types of material too).

An interest in reframing and re-problematising policy and practice concerns is also likely to emphasise different ways of producing evidence: for example, through engaged scholarship, ongoing dialogue and iterative co-production processes (Nutley et al., 2007). These diverse ways of producing evidence, involving multiple different groups (e.g. researchers, practitioners, policymakers), add to the challenges of assessing the quality of the evidence so produced.

6 THE USE AND IMPACT OF STANDARDS OF EVIDENCE AND ENDORSING PRACTICES

Key questions

What do we know about how standards of evidence are used in practice and what their impact has been? Have they changed research practice? Have they influenced policymakers' and practitioners' views on what counts as good evidence? Have they influenced decision–making processes? What have been the impacts of schemes that have certified the quality of particular intervention programmes or particular evidence providers? Do various stakeholders have confidence in such schemes? Have they changed the ways funds are allocated or the way services are commissioned? What impact have they had on service delivery?

The evidence base for answering these and other related questions is very patchy. We know of no systematic or realist research reviews that have sought to tackle such questions. For this reason, we are not able to address all of the above questions and where we do offer comments these are somewhat speculative as they rely on a limited collection of studies, supplemented by personal experience.

Impact on perceptions of what counts as good evidence

Standards of evidence may well have changed the practice of research, particularly where there is a degree of consensus about what constitutes good evidence. Such impacts are likely to be around both choice of methodology and the detailed conduct of studies to address quality concerns. However, in the social sciences there is only limited consensus as to what constitutes good evidence (Rutter et al., 2010, Sempik et al., 2007). Moreover, consensus tends to be greater in relation to quality criteria for quantitative research than it is for qualitative research (Rutter et al., 2010; Spencer, Ritchie et al., 2003; Dixon–Woods, Bonas et al., 2006). Indeed, a survey of social policy researchers in the UK found that they did sometimes think in terms of a hierarchy of methods, but that this was an inverse of the traditional hierarchies of evidence by study design referred to above: qualitative methods were placed at the top and experimental methods at the bottom (Sempik et al., 2007).

Policymakers' and practitioners' views on what counts as good evidence seem to be reasonably persistent and resilient to explicit standards of evidence (perhaps because of their contested nature). In general, they are interested in persuasive and actionable evidence but these qualities are not necessarily linked to particular study designs (Cameron et al., 2011). A former Deputy Chief Social Researcher in UK central government has reflected that policymakers' hierarchy of evidence tends to place research evidence at the bottom of the hierarchy, below 'cabbies' evidence' (see Box 5). While fairly tongue–in–cheek, such observations suggest that for all the technical debate over evidence quality, more work may need to be done with potential users.

Box 6: One insider's view of policymakers' hierarchy of evidence

1. Expert evidence (including consultants and think tanks).
2. Opinion–based evidence (including lobbyists/pressure groups).
3. Ideological evidence (party think tanks, manifestos).
4. Media evidence.
5. Internet evidence.
6. Lay evidence (constituents' or citizens' experiences).
7. Street evidence (urban myths, conventional wisdom).
8. Cabbies' evidence.
9. Research evidence.

Source: Phil Davies, former Deputy Chief Social Researcher, 2007.

Use of practices and programmes that are endorsed as evidence–based

Perceptions of interventions that have been labelled as proven, promising or recommended seem to vary. Given that many such interventions in the social policy field originate from the US, there is scepticism about their transferability to other countries and contexts. Even homegrown programmes can suffer from concerns about whether they are transferable from one area of the country to another.

Evidence from the health field is similarly discouraging. There is quite a lot of literature documenting the non–implementation of NICE guidelines (e.g. Spyridonis and Calnan 2011; Kidney et al., 2011; Al–Hussaini et al., 2012). Moreover, Kidney et al., (2011) found that the level of evidence underpinning NICE recommendations was not an important factor influencing their adoption in practice.

More encouragingly, there is evidence from the US that advocacy groups promoting the importance of investing in 'proven' programmes, such as the Coalition for Evidence–Based Policy, have influenced the funding patterns of several Federal Government departments (Haskins and Baron 2011). They also seem to have had a profound effect on the demand for particular forms of research and the ways in which research is supplied. There is, however, less agreement about whether this is wholly a good thing (Mason, 2011). These initiatives may have encouraged the increased adoption of proven or promising programmes. They seem to have been less effective in encouraging state and local governments to stop doing things for which there is no evidence of effectiveness or which have been shown to be ineffective (Weiss et al., 2008; Haskins and Baron 2011).

Impact on innovation and encouraging evidence–based practice

In several evidence–based funding regimes a proportion of funds are reserved for new interventions that do not yet meet the standards of evidence required of recognised programmes, in order to facilitate ongoing innovation. Nevertheless top–down schemes that endorse 'proven' practices may be discouraging because so few practices and programmes appear to reach the standards required for recognition. For example, the Blueprints for Violence Prevention initiative has reviewed 900 programmes and only 11 have been designated as model programmes, with a further 19 being rated as promising (Taxman and Belenko 2012).

In response to concerns about top–down schemes, there are advocates for a more bottom–up approach to defining and encouraging evidence–based practices. For example, Hogan et al., (2011) discuss the approach taken by the Singapore national government as it seeks to maintain the excellent reputation of its education system. Here the emphasis has been on facilitating local autonomy at school and school cluster levels. A top–down process of knowledge dissemination around effective practices is rejected in favour of shifting the locus of knowledge production to schools so that they co–produce the research agenda and collaborate with researchers in knowledge creation and on–going learning.

Use of bottom–up schemes for encouraging evidence–based practice

Within the UK, several initiatives have used the standards of evidence produced by the Social Research Unit at Dartington to develop a bottom–up approach to encouraging and facilitating evidence–based practice (Annex 2). These schemes focus on recognising and accrediting the developmental stages of an intervention.

For example, in Project Oracle service providers conduct a self–assessment of their interventions using a practitioner guidebook. This sets out five evidence levels. Level 1 is the entry level and this requires a sound theory of change or logic model with clear plans for evaluation. Level 5 is the highest level and signifies that an intervention has been subject to multiple independent evaluations and cost–benefit analysis.

Service providers submit evidence to Project Oracle to justify their self–assessment. Oracle staff then validate the evidence level for an intervention and work with the provider to agree a detailed action plan to improve the evidence–base for the intervention. The rationale is that improving the evidence for an intervention will also improve the practice itself.

One of the potential limitations of the Project Oracle approach is that it relies on self–nomination. It is too early to tell yet whether there will be sufficient interest amongst providers, and sufficient resources within the project itself, to make it work effectively on a large scale in the longer term (Ilic and Bediako 2011).

A similar developmental approach is proposed by Nesta in its standards of evidence for impact investing (Puttick and Ludlow 2012). Nesta have adopted a modified version of the standards of evidence used by Project Oracle. The rationale for this is that these standards are seen to retain academic standards of rigour whilst ensuring that the evidence requirements are appropriate to ongoing innovation and development of services and products.

Impact of accrediting evidence providers

Schemes which focus on accrediting evidence providers, such as NHS Evidence and the Information Standard (see Annex 1) seem to be popular with the accredited institutions. However, we are not aware of any independent evaluation of their effectiveness. For example, are accreditation processes suitably rigorous? Do the endorsements associated with accreditation steer evidence users in the direction of these evidence providers?

A study by Fackrell et al., (2012) suggests that website accreditation may be a blunt instrument. The study used the DISCERN instrument to score websites according to the reliability, quality and trustworthiness of the healthcare information they provided. It found that both the highest and lowest ranked websites in the study had received accreditation under the Information Standard.

In summary, our knowledge about the impact (positive or negative) of standards of evidence and endorsement schemes is limited. It is important that we improve our knowledge on the impact of existing schemes and build in sensitive evaluations when embarking on new schemes. What we do know is not wholly encouraging and this suggests that there is a need to revisit and reconsider the, often implicit, 'programme theories' underpinning various schemes.

7 CONCLUSIONS AND WAYS FORWARD

There is no simple answer to the question of what counts as good evidence. It depends on what we want to know, for what purposes, and in what contexts we envisage that evidence being used. Research data only really become information when they have the power to change views, and they only really become evidence when they attract advocates for the messages they contain. Thus endorsements of data as 'evidence' reflect judgements that are socially and politically situated.

Standards of evidence should pay attention to the need for different types and qualities of knowledge when addressing a necessarily diverse range of policy and practice questions. Developing standards of evidence that respond to such concerns is not likely to be a straightforward task.

Matrices of evidence offer a helpful way forward. However, for any policy or practice question, there will be conflicting views about the merits of different forms of evidence. As we have discussed in relation to the 'what works' question, there are divergent views about the role of theory, the use of meta-analysis, and the relative merits of observational studies vis-à-vis experimental studies. It is not likely that all these differences of perspective can be resolved through dialogue and debate: choices are necessary.

In theory we may be able to separate different policy and practice questions in order to match them with appropriate standards of evidence. However, there is a need to recognise that, in practice, decision makers need to consider evidence in the round. Policymakers and practitioners need to weigh evidence relating to what works alongside evidence about cost, acceptability and distributional effects.

The purposes served by standards of evidence should also be clarified. These will affect the criteria used and the labels attached to different levels of evidence. Are standards mainly used to endorse what are considered to be best practices and proven programmes? Or do they serve a more developmental purpose aimed at improving both practices and the available evidence? We see a lot of merit in a developmental approach that seeks to encourage progress through some assessed stages on an evidence journey. Our view is that there is no natural end to this journey: all evidence is partial, provisional and contingent, and thus needs to be used as part of an ongoing process of evaluation, learning, adaptation and innovation (Sanderson, 2009).

What does all this mean for the prospects of setting up a respected and authoritative voice on what works in the social policy field: a social policy NICE? It should come as little surprise that while we recognise the initial attractions of establishing such a body, we have significant concerns about the challenges that it would face. We would echo many of the concerns raised by several commentators (e.g, Walker 2012; Corry 2012). In particular, the complex and contested nature of social research sets it apart from the clinical research evidence that is typically considered by NICE. A social policy NICE would face a tougher challenge in developing evidence- and consensus-based guidelines for practice. There is also the question of whether such guidelines would actually influence decision making: evidence on the extent to which NICE guidelines are implemented is not encouraging (Spyridonis and Calnan 2011; Kidney et al., 2011; Al-Hussaini et al., 2012). This is perhaps unsurprising. It has long been acknowledged that policymakers and practitioners make decisions in environments in which they are subject to multiple, often competing, influences and concerns – of which 'evidence' is only one, and a highly-contested one at that (Nutley et al., 2007).

Our overall conclusion is that there is a need to debate standards of evidence in order to develop understanding of different viewpoints. The outcome is likely to be a range of standards of evidence schemes – one size will not fit all the purposes and perspectives that such schemes serve. There is no doubt in our mind that standards of evidence are an essential component of developing more evidence–informed policy and practice, but there are dangers in these becoming too fixed, rigid and prescriptive. Moreover, experience shows that we should remain realistic about the extent to which they will actually shape decision making on the ground.

REFERENCES

Al-Hussaini , A., Owens, D. and Tomkinson, A. (2012) Have two UK national guidelines had any effect on grommets day-case utilisation and rate over the last 10 years? 'Eur Arch Otorhinolaryngol.' 269: 2053-2056.

Atkins, D., Eccles, M. et al., (2004) Systems for grading the quality of evidence and the strength of recommendations 1 : critical appraisal of existing approaches. The Grade Working Group. ' BMC Health Services Research.' 4(1): 38.

Bagshaw, S. and Bellomo, R. (2008) The need to reform our assessment of evidence from clinical trials: A commentary. 'Philosophy, Ethics, and Humanities in Medicine.' 3:23 (doi:10.1186/1747-5341-3-23).

Brechin, A. and Siddell, M. (2000) 'Ways of knowing', In Gomm, R. and Davies, C. (eds) 'Using evidence in health care.' Buckingham: Open University Press.

Cameron, A., Salisbury, C., Lart, R., Stewart, K., Peckham, S., Calnan, M., Purdy, S. and Thorp, H. (2011) Policy makers' perceptions on the use of evidence from evaluations. 'Evidence & Policy.' 7(4): 429-47.

Chatterji, M. (2008) Synthesizing evidence from impact evaluations in education to inform action. 'Education Researcher.' 37 (1): 23-26.

Cook,T., Shadish, W. and Wong, V. (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. 'Journal of Policy Analysis and Management.' 27: 724-50.

Corry. D, (2012) A Nice idea for social policy – but maybe not for charities. 'The Guardian Policy Hub.' (http://www.guardian.co.uk/public-leaders-network), 22 June 2012.

Davies, P. (2007) 'Types of Knowledge for Evidence-Based Policy.' Presentation to NORFACE Seminar on Evidence and Policy. University of Edinburgh, 26 November 2007.

Dixon-Woods, M., Bonas, S. et al., (2006) How can systematic reviews incorporate qualitative research? A critical perspective. 'Qualitative research.' 6(1): 27-44.

Fackrell, K., Hoare, D., Smith, S., McCormack, A. and Hall, D. (2012) An evaluation of the content and quality of tinnitus information on websites preferred by general practitioners. 'BMC Medical Informatics and Decision Making.' 12: 70.

Foucault, M. (1997) 'Discipline and punish.' Harmondsworth: Penguin.

Haskins, R. and Baron, J. (2011) 'Building the connection between policy and evidence: the Obama evidence-based initiatives.' In Puttick, R. (ed) 'Using Evidence to Improve Social Policy and Practice.' London: Nesta/Alliance for Useful Evidence.

Hogan, D., Teh, L. and Dimmock, C. (2011) 'Educational Knowledge Mobilization and Utilization in Singapore.' Paper prepared for the 2011 Conference of the International Alliance of Leading Educational Institutions, OISE, University of Toronto.

Huxley, P., Maegusuku-Hewitt, T., Evans, S., Cornes, M., Manthorpe, J. and Stevens, M. (2010) Better evidence for better commissioning: a study of the evidence base of generic social care commissioning guides in the UK. 'Evidence & Policy.' 6(3): 291-308.

Ilic, M. and Bediako, S. (2011) 'Project Oracle: understanding and sharing what really works', in Puttick, R. (ed) 'Using Evidence to Improve Social Policy and Practice.' London: Nesta/Alliance for Useful Evidence.

Johnston, M., Vanderheiden, G., Farka, M., Rogers, E., Summers, J. and Westbrook, J. (2009) 'The challenge of evidence in disability and rehabilitation research and practice: a position paper.' NCDDR Task Force on Standards of Evidence and Methods, Austin, TX: SEDL.

HM Government (2012) 'Civil Service Reform Plan.' London: HM Government. http://www.civilservice.gov.uk/wp-content/uploads/2012/06/Civil-Service-Reform-Plan-acc-final.pdf

Kidney, E., Jolly, S. and Kenyon, S. (2011) Does the level of evidence underpinning NICE recommendations influence adoption into Trust maternity guidelines? 'Arch Dis Child Fetal Neonatal Ed.' 96(suppl 1): Fa75-Fa96.

Konnerup, M. and Kongsted, H. (2012) Do Cochrane reviews provide a good model for social science? The role of observational studies in systematic reviews. 'Evidence and Policy.' 8(1): 79-86.

Mason, S. (2011) 'The Federal Challenge to University-Based Education Research in the United States: Turning Research into Policy and Practice.' Paper prepared for the 2011 Conference of the International Alliance of Leading Educational Institutions, OISE, University of Toronto.

Nutley, S.M., Walter, I. and Davies, H.T.O. (2007) Using Evidence: 'How Research Can Inform Public Services.' Bristol: The Policy Press.

Ogilvie, D., Egan, M., Hamilton, V. and Petticrew, M. (2005) Sytematic review of health effects of social interventions: 2. Best available evidence: how low should you go? 'Journal of Epidemiology and Community Health.' 59: 886–892.

Pawson. R, (2003) 'Assessing the quality of evidence in evidence–based policy: when, why, how and when.' ESRC Research Methods Programme Conference, Buxton, May 2003.

Pawson, R., Boaz, A. et al., (2003) 'Types and quality of knowledge in social care.' London: SCIE.

Pawson, R., Greenhalgh, T., et al., (2005) Realist review – a new method of systematic review designed for complex policy interventions. ' Journal of Health Services Research and Policy.' 10(3 Supplement): 21–34.

Perkins, D. (2010) 'Fidelity–Adaptation and Sustainability.' Presentation to seminar series on Developing evidence informed practice for children and young people: the 'why and the what.' Organised by the Centre for Effective Services (www.effectiveservices.org) in Dublin, Cork and Galway in October 2010.

Petticrew, M. and Roberts, H. (2003) Evidence, hierarchies, and typologies: horses for courses. 'Journal of Epidemiology and Community Health.' 57: 527–529.

Petticrew, M. and Roberts, H. (2006) 'Systematic Reviews in the social sciences: a practical guide.' Oxford: Blackwell.

Puttick, R. and Ludlow, J. (2012) 'Standards of evidence for impact investing.' London: Nesta.

Rutter, D., Francis. J., Coren, E. and Fisher, M. (2010) 'SCIE systematic research reviews: guidelines (2nd edition).' London: Social Care Institute for Excellence.

Sanderson, I. (2009) Intelligent Policy Making for a Complex World: Pragmatism, Evidence and Learning. 'Political Studies.' 57(4): 699–719.

Sempik, J., Becker, S. and Bryman, A. (2007) The quality of research evidence in social policy: consensus and dissension among researchers. 'Evidence and Policy.' 3(3): 407–23.

Spencer, L., Ritchie, J., et al., (2003) 'Quality in qualitative evaluation: a framework for assessing research evidence.' London: Government Chief Social Researcher's Office.

Spyridonidis, D. and Calnan, M. (2011) Opening the black box: A study of the process of NICE guidelines implementation. 'Health Policy.' 102: 117–125.

Taxman, F. and Belenko, S. (2012) 'Identifying the evidence base for "what works" in community corrections and addiction treatment.' In chapter 2 of Taxman, F. and Belenko, S. 'Implementing evidence–based practice in Community Corrections and Addiction Treatment.' Springer Science+Business Media.

Walker, D. (2012) Would a version of the health standards body NICE work for social policy? 'The Guardian Policy Hub.' (http://www.guardian.co.uk/public–leaders–network), 11 June 2012.

Weiss, C., Murphy–Graham, E., Petrosino, A. and Gandhi, A. (2008) The Fairy Godmother—and Her Warts: Making the Dream of Evidence–Based Policy Come True.' American Journal of Evaluation.' 29 (1): 29–47.

Weiss, C. (1980) Knowledge creep and decision accretion. 'Knowledge: Creation, Diffusion, Utilization.' 1(3), 381–404.

WSIPP (2012) 'Inventory of Evidence–Based, Research–Based, and Promising Practices.' Washington State Institute for Public Policy (WSIPP) and University of Washington Evidence–based Practice Institute, September 2012, www.wsipp.wa.gov

ANNEX 1 EXAMPLES OF STANDARDS OF EVIDENCE SCHEMES

Name	Country and sector	Type/purpose
GRADE: Grading of Recommendations Assessment, Development and Evaluation (c. 2004) http://www.gradeworkinggroup.org/index.htm	UK, healthcare	The GRADE system categorises quality of evidence and strength of recommendations (and clearly separates the two). Quality of evidence is classified as high, moderate, low and very low. Evidence based on RCTs starts as 'high' evidence but may be moved down the scale (e.g. poor quality study, reporting bias). Observational studies (e.g. cohort and case-control studies) start with a 'low quality' rating but may be graded upwards (e.g. if the magnitude of treatment effect is very large or if all plausible biases would decrease the magnitude of an apparent treatment effect). The strength of recommendations is classified as 'strong' or 'weak.' Increasingly being adopted worldwide (sometimes with modifications e.g. combining the low and very low categories). The Scottish Intercollegiate Guidelines Network (SIGN) has incorporated the GRADE approach within its guideline development methodology.
National Institute for Health and Clinical Excellence (NICE) Guidelines (founded 1999) http://www.nice.org.uk/	UK, health and social care	Incorporates elements of the GRADE system; in addition NICE integrates a review of cost-effectiveness studies. Levels of evidence are classified from 1a (systematic review or meta-analysis of RCTs) to 4 (expert committee reports or opinions and/or clinical experience of respected authorities). Recommendations are graded from A (based directly on level 1 evidence) to D (based directly on level 4 evidence or extrapolated from level 1, level 2, or level 3 evidence); additional categories are GPP (Good practice point on the view of the guideline development group) and NICE TA (Recommendation taken from a NICE Technology Appraisal).
Project ORACLE (2010) http://www.project-oracle.com/	UK (London), youth services	The ORACLE standards offer five levels of evidence in assessing interventions. 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 replication evaluations and cost-benefit analysis. Oracle self-assessment is carried out by the provider alongside a practitioner guidebook; the organisation then submits evidence to justify its self-assessment at a given level. Oracle staff validate the level and work with the provider to agree a detailed action plan to improve their evidence base.
SCIE Good Practice Framework (date of introduction not known) http://www.scie.org.uk/goodpractice/learnmore.aspx	UK, social care	The principles and rationale on which the Framework is based include the concepts that practice-based knowledge can complement research evidence and that the experiences of service users and carers are essential measures of effective practice. The Good Practice Framework combines guided self-evaluation and external review and classification of submitted examples. Links to relevant research studies are given.

NHS Evidence Accreditation Mark (introduced 2009) http://www.evidence.nhs.uk/accreditation	UK, health and social care	Accreditation scheme for producers of guidance (including clinical guidelines, clinical summaries and best practice statements). The Accreditation Mark shows that the guideline producers meet a defined set of criteria in the processes they use to develop their products; it does not accredit the content of individual products. The scheme is based on internationally agreed criteria for guideline development (i.e. the AGREE instrument (2001)).
The Information Standard (introduced 2009) http://www.theinformationstandard.org/	England, health and social care	Certification scheme for all organisations producing evidence-based health and social care information for the public. Guidance states that RCTs and double blind trials "*are considered the most reliable form of primary research in the field of health and social care interventions*" but that in practice there are many situations where relevant research studies have not yet been done and that in those cases it is appropriate for organisations to base their information on the best available evidence, or on health professionals' experience or expertise or on the personal experiences of patients/service users, so long as this is clearly acknowledged.
Maryland Scale of Scientific Methods (1997) https://www.ncjrs.gov/pdffiles/171676.PDF	US, criminal justice	Five-point scale from Level 1 (lowest) to Level 5 (highest) for classifying the strength of methodologies used in 'what works' studies. Level 1 is correlation between a crime prevention programme and a measure of crime or crime risk factors at a single point in time. Level 5 is random assignment and analysis of comparable units to the programme and comparison groups. Level 3 (a comparison between two or more comparable units of analysis, one with and one without the programme) is deemed to be the minimum level to draw conclusions about effectiveness.
RAND Promising Practices Network (founded 1998) http://www.promising-practices.net/	US, children, families and communities	Programmes are assessed as 'Proven' or 'Promising' according to a range of evidence criteria e.g. effect size, statistical significance, use of comparison groups.
Top Tier Evidence Initiative (launched 2008) (Coalition for Evidence-Based Policy) http://toptierevidence.org/wordpress/	US, social policy	Identifies and validates 'Top Tier' and 'Near Top Tier' interventions; Top Tier Interventions are *"Interventions shown in well-designed and implemented randomized controlled trials, preferably conducted in typical community settings, to produce sizable, sustained benefits to participants and/or society."* The Top Tier initiative recognises that not all social problems currently have interventions that would meet the Top Tier (e.g. because of research gaps) and that public officials may need to rely on evidence that falls below the Top Tier. In these cases the initiative refers users to other high-quality resources that do review such evidence.
Blueprints for Violence Prevention (Centre for the Study and Prevention of Violence, University of Colorado at Boulder) (1996) http://www.colorado.edu/cspv/blueprints/	US, criminal justice	Reviews programmes according to a range of criteria including these three key criteria: evidence of deterrent effect with a strong research design; sustained effect; multiple replication. 'Model programs' must meet all three; 'promising programs' must meet at least the first.

What Works Clearinghouse (WWC) (Department of Education) (created 2002) http://ies.ed.gov/ncee/wwc/	US, education	Provides reviews of the research literature; also provides external users with templates that can be used to assess the quality of research studies according to WWC evidence standards.
Washington State Institute for Public Policy (WSIPP) Inventory of Evidence-Based, Research-Based and Promising Practices (2012) http://www.wsipp.wa.gov/pub.asp?docid=E2SHB2536	US, child welfare, juvenile justice and mental health	The inventory assigns programmes and practices to the relevant category (evidence-based, research-based and promising) according to current-law definitions of these terms and also to alternative definitions developed by WSIPP in consultation with stakeholders. The inventory builds on work previously conducted by WSIPP and will be updated at intervals.
NREPP (National Registry of Evidence-Based Programs and Practices (1997, remodelled in 2004) http://www.nrepp.samhsa.gov/	US, mental health and substance abuse	Searchable online registry of mental health and substance abuse interventions reviewed and rated by independent reviewers. NREPP rates the quality of the research supporting intervention outcomes and the quality and availability of training and implementation materials; NREPP ratings do not reflect an intervention's effectiveness.
Evidence-Based Policing Matrix (Centre for Evidence-Based Crime Policy) (c. 2009) http://gemini.gmu.edu/cebcp/Matrix.html	US, criminal justice	The Matrix is a 'research-to-practice' translation tool that presents the findings of stronger studies (i.e. experimental and quasi-experimental studies) visually to guide police agencies in developing future tactics or strategies. Evaluation criteria are based on the Maryland Scale but with modifications.
NHMRC (National Health and Medical Research Council) 'Designation of Levels of Evidence' (1999) http://www.biomedcentral.com/1471-2288/9/34	Australia, healthcare	The original hierarchy was developed in relation to interventions (clinical guidelines and health technology assessment) and ranks the body of evidence into four levels. Level I is evidence from a systematic review of RCTs and Level IV is evidence from case series. The hierarchy was revised in the mid 2000s to increase its relevance for assessing the quality of other types of studies (e.g. prognostic, aetiologic and screening studies).

ANNEX 2 STANDARDS OF EVIDENCE DEVELOPED BY THE SOCIAL RESEARCH UNIT AT DARTINGTON

A set of standards of evidence have been developed by the Social Research Unit at Dartington (see Box 7). These standards were used as a guide by the Allen Review on Early Intervention. They are also being used to underpin the UK's source of information on evidence-based programmes called Evidence2Success.

Any intervention that has the intention to improve children's health and development can be assessed against the four dimensions of the standards:

A. Evaluation quality

Many interventions have been evaluated, but the quality of evaluation varies considerably. The standards value evaluations that give a reliable indication of impact on child outcomes. (Other types of evaluation, such as consumer satisfaction and implementation quality are valued but fall outside the focus of the standards.) Interventions that meet this test typically:

- Are supported by people who have a genuine interest in finding out whether the intervention is effective;
- Have been subjected to an evaluation that compares outcomes for children receiving the intervention with children with the same needs who do not receive the intervention;
- Ideally, have been independently evaluated using a well-executed randomised controlled trial.

B. Intervention impact

The standards value interventions that can be clear about how much impact will typically be achieved on specific dimensions of children's health and development. The standards emphasise two dimensions of impact:

- A positive effect size, a standard measure of impact that provides comparable data regardless of the outcomes assessed;
- No harmful effects or negative side-effects of the intervention.

C. Intervention specificity

This is an estimation of whether the intervention might logically be expected to have an impact on children's health and development. Programmes and practices that meet this test typically are clear about:

- Who is being served;
- What impact on which aspects of children's health and development will be achieved;
- The reason – the logic behind – why the intervention will achieve the outcome.

D. System readiness

Many of the most effective interventions are not ready for the real world. Meeting this test typically involves:

- Having a clear indication of unit cost and staffing requirements;
- Explicit processes to measure the fidelity of implementation and to address common implementation problems.

Application of the standards in relation to Realising Ambition

Each application for Realising Ambition support will be screened against the standards. It is anticipated that very few interventions will meet the standards in full. Most of the interventions in the Realising Ambition portfolio will score well against at least one dimension of the standards, and will have a clear plan for improving against other dimensions.

Applicants seeking recognition from Realising Ambition need to bear in mind that their work will be regularly scrutinised against the standards during the three to five years of the Realising Ambition programme. The objective is not to meet all of the standards at the outset, but to plan for continuous improvement.

Source: http://www.dartington.org.uk/sites/default/files/Standards%20of%20Evidence.pdf

Box 7: Dartington standards of evidence criteria

A. Evaluation quality

Good enough

A1. One randomised controlled trial (RCT) or two quasi-experimental design (QED) evaluations (initial quasi-experimental evaluation and replication) with the following characteristics (see A1a-A1e):

A1a. Assignment to the intervention is at a level appropriate to the intervention.

A1b. There is use of measurement instruments that are appropriate for the intervention population of focus and desired outcomes.

A1c. Analysis is based on 'intent to treat'.

A1d. There are appropriate statistical analyses.

A1e. Analyses of baseline differences indicate equivalence between intervention and comparison.

A2. There is a minimum of one long-term follow-up (at least six months following completion of the intervention) on at least one outcome measure indicating whether results are sustained over time.

A3. There is a clear statement of the demographic characteristics of the population with whom the intervention was tested.

A4. There is documentation regarding what participants received in the intervention and counterfactual conditions.

A5. There is no evidence of significant differential attrition.

A6. Outcome measures are not dependent on the unique content of the intervention.

A7. Outcome measures reflect relevant outcomes.
Reqires evidence that one or more of the outcome measures reflects one or more relevant outcomes.

A8. Outcome measures are not rated solely by the person or people delivering the intervention.

Best

A9. There are two RCTs or one RCT and one QED evaluation (in which anaylisis and controls rule out plausible threats to internal validity).
Requires evidence that at least two RCTs or one RCT and one QED evaluation were conducted on the intervention in question and, critically, that they meet the methodological requirements spelled out in all 'good enough' evaluation quality criteria (A1 – A8).

A10. The evaluation results indicate the extent to which fidelity of implementation affects the impact of the intervention.

A11. Dose-response analysis is reported.

A12. Where possible or appropriate there is analysis of the impact on sub-groups (e.g. do the results hold up for different age groups, boys and girls, ethnic minority groups?)

A13. There is verification of the theoretical rationale underpinning the intervention, provided by mediator analysis showing that effects are taking place for the reasons expected.

B. Impact

Good enough

B1. There is a positive impact on a relevant outcome.
Requires evidence that in a majority of studies complying with the 'good enough' evaluation quality criteria set out in section A, programme group participants did better relative to the control group participants on a relative outcome, and that the difference is statistically significant.

B2. There is a positive and statistically significant effect size, with analysis done at the level of assignment (or, if not, with appropriate correction made.)

or

There is a reported sample size of 0.2 with a sample size of more than 500 individuals across all studies.

B3. There is an absence of iatrogenic effects for intervention participants. (This includes all sub-groups and important outcomes.)

Best

B4. If two or more RCTs or at least one RCT and one QED evaluation have been conducted, and they meet the methodological criteria stipulated in section A (see criterion A9), there is evidence of a positive effect (criterion B1) and an absence of iatrogenic effects (criterion B3) from a majority of the studies.

B5. There is evidence of a positive dose-response relationship that meets the methodological standard stated in A11.

C. Intervention specificity

Good enough

C1. The intended population of focus is clearly defined.

C2. Outcomes of intervention are clearly specified and meet one of the relevant outcomes.

C3. The risk and promotive factors that the intervention seeks to change are identified, using the intervention's logic model or theory explaining why the intervention may lead to better outcomes.

C4. There is documentation about what the intervention comprises.

Best

C5. There is a research base summarising the prior empirical evidence to support the casual mechanisms (risk and protective factors) that underlie the change in outcomes being sought.

D. System readiness

Good enough

D1. There are explicit processes for ensuring that the itervention gets to the right people.

D2. There are training materials and implementation procedures.

D3. There is a manual(s) detailing the intervention.

D4. There is reported information on the financial resourses required to deliver the intervention.

D5. There is reported information on the human resourses required to deliver the intervention.

D6. The intervention that was evaluated is still available.

Best

D7. The intervention is currently being widely disseminated.

D8. The intervention has been tested in 'real world' conditions.

D9. Technical support is available to help implement the intervention in new settings.

D10. Absolute financial investment is stated.

D11. There is a fidelity protocol or assessment checklist to accompany the intervention.

Note: More detailed explanations of these standards exist for those completing programme reviews: they are available from Nick Axford at the Social Research Unit, Dartington (naxford@dartington.org.uk).

Source: Annex C of Allen (2011) 'Early Intervention: The Next Steps. An Independent Report to Her Majesty's Government.' London: HM Government.

ANNEX 3 INCLUDING FIVE TYPES OF KNOWLEDGE IN SYSTEMATIC RESEARCH REVIEWS IN SOCIAL CARE (EXTRACT FROM RUTTER ET AL., 2010, PP 17–19)

36. A systematic review includes any knowledge that exists in answer to a particular question. The aim is comprehensive coverage. In practice, explicit and comprehensive electronic and manual searches are undertaken to find relevant research literature, user testimony, economic data and other sources of material to be included in the review.

37. It is important that the five types of knowledge identified in social care (Pawson, Boaz et al., 2003) are incorporated into knowledge reviews. Below is a list of these knowledge types and possible ways they can be incorporated into knowledge reviews. Most of the literature included in a research review is likely to be research evidence, and there are systematic approaches to the review of such evidence. However, such research evidence may uncover or expand on any of the following types of knowledge, depending on the focus and participants in the study.

Policy knowledge

38. Policy guidance, legislation and other policy information should be incorporated into the background section of a review report, to ensure that the appropriate context for the review topic is identified and described.

Organisational knowledge

39. Any relevant information from providers and regulatory bodies would be summarised in the background section. Where services have been evaluated, information from specific organisations may be included in the research review. This might include information on barriers and facilitators to improving the intervention or service, and other organisational information in relation to working practices or service delivery, where these have an impact on the review question. It would be likely in most cases that the practice enquiry element of the knowledge review would capture specific perspectives from organisational experience.

Practitioner knowledge

40. Practitioners may be involved either as part of the team conducting the review or as members of advisory or stakeholder groups. Additionally, practitioner knowledge might be captured in the research review through the incorporation of any relevant research or other published material. This knowledge might include information on barriers and facilitators to implementing or improving an intervention or service, and other practitioner–level information in relation to working practices that have an impact on the review question. Practice enquiries also capture practitioner knowledge and experience, and this is a key area where practitioner views are included in SCIE knowledge reviews.

User and carer knowledge

41. Service users and carers should be involved, ideally as part of the team conducting the review, or as members of advisory or stakeholder groups. Additionally, as specified in the section on searching (Paragraph 137 onward), specific attempts should be made to locate sources of user testimony in searches. Similarly, such knowledge might be captured in searches through the incorporation of any research or other published material that presents user views or experiences.

42. The purpose of collecting this data is always to ensure that user and carer views are represented so that their perspectives on access, impact and utility of the intervention or the processes being reviewed are included in the evidence base.

Research knowledge

43. Research knowledge is primarily captured in knowledge reviews through searching databases of published and unpublished research studies, and the incorporation of this in the research review component of the knowledge review.

Alliance for Useful Evidence

1 Plough Place
London EC4A 1DE

www.alliance4usefulevidence.org

February 2013