Alliance for Useful Evidence

Mapping the Standards of Evidence used in UK social policy

Ruth Puttick

May 2018

Funded by:

nesta

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

BIG LOTTERY FUND

NATIONAL LOTTERY FUNDED

Acknowledgements

We are incredibly grateful to the Standards of Evidence Working Group, convened by the Alliance for Useful Evidence, who all generously shared their experiences and insights to help shape and develop this project. Thank you to **Tom McBride** at the Early Intervention Foundation, **Leon Feinstein**, **Sue Holloway** at Project Oracle, **Tamsin Shuker** at the Big Lottery Fund, **Nick Axford** at Dartington Service Design Lab, and **Reuben Saxon** at The Social Innovation Partnership.

Our warmest thanks to the other organisations who shared their experiences of developing and using standards of evidence, including **Dr Kay Nolan** at NICE, **André Clarke** at BOND, **Sara MacLennan** at the What Works Centre for Wellbeing, **Joanne Moore** at the Aim Higher Network, **Nerys Thomas** and **Julia Morris** at the College of Policing, **Danielle Mason** and **Triin Edovald** at the Education Endowment Foundation, **Laura Carmody** at HACT, **Max Nathan** at the What Works Centre for Local Economic Growth, and **Richard Lynas** at Mentor UK.

Thanks also to Nesta colleagues, particularly **Jonathan Breckon** and **Lucy Heady**, who provided thoughtful comment and critique throughout.

As ever, all errors and omissions remain the author's own.

About The Alliance for Useful Evidence

The Alliance for Useful Evidence is a network, hosted by Nesta, that champions the smarter use of evidence in social policy and practice. We do this through advocacy, convening events, sharing ideas and resources, and supporting individuals and organisations through advice and training. We promote our work through our network of more than 3,500 individuals from across government, universities, charities, businesses, and local authorities in the UK and internationally. Anyone can join the Alliance network at no cost.

To sign up please visit: **www.alliance4usefulevidence.org/join**

We are funded by the Big Lottery Fund, the Economic and Social Research Council and Nesta.

About Nesta

Nesta is a global innovation foundation. We back new ideas to tackle the big challenges of our time.

We use our knowledge, networks, funding and skills - working in partnership with others, including governments, businesses and charities. We are a UK charity but work all over the world, supported by a financial endowment.

To find out more visit **www.nesta.org.uk**

Alliance for Useful Evidence

Funded by:

nesta

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

BIG LOTTERY FUND

NATIONAL LOTTERY FUNDED

Mapping the Standards of Evidence used in UK social policy

Ruth Puttick

May 2018

Alliance for Useful Evidence

Funded by: nesta ESRC Economic & Social Research Council BIG LOTTERY FUND NATIONAL LOTTERY FUNDED

Foreword

Not all evidence is born equal. Some evidence is stronger, more trustworthy, and more relevant, than other evidence. But figuring out what is good can be tricky. For the specialist, there are things like *Magenta* and *Green Books*, the 100-page bibles on appraisals and evaluations produced by Her Majesty's Treasury. They talk you through such things as cost-benefit analysis, randomised controlled trials, quasi-experimental impact evaluations, and the like.

Such detailed guides are fine and needed. But what is also needed are easy-to-grasp frameworks, for both specialists and lay audiences, that judge the evidence backing your policy, programme or practice. This report sets out these frameworks – currently 18 in the UK, and rising - and helps you navigate between their different approaches.

Usually the frameworks have some sort of one to five scale, and usually the focus is on evidence of impact. But what our mapping reveals is that there are important variations around this stereotype. A good example of this is that some standards, like those produced by Nesta or Project Oracle, look at the strength of the designs of single evaluations or interventions, while others look at whole bodies of evidence, such as the meta-analysis and systematic reviews used by the Education Endowment Foundation. BOND's international development evidence principles go even further. It has added new criteria such as 'voice and inclusion'; taking into account the perspectives of the marginalised or people living in poverty.[1] There is no carbon copy for these standards. All have been tweaked or rebuilt to meet the needs of their sector, and that is a good thing.

However, this diversity can create confusion. It is alarming to hear that some interventions have been rated a decent two on one scale, but a poor zero by another. That sends a garbled message to our sector. We need to standardise the standards, set up an independent accreditation system, like those curated by the Kitemarking body the British Standards Institute, or the ISO. Another option is to create something like the US Results First Clearing House Database,[2] a one-stop online shop that compares the ratings of interventions across eight national what works-type centres. We will certainly endeavour to encourage knowledge sharing amongst providers of standards, including working internationally with OECD and others, and create a toolkit or good practice guide. We also hope to do more to understand the 'demand side' of users – such as grantees and funders, frontline professionals, commissioners of services, and others. But for now, we should reflect on how 18 standards represent our still nascent sector. We should be proud of how they have grown, learnt from each other, adapted, and flourished.

Jonathan Breckon, Director, Alliance for Useful Evidence, Nesta.

1

Introduction

We all know that evidence is vital if good decisions are to be made. Yet evidence can be lacking, there can be confusion about what evidence looks like, or there can be different types of evidence making it unclear how confident we can be in what the data is saying.

This is where standards of evidence come in.

Over the past decade, there has been a growing interest in standards of evidence, and other frameworks that help us understand what is working, and what isn't, by grading effectiveness or impact against a scale or level. Typically, the lower levels indicate that there is some evidence, and as the scale or level is ascended, more evidence is available to increase confidence in deciding whether the intervention or practice is working.

This paper maps the 18 standards of evidence currently used in UK social policy. It is intended to help all of us - innovators, commissioners, providers, policymakers, services users, and practitioners – navigate the field and understand where standards of evidence exist, what they do, similarities between them, and where they differ.

In summary we have found:

- There has been a rapid proliferation of standards of evidence and other evidence frameworks since 2000. This is a very positive development and reflects the increasing sophistication of how evidence is generated and used in social policy.
- There are common principles underpinning them, particularly the shared goal of improving decision-making, but they often ask different questions, are engaging different audiences, generate different content, and have varying uses. This variance reflects the host organisation's goals, which can be to inform its funding decisions, to make recommendations to the wider field, or to provide a resource for providers to help them evaluate.
- It may be expected that all evidence frameworks assess whether an intervention is working, but this is not always the case, with some frameworks assessing the quality of evidence, not the success of the intervention itself.
- The differences between the standards of evidence are often for practical reasons and reflect the host organisation's goals. However, there is a need to consider more philosophical and theoretical tensions about what constitutes good evidence. We identified examples of different organisations reaching different conclusions about the same intervention; one thought it worked well, and the other was less confident. This is a problem: Who is right? Does the intervention work, or not? As the field develops, it is crucial that confusion and disagreement is minimised.
- One suggested response to minimise confusion is to develop a single set of standards of evidence. Although this sounds inherently sensible, our research has identified several major challenges which would need to be overcome to achieve this.
- We propose that the creation of a single set of standards of evidence is considered in greater depth through engagement with both those using standards of evidence, and those being assessed against them. This engagement would also help share learning and insights to ensure that standards of evidence are effectively achieving their goals.

2

Current landscape of standards of evidence in social policy

We have analysed the 18 standards of evidence currently used in UK social policy.[3] We have only looked at impact standards and those relevant to the social sector, we have also only included those currently used in the UK.

Our analysis focuses on **18 frameworks** used by 16 UK organisations for **judging evidence used in UK domestic social policy which are relevant to government, charities, and public service providers**. These are:

Organisation	Framework
Big Lottery Fund's Realising Ambition programme	The Confidence Review
Bond	Evidence Principles
Centre for Analysis of Youth Transitions (CAYT)	Standards of Evidence
Dartington Service Design Lab (formerly known as Dartington Social Research Unit)	Standards of Evidence
Early Intervention Foundation	Evidence Standards
Education Endowment Foundation	Teaching and Learning Toolkit
Education Endowment Foundation	Evaluation Padlock Rating
HACT	Standards of Evidence
National Institute for Health and Care Excellence (NICE)	GRADE (Grading of Recommendations Assessment, Development and Evaluation)
National Institute of Health Research's Health Services and Delivery Research (NIHR HS&DR) Programme	Realist and Meta-Narrative Evidence Syntheses: Evolving Standards (RAMESES)
National Academy for Parenting Practitioners	Parenting Programme Evaluation Tool (PPET)
Nesta	Standards of Evidence
Office for Fair Access (OFFA)	Standards of Evidence[4]
Project Oracle	Standards of Evidence
What Work Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)
What Works Centre for Crime Reduction	EMMIE Framework
What Works Centre for Wellbeing	GRADE (Grading of Recommendations Assessment, Development and Evaluation)
What Works Centre for Wellbeing	CERQual (Confidence in the Evidence from Reviews of Qualitative research)

The standards of evidence we have analysed **are in a range of charitable and government organisations**. It is worth noting that three of the UK What Works network and its affiliates – the Centre for Ageing Better, the Wales Centre for Public Policy, and What Works Scotland - are omitted from the list. This is because they currently do not use a standards of evidence framework. We have also excluded the standards of evidence used by Pearson Plc[5] as these are the same as Nesta's Standards of Evidence.

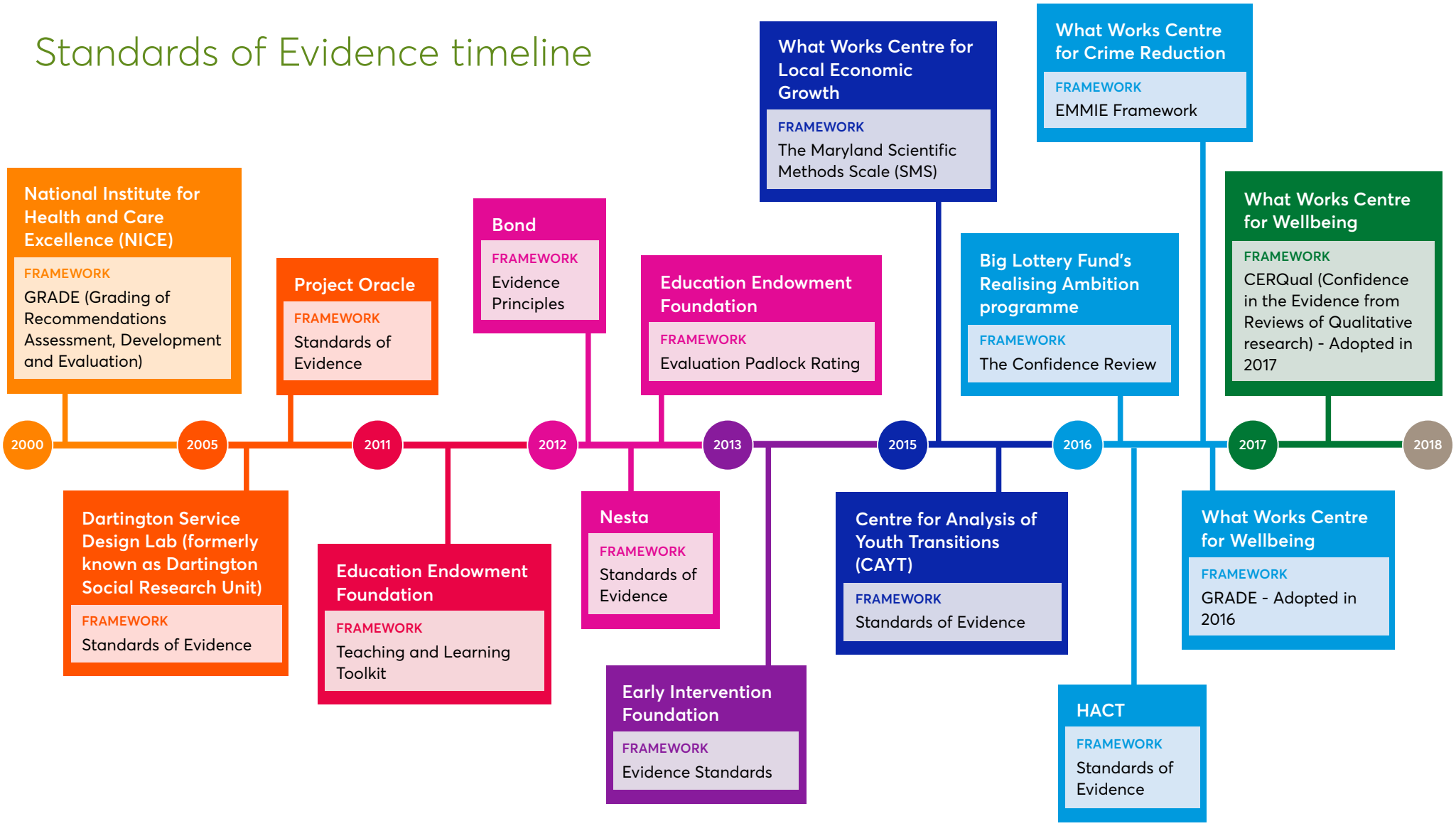
Standards of Evidence have a long history, and there has been a rapid proliferation of standards of evidence since 2000. The oldest that we have examined is the GRADE system used by NICE and the What Works Centre for Wellbeing, followed by the Standards of Evidence for London, commissioned by the Greater London Authority for Project Oracle and delivered by the Dartington Social Research Unit in collaboration with international experts such as Delbert Elliot at the University of Colorado, and Steve Aos at the Washington State Institute for Public Policy (WSIPP).[6, 7] The most recent to emerge was the What Works Centre for Wellbeing's CERQual framework in 2017, and in 2016, HACT's Standards of Evidence, Realising Ambition's Confidence Review, and the What Works Centre for Crime Reduction's EMMIE framework.

The standards of evidence **span numerous social policy areas**. Some, such as those used by Nesta and Realising Ambition, are designed to accommodate all policy areas. Three of the frameworks – developed by Project Oracle, the Early Intervention Foundation, and Dartington Service Design Lab – focus upon children and young people. Bond focuses on international development, with the other frameworks focusing on health, local economic growth, housing, education, wellbeing, social care, and crime. It is worth noting, that although most have a specific policy focus, all of them could be used in other policy areas.

The evolution of standards of evidence is a very positive development and reflects the increasing sophistication of how evidence is generated and used in social policy. And it also reflects the determination of those organisations involved. Often the creation of evidence and the increased scrutiny which accompanies it, is met with controversy and resistance.[8]

Although there has been a rapid uptake of standards of evidence, they are not without their critics. We do not have space to go into these here, but it is worth noting the ongoing debates about whether standards of evidence focus too heavily on quantitative methods, and in particular, randomised controlled trials (RCTs), or whether lists of interventions to tackle complex social challenges is too simplistic a response. These discussions are valuable in helping sharpen the use of evaluation across social policy, but they should not distract from the value that standards of evidence are providing in the quest of finding what is working.

Standards of Evidence timeline



Office for Fair Access Standards of Evidence[9]

In 2017, the Office for Fair Access (OFFA), the independent regulator of fair access to higher education in England developed standards of evidence. To create these, OFFA adapted Nesta's Standards of Evidence and those developed by The Social Innovation Partnership (TSIP). The OFFA Standards of Evidence are on a 1 to 3 scale as follows:

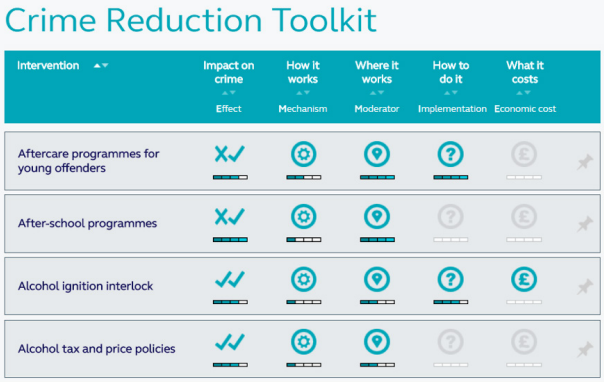
Level 1
The Higher Education Institution (HEI) can provide a narrative to motivate its selection of outreach activities in the context of a coherent outreach strategy.

Level 2
In addition to a narrative account, the HEI has collected data on impact and can report evidence that those receiving an intervention treatment have better outcomes, though this does not establish any direct causal effect.

Level 3
The HEI has implemented an evaluation methodology which provides evidence of a causal effect of an intervention.

The Crime Reduction Toolkit

The What Works Centre for Crime Reduction has developed the Crime Reduction Toolkit. The EMMIE Framework - discussed in this paper – is part of this Toolkit. The Toolkit ranks interventions that have undergone at least one systematic review on the impact on crime, how it works, whether it works, how to do it, and how much it costs. A snapshot of the Toolkit is shown below.



Source: whatworks.college.police.uk/toolkit/Pages/Toolkit.aspx

Realist and Meta-narrative Evidence Syntheses (RAMESES)

In 2011, the National Institute of Health Research Health Services & Delivery Research Programme funded the RAMESES project to develop quality and reporting standards and training materials for realist reviews. The aims were to increase the quality of realist reviews, its reporting and to help build realist review research capacity. In 2015 it funded the RAMESES II project to do the same for realist evaluations.[10]

RAMESES was the first attempt to provide standards for reporting realist synthesis. Realist synthesis is seen as an alternative to the systematic review method[11] and draws on theory-driven, qualitative, and mixed methods approaches. These approaches can help expand the knowledge base by explaining the success, failure, or mixed fortunes of complex interventions.[12]

What they do

The standards of evidence vary widely in their aims, scope, and objectives. The table below summarises what they do. It is ordered chronologically so that the evolution and progression of the landscape of standards of evidence can be observed.

Organisation	Framework	Date the framework was created	Aim	Summary
National Institute for Health and Care Excellence (NICE)	GRADE (Grading of Recommendations Assessment, Development and Evaluation)	2000	*"Rating the quality of evidence across outcomes in systematic reviews and guidelines… GRADE rates the quality of evidence for a particular outcome across studies and does not rate the quality of individual studies."* GRADE is used to develop NICE clinical guidelines.[13]	The certainty of evidence is classified as high, moderate, low, or very low.[14] • High: we are very confident that the true effect lies close to that of the estimate of the effect. • Moderate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. • Low: our confidence in the effect estimate is limited. The true effect may be substantially different from the estimate of the effect. • Very low: we have very little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of effects.
Dartington Service Design Lab (formerly known as Dartington Social Research Unit)	Standards of Evidence	2005	*"To determine which programmes work in improving children's outcomes".*[15]	It evaluates specific interventions across four dimensions: 1) Evaluation Quality; 2) Impact; 3) Intervention Specificity; 4) System Readiness
Project Oracle	Standards of Evidence	2005	*"Validation against the standards reflects how an organisation currently gathers and uses evidence on the interventions they deliver, and identifies how they might improve over time."* [16]	A five-step process, as the scale is ascended confidence increases that the change is happening because of the intervention.

Organisation	Framework	Date the framework was created	Aim	Summary
Education Endowment Foundation	Teaching and Learning Toolkit	2011	*"Provides guidance for teachers and schools on how to use their resources to improve the attainment of all pupils, including the most disadvantaged".*[17]	The toolkit combines impact estimates from high quality research studies into a single average for a topic area, such as 'one-to-one tuition', or 'classroom assistants'.
Nesta	Standards of Evidence	2012	*"An approach used to measure the impact of a range of our practical innovation programmes and investments."*[18]	A five-level framework. As products and services move up the five levels, so does certainty that they will have a positive impact on the intended outcome.
Bond	Evidence Principles	2012	*"…a tool for assessing the quality of evidence collected and used by NGOs to measure, learn from and demonstrate their impact. They can also be used to review and quality-assure existing evidence (e.g. an evaluation report), and as a reference point when thinking about how evidence will be generated (e.g. to set an evaluation terms of reference)".*[19]	The five principles are: voice and inclusion, appropriateness, triangulation, contribution, and transparency. For each principle the checklist has four questions that can be used to test the quality of a piece of evidence. Each question is scored on a 1 – 4 scale, allowing the user to test the quality of a piece of evidence on a scale: 1) weak evidence, 2) minimum standard, 3) good practice, 4) gold standard. An overall score and colour (red, amber, light green or dark green) is then assigned to each principle to provide a holistic picture of the robustness of a piece of evidence.
Education Endowment Foundation	Evaluation Padlock Rating	2012	*"To judge the security of findings from EEF evaluations […] and communicate the likelihood of finding the same result in a similar context again in a well conducted evaluation".*[20]	The padlock system aims to differentiate between EEF evaluations, most of which are randomised controlled trials (RCTs) and gives 5 padlocks to the best kind of evidence expected from a single study, to 0 padlocks, where the study adds little or nothing to the evidence base.
Early Intervention Foundation	Evidence Standards	2013	*"To inform judgements about the extent to which a programme has been found effective in at least one rigorously conducted evaluation study".*[21]	The rating system distinguishes five levels of strength of evidence of impact. This is not a rating of the scale of impact but of the degree to which a programme has been shown to have a positive, causal impact on specific child outcomes.
Centre for Analysis of Youth Transitions (CAYT)	Standards of Evidence	2015	*"To provide educators and prevention practitioners with evidence of what has proved – or is promising – to be of good practice, and to highlight those programmes showing high effectiveness and rigorous evidence."*[22]	An overall performance score is given with two separate scores combined for a) Standard of Evidence and b) Programme Impact.[23]

Organisation	Framework	Date the framework was created	Aim	Summary
What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	2015	*"To produce its reviews and toolkits, the WWC screens an initial long-list of evaluations on relevance, geography, language and methods, keeping impact evaluations from the UK and other OECD countries, with no time restrictions on when the evaluation was done. They then screen the remaining evaluations on the robustness of their research methods, keeping only the more robust impact evaluations. They use the Maryland Scientific Methods Scale (SMS) to do this."* [24]	The SMS is a five-point scale ranging from 1, for evaluations based on simple cross-sectional correlations, to 5 for randomised control trials. WWC Reviews are designed to be overviews and use evidence at SMS3 and above. Toolkits are designed as more practical/implementation guidance, so use evidence at SMS2 and above.[25] (Note: These levels are based on, but not identical to, the original Maryland SMS. The levels here are generally a little stricter than the original scale to help to clearly separate levels 3, 4 and 5 which form the basis for the evidence reviews.)[26]
Big Lottery Fund's Realising Ambition programme	The Confidence Review	2016	*"To help service delivery organisations identify areas of strength and areas for improvement. It focuses on both the service that is delivered and on the organisation itself."*[27]	It examines five dimensions it views as essential for replication – service design, service delivery, ability to monitor impact, ability to determine benefit and the prospects for sustainability. Each dimension contains four indicators.
HACT	Standards of Evidence	2016	*"To provide an agreed, repeatable way of doing something – in this case, a consistent way of producing evidence of the effectiveness of the interventions".*[28]	A seven-step process to help providers generate evidence on their intervention.
What Works Centre for Crime Reduction	EMMIE Framework	2016	*"…summarises the best available research evidence on what works to reduce crime. [and] to present evidence from systematic reviews of research on crime reduction interventions in a format that helps users to access and understand it quickly."*[29]	**EMMIE** rates each intervention against the following five dimensions: **E**ffect - Impact on crime: Whether the evidence suggests the intervention led to an increase, decrease or had no impact on crime. **M**echanism - How it works: What is it about the intervention that could explain its effect? **M**oderators - Where it works: In what circumstances and contexts is the intervention likely to work/not work? **I**mplementation - How to do it: What conditions should be considered when implementing an intervention locally? **E**conomic cost - How much it costs: What direct or indirect costs are associated with the intervention and is there evidence of cost benefits?

Organisation	Framework	Date the framework was created	Aim	Summary
What Works Centre for Wellbeing	GRADE	Adopted in 2016	*"…used for grading the quality of evidence from quantitative systematic reviews."*[30]	The evidence is graded as: • High quality: Further research is very unlikely to change our confidence in the estimate of effect. • Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. • Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. • Very low quality: any estimate of effect is very uncertain.
What Works Centre for Well being	CERQual (Confidence in the Evidence from Reviews of Qualitative research)	Adopted in 2017	*"…to rate the quality of evidence for findings in qualitative evidence reviews".*[31]	CERQual grades the evidence of qualitative data in systematic reviews. It is conceptually like GRADE but is tailored to suit qualitative evidence. It bases its qualitative reviews on four components: • Methodological limitations of the qualitative studies contributing to a review finding. • Relevance to the review question of the studies contributing to a review finding. • Coherence of the review finding. • Adequacy of data supporting a review finding.

Role and remit

Although all the evidence frameworks have the ultimate goals of improving understanding of what is working, and what isn't, they don't all have the same focus. The frameworks assess different things, with differing focus and aims, reflecting how the host organisation is using the framework. The focus of evidence frameworks can be on:

- **Specific interventions**. Including well-defined programmes, through to thematic topics and areas, such as 'homework'.
- **Bodies of evidence**. Such as an evidence review or meta-analyses on a topic.
- **Organisation's readiness**. Such as an organisation's ability to evaluate or replicate.
- **Quality of an individual evaluation**. Such as how robust the study is and how confident we can be in its findings.

Although there can be overlap between the categories, **the principal usage** of each of the frameworks can be divided into three groups:

1. To inform the strategy and funding decisions of the organisation using the standards of evidence.
2. To make recommendations to the wider field – including policymakers, academics, commissioners, funders, and others - about what works and what doesn't.
3. As a resource for providers of interventions and projects to help deepen their understanding of their organisation and help them to evaluate.

In the first category, when evidence frameworks help inform the strategy and funding decision of the host organisation, is Nesta's Standards of Evidence, where the standards of evidence are primarily used to decide whether a grant is awarded, or an investment is made, and is then continued to be used to help assess the progress of the intervention, and to plan future evaluations. Once an evaluation is completed, Nesta uses the standards when deciding to make future grants or investments and to assess how well the evidence base has improved across a cohort of investees or grantees. NICE also fits into this category, with GRADE helping signal to the wider field what is working, but its primary focus being to develop clinical guidelines to inform the NHS.

The second category uses standards of evidence to make recommendations to the wider field. In this group is the What Works Centre for Crime Reduction's EMMIE framework; the What Works Centre for Economic Growth's Maryland Scientific Methods Scale (SMS); the Early Intervention Foundation's Evidence Standards; Dartington Service Design Lab's Standards of Evidence; the Education Endowment Foundation's (EEF) Teaching and Learning Toolkit; the EEF's Evaluation Padlock Rating System. Although all these evidence frameworks vary greatly, and some focus on individual interventions and others focus on bodies of evidence, what they all have in common is that they are used to signal to the wider field whether a specific intervention is working or how much confidence can be placed on the findings.

The third category are standards of evidence being developed as a resource for those developing interventions and projects to help them deepen their understanding of their organisation and to help them evaluate. These resources are accessible online and are free to use. There is no attempt to record the evidence generated, or to share it with the wider

field. Within this category are Bond's Evidence Principles, HACT's standards of evidence, and Realising Ambition's Confidence Framework. Project Oracle's Standards of Evidence also fit within this category, however, it takes a very different approach. Rather than providing guidance on how to conduct an evaluation, Project Oracle supports organisations to understand and assess the evidence behind their interventions. This review, or 'validation' as Project Oracle term it, doesn't endorse the organisation or their work, but instead confirms that plans and evidence are in place to show that the evidence submitted is of a sufficiently high quality.[32] Project Oracle then acknowledges where organisations are positioned on their Standards of Evidence within its Evidence Hub, an open access website.[33]

Type of evaluation and methods

A similarity across each of the standards of evidence is that they are interested in impact, however, as the section above demonstrated, they vary in what they are assessing the impact of. For instance, some look at specific interventions, specific evaluations, or bodies of evidence.

All the evidence frameworks include impact evaluations, and only three also include process evaluation. It may not be surprising that only a few incorporate process evaluations, as many are designed to only look at evidence of impact. As an example, the What Works Centre for Local Economic Growth only considers impact evaluations and uses this specific type of evidence to understand the causal effects of policy interventions and to establish their cost-effectiveness.[34]

Quantitative methods are a feature of all the evidence frameworks, and 12 out of 15 explicitly consider the effect size of the impact. The use of qualitative method is less prevalent, with only a third explicitly mentioning that it uses or considers qualitative methods. An exception worth highlighting is the What Work Centres for Wellbeing's CERQual framework, which focuses exclusively on qualitative studies in an evidence review and it sits alongside the Centre's use of GRADE to assess quantitative studies.

For more details on the types of evaluation and methods used, see Table 1 in the Appendix.

Ranking and rating

A similarity across nearly all the standards of evidence is the use of a scale. However, despite interventions, evaluations, or meta-analyses being assessed on a tiered system there are differences between them, and some frameworks have multiple dimensions within their tiered system. (Table 3 in the Appendix provides more detail.)

In summary, 13 of the frameworks use an ascending scale. Nesta and Project Oracle use five levels, whilst the What Works Centre for Crime Reduction award a score out of five to signal how well an intervention scores on five components: effect, mechanism, moderators, implementation, and cost. Dartington Service Design Lab's Standards of Evidence does not have a tiered system, but instead analyses four dimensions of an intervention's evidence and rates each as 'good enough' or 'best'. The What Works Centre for Wellbeing's CERQual and GRADE frameworks rate evidence as high quality, moderate quality, low quality, or very low quality. The two which do not use a scale are Realising Ambition's Confidence Review and HACT's Standards of Evidence.

Spread across the standards of evidence

Where a graded scale is used in a standards of evidence framework, it may be assumed that there are interventions, evaluations, and meta-analyses, at each level. Yet what we can see when we examine the spread is that there are fewer at the higher levels. This is the case with the What Works Centre for Wellbeing's use of GRADE and CERQual, where most of the reviews are rated as moderate quality or low quality, and only a few have reached the high quality threshold.[35] For example, for the reviews undertaken by the What Works Centre for Economic Growth, out of the 15,000 studies examined, only around 2.5 per cent are at SMS Level 3 or above.[36, 37] A similar pattern is seen with the Early Intervention Foundation's (EIF) Evidence Standards. On the EIF Evidence Standards, six are ranked at Level 4, 38 at Level 3, and 37 at Level 2.[38] Fewer programmes reaching the higher levels is perhaps not surprising, when the early intervention field is still building its evidence and programmes are still working towards reaching the higher levels.

In the case of Project Oracle, interestingly nothing has yet been graded as reaching Level 4 or 5, and only 2 per cent of projects (six projects) have reached Level 3. This means that 98 per cent of projects are at Levels 1 and 2.[39] This serves as a reminder of how nascent the field is and how robust evidence of impact can be lacking, it also highlights the support that charities need in undertaking evaluation and building the evidence behind their work.

In addition, some of the standards of evidence do not have data available about the spread of interventions or evaluations across the different levels. This may be because they don't track or record it. Bond, for example, produced its Evidence Principles for use by its wider network. This means that they produce the tools, but then don't track who is using them, or how different interventions are rated on the framework.

Assessment of whether it works, or not

One thing that may be expected that all evidence frameworks have in common is that they assess whether an intervention or project is working or not. Interestingly, this is not the case across the board. The EEF's Evaluation Padlock Rating, for example, does not consider whether the intervention has worked. In other words, they are assessing the quality and robustness of the evaluation or body of evidence, not the success of the intervention.

In addition, more dimensions than simply 'does it work' should be taken into consideration. It has been argued before that the use of 'what works' is too simplistic, and instead, we should be considering if it works, for whom, under what circumstances, how, why, and at what cost.[40]

Yet only five of the frameworks answer all these questions. This group includes Dartington Service Design Lab's Standards of Evidence, EIF's Standards of Evidence, NICE's GRADE framework, What Work Crime's EMMIE, and EEF's Teaching and Learning Toolkit.

Table 2 in the Appendix shows how the other frameworks answer these questions.

As well as looking at what is working, there is also a need to consider what is not working. Publishing negative findings is not common across the standards. The one exception is the Early Intervention Foundation's (EIF) standards of evidence, which has a category called 'NE' which denotes programmes found to not be effective in at least one rigorously conducted study. The EIF make clear that this does not mean that the programme will never work, but instead that it needs adapting and improving, based upon the evaluation findings.

Validation and quality assurance of evidence

Standards of evidence can help decision-makers understand whether something is working, the quality of the evidence, such as the sample size and the data collection method, or quality of the institution supplying the evidence. This guidance can be provided by the organisation developing the intervention as self-reporting, or it can be undertaken by a third party that is not involved in the intervention.

Looking across the different standards of evidence, a variety of evidence sources is considered. Some, like the Education Endowment Foundation's Teaching and Learning Toolkit and the What Works Centre for Crime Reduction's EMMIE Framework, only consider academic research, which as part of its generation, has undergone a peer review process.

Project Oracle's Standards of Evidence and the Education Endowment Foundation's Evaluation Padlock Rating system approach peer review slightly differently. Here the findings of the evaluation are assessed and validated by an external peer review panel, which they appoint. The Education Endowment Foundation uses external peer review for all interventions, whereas Project Oracle uses this for interventions at Level 3 and above.

Peer review, or another form of validation, is not a core component of the other standards of evidence frameworks. As an example, Nesta's standards of evidence considers provider-generated research, and then internally assesses it, to ensure its quality and rigour.

In comparison, Realising Ambition's Confidence Review and HACT's standards of evidence rely mainly – if not wholly – on the evidence generated by the provider developing and delivering the intervention. There is no assessment of the quality of the claims made, or how trustworthy the data is. The Confidence Review goes as far as to say that there are 'no right or wrong answers'.[41]

Follow-up studies

Many of the standards of evidence provide a 'snap shot' of impact. This group includes HACT, Project Oracle, and Realising Ambition.[42] Here the evaluation is considered in one assessment, and there is no requirement for further or follow-up studies.

In contrast, to get a top rating with Dartington Service Design Lab's Standards of Evidence, the Early Intervention Foundation, and the Education Endowment Foundation, all require at least one follow-up study to assess how results are sustained over time.

Nesta is slightly different in that an intervention is ranked on the scale as a snap shot in time, but then the standards of evidence are used to help plan future evaluation design, and so evaluation is not a one-off exercise.

Linked to the section above about validation, it is worth noting that there is no clarity on how long validation lasts, and if there is a need for organisations or projects to develop and build more evidence. For example, a project graded at a certain level by Project Oracle, could potentially stay at that level indefinitely.

Cost of the intervention

The price of the intervention is a crucial factor in deciding whether an intervention can be adopted and used elsewhere.

Cost is an explicit component of the assessment made by the Early Intervention Foundation, Education Endowment Foundation, Nesta and the What Works Centre for Crime Reduction. Yet, the cost of the intervention is not an explicit consideration for all the standards of evidence. For example, Project Oracle and HACT's standards of evidence and Realising Ambitions' Confidence Framework do not take it into consideration at all.

Transferability and intervention readiness

We have analysed each of the standards of evidence to assess whether they look at intervention readiness, and its potential to be replicated, used elsewhere, and scaled. Dartington Service Design Lab's Standards of Evidence has one of the most comprehensive approaches and examines 'system readiness'. To meet this test, the intervention needs to have a clear indication of unit cost, staffing requirements, and explicit processes to measure the fidelity of implementation and to help address common implementation problems.

3

Reflections on the analysis

This report has mapped the standards of evidence used across social policy.

We have shown that the standards of evidence vary along two main lines. Firstly, the focus of the framework and its unit of analysis can be an intervention, an organisation, an individual evaluation, or a body of knowledge. This differing focus links to the second main difference: purpose. The principal purpose can be divided into three groups: to inform the host organisation's strategy and funding decisions; to make recommendations to the wider field about what works and what doesn't, and as a resource for providers to deepen their understanding of their organisation and to help them to evaluate. The focus and purpose then inform the questions the framework is trying to address, its unit of analysis, the body of knowledge it wants to explore, and the audience it's trying to advise. This difference is understandable.

However, there is scope for learning and knowledge sharing across the landscape of standards of evidence. This includes consideration of the following:

- **Validation of findings**. There is no consistent approach to how findings are validated. Some of the frameworks consider academic research which has undergone peer review. Others consider both academic and grey literature, and some consider data wholly generated by the developer of the intervention, without any checking of this.
- **Cost**. Not all frameworks consider cost when making an assessment. How much an intervention or new way of working is going to cost is often a crucial element of the decision-making process.
- **Implementation and transferability**. Only a few of the frameworks consider whether an intervention is ready to be implemented, transferred, and scaled.
- **Avoiding confusion**. A more fundamental challenge which needs discussion is the potential for the same intervention to be graded differently by different frameworks. The textbox provides an example. As the use of standards of evidence is still relatively small compared to the size of the sector, this issue is relatively minor, but as the field evolves, it is important that there is clarity on why different decisions are being reached. If left unresolved, there is a risk that confusion can undermine the collective efforts underway to improve decision-making.

Case study of confusion

One provider we spoke to said that their intervention was Level 2. The same provider with the same intervention was then assessed by a different organisation as part of a different grant programme. The second organisation judged their evaluation differently, viewing it as of poorer quality, and gave it a Level 0, and were bemused that the first assessment had been so favourable. The provider developing the intervention was then left with two scores, and two very different assessments and was left confused by the whole experience.

Creation of a single set of standards of evidence

It has been proposed that one way of overcoming the challenges associated with confusion and disagreement would be to generate a single set of standards, which are understandable by all involved, and recognised as high quality, trustworthy and robust.

This sounds inherently sensible. However, from our research, we have identified three potential obstacles which would need to be overcome:

- Ensuring that any new standards of evidence are appropriate across multiple contexts and to suit different purposes, such as by taking into consideration when interventions, evaluations, organisations, or bodies of evidence, are being assessed and ranked.
- Ensuring that organisations which have developed their standards of evidence are willing and able to share their technical knowledge and experience with others. For some organisations who have spent time and money in developing their standards of evidence, there may be an understandable reluctance to then give this away for free.
- Ensuring organisations are willing to adopt and recognise the new standards. There are organisational incentives to develop your own brand and own IP, and we are aware that some organisations are potentially reluctant to change their practice and adopt standards of evidence developed elsewhere, by different organisations.

These obstacles could be viewed as 'supply side' issues. They focus upon what the organisations using standards of evidence would need to do for a single standard of evidence to be developed, recognised, and used.

In addition, there is also a need to consider the demand side – what do people, commissioners, practitioners, providers, and others require from standards of evidence? Do people know how to use them? Do standards cover all elements that different audiences require, such as clear guidance on what is working or how to implement new interventions?

Next steps and recommendations

Now that the landscape of standards of evidence has been mapped, the question is: what do we do now?

Our research has raised several issues and questions which warrant further exploration in the longer term. Our next steps are:

1. **Facilitated knowledge sharing about using standards of evidence**

   The Alliance for Useful Evidence will organise workshops and other discussion fora to help share knowledge and learning about standards of evidence to help those using them to improve their practice. This will also explore why different conclusions are reached about the same interventions, and what can be done to minimise this confusion.

2. **Analyse the 'demand side'**

   The Alliance for Useful Evidence will engage the wider field in a conversation about how to make standards of evidence as useable and useful as possible. This will involve understanding what different audiences – be that commissioners, policymakers, charities, researchers, and others – want and need from standards of evidence, where the current landscape is meeting these requirements and where additional work is needed.

3. **Creation of a best practice guide**

   From this engagement, a detailed practice guide will be developed. The guide will help set out, in depth, how different audiences can navigate and use standards of evidence, based upon what they want to achieve, and who they want to influence. This will incorporate – if deemed appropriate – the creation of a single set of standards of evidence.

4. **Forge partnerships internationally**

   This report has focused on standards of evidence used in UK social policy, but there are frameworks being used in different areas of the world. We are keen to explore and learn from these, and to work with international partners, such as the OECD, to ensure we widely share the best practice.

Get involved

If you would like to share your experiences of using standards of evidence, or would like to know more about this work, please contact us at **Alliance.4usefulevidence@nesta.org.uk**

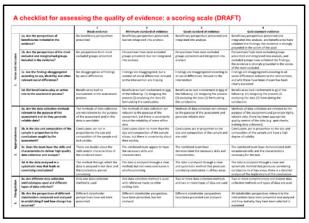
Appendix 1: A note on the data

Data for this report has been compiled from publicly available sources, supplemented with papers that have been shared with us directly from organisations about their standards of evidence. We have also discussed the findings with the organisations featured in this report. In addition, we have talked to many of the organisations cited here and other experts in the field.

All references are listed in the Endnotes. Unless otherwise indicated, the data tables and other information on the standards of evidence contained within this report are derived from organisational websites and reports. The sources are:

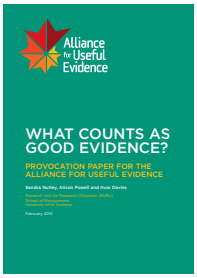
Bond

www.bond.org.uk/data/files/Effectiveness_Programme/120828Full_Bond_checklist_and_guide.pdf



Dartington Service Design Lab

www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf



Early Intervention Foundation

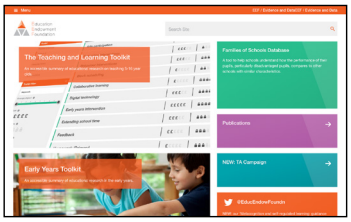
guidebook.eif.org.uk/eif-evidence-standards

Education Endowment Foundation

educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit

educationendowmentfoundation.org.uk/uploads/pdf/Teaching_and_Learning_Toolkit_(July_12).pdf

educationendowmentfoundation.org.uk/uploads/pdf/Classifying_the_security_of_EEF_findings_FINAL.pdf

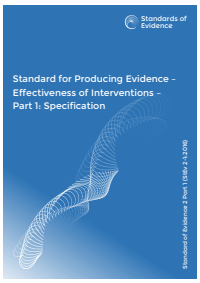


Hact

www.hact.org.uk/sites/default/files/StEv2-1-2016%20Effectiveness-Specification.pdf

www.hact.org.uk/standards-evidence-housing

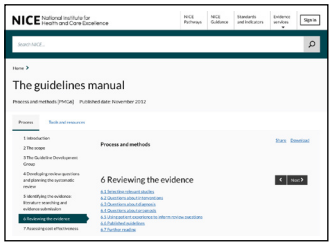
www.hact.org.uk/sites/default/files/StEv2-2-2016%20Effectiveness-Explanation.pdf



National Institute for Health and Care Excellence (NICE)

www.nice.org.uk/process/pmg6/chapter/reviewing-the-evidence

www.jclinepi.com/article/S0895-4356(10)00332-X/fulltext



Nesta

www.nesta.org.uk/report/nesta-standards-of-evidence

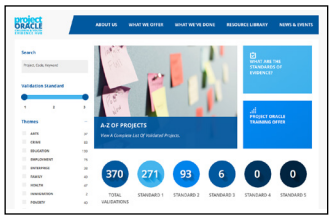
www.nesta.org.uk/centre-social-action-innovation-fund-evaluations/nesta-standards-evidence

Project Oracle

www.nesta.org.uk/sites/default/files/using_evidence_to_improve_social_policy_and_practice.pdf

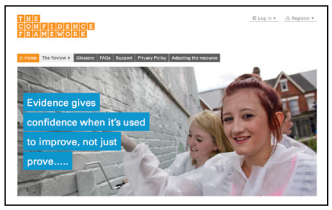
project-oracle.com/what-we-do

project-oracle.com/projects/standards-of-evidence



Realising Ambition

www.theconfidenceframework.org.uk
www.theconfidenceframework.org.uk/faqs



What Works Centre for Crime Reduction

whatworks.college.police.uk/toolkit/About-the-Crime-Reduction-Toolkit/Pages/About.aspx

whatworks.college.police.uk/toolkit/Pages/Toolkit.aspx

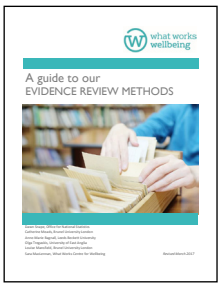
whatworks.college.police.uk/toolkit/About-the-Crime-Reduction-Toolkit/Pages/About.aspx

What Works Centre for Economic Growth

www.whatworksgrowth.org/resources/the-scientific-maryland-scale

What Works Centre for Wellbeing

www.whatworkswellbeing.org/product/a-guide-to-our-evidence-review-methods



Appendix 2: Data tables

Table 1: Type of Evaluation and Methods

Organisation	Framework	Impact evaluation	Process evaluation	Quantitative methods	Recognition of effect size	Qualitative methods
Big Lottery Fund's Realising Ambition programme	The Confidence Review	Yes	Yes	Yes	No	Yes
Bond	Evidence Principles	Yes	Yes	Yes	Yes	Yes
Centre for Analysis of Youth Transitions	Standards of Evidence	Yes	No	Yes	Yes	Unclear if considered or used
Dartington Service Design Lab	Standards of Evidence	Yes	No	Yes. At least one well conducted RCT	Yes	Unclear if considered used
Early Intervention Foundation	Evidence Standards	Yes	No	Yes	Yes	No
Education Endowment Foundation	Evaluation Padlock Rating	Yes	No	Yes	Yes	No
Education Endowment Foundation	Teaching and Learning Toolkit	Yes	No. (Each Toolkit impact estimate is based on impact evaluation but the full Toolkit entry for each topic may draw on process evaluation as well.)	Yes. Meta-analyses to produce a single average.	Yes	No. (However, the full Toolkit entry for each topic may draw on both qualitative methods as well.)
HACT	Standards of Evidence	Yes	Yes	Yes	No	Yes

Organisation	Framework	Impact evaluation	Process evaluation	Quantitative methods	Recognition of effect size	Qualitative methods
Nesta	Standards of Evidence	Yes	No	Yes	Yes (from Level 3)	Yes
Project Oracle	Standards of Evidence	Yes	No	Yes	Yes	Yes
What Work Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	Yes	No (although it is considered as background material, it is not formally reviewed).	Yes	Yes	No
What Works Centre for Crime Reduction	EMMIE Framework	Yes	Yes (?)	Yes	Yes	Unclear if considered used
National Institute for Health and Care Excellence (NICE)	GRADE	Yes	No	Yes	Yes	Unclear if considered used
What Works Centre for Wellbeing	CERQual	Yes	Yes	No	No	Yes
What Works Centre for Wellbeing	GRADE	Yes	Yes	Yes	Yes	Yes (but it depends on the protocol and selection criteria)

Table 2: Questions asked in the assessment process

Organisation	Framework	Does it work?	Who does it work for?	Under what circumstances?	How does it work?	At what cost?
Bond	Evidence Principles	Yes	Yes	Yes	Yes	No
Centre for Analysis of Youth Transitions	Standards of Evidence[43]	Yes	No	No	No	No
Dartington Service Design Lab	Standards of Evidence	Yes	Yes	Yes	Yes	Yes
Early Intervention Foundation	Evidence Standards	Yes	Yes	Yes	Yes	Yes
Education Endowment Foundation	Teaching and Learning Toolkit	Yes	Yes	Yes	Yes	Yes
Education Endowment Foundation	Evaluation Padlock Rating	No	No	No	No	No
HACT	Standards of Evidence	Yes	Yes	No	No	Yes
National Institute for Health and Care Excellence (NICE)	GRADE	Yes	Yes	Yes	Yes	Yes
Nesta	Standards of Evidence	Yes	Yes	No	No	Yes
Project Oracle	Standards of Evidence	Yes	No	No	No	No
Realising Ambition	The Confidence Review	Yes	No	No	No	No
What Works Centre for Crime Reduction	EMMIE Framework	Yes	Yes	Yes	Yes	Yes
What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	Yes	Yes (where evidence base allows)	Yes	Yes (where evidence base allows)	Yes (where evidence base allows)
What Works Centre for Wellbeing	GRADE	Yes	Yes (but depends on how narrowly the question is defined)	Yes (but depends on how narrowly the question is defined)	No[44]	Yes (but depends on how narrowly the question is defined)[45]
What Works Centre for Wellbeing	CERQual[46]	Yes	Yes	Yes	Yes	No

Table 3: Ranking and rating

Organisation	Framework	Use of a numbered or ascending scale
Bond	Evidence Principles	Yes. The five principles are: voice and inclusion, appropriateness, triangulation, contribution, and transparency. For each principle the checklist has four questions that can be used to test the quality of a piece of evidence. Each question is scored on a 1 – 4 scale, allowing the user to test the quality of a piece of evidence on a scale: 1) weak evidence, 2) minimum standard, 3) good practice, 4) gold standard. An overall score and colour (red, amber, light green or dark green) is then assigned to each principle to provide a holistic picture of the robustness of a piece of evidence, and to help people understand where they are in the data collection process.
Centre for Analysis of Youth Transitions	Standards of Evidence	Yes.[47] There are two scales. a) Programme Impact Rating When assessing impact grades, two components are considered: a) Reach: the extent to which the programme attracts its intended audience and b) Significance: the effect that the programme is having on young people to influence health and wellbeing. This is on a 0 – 4 scale: • 0: No impact in terms of reach and significance; or the impact was not eligible; or the impact was not underpinned by quality research produced by the submitted research outputs. • 1: Impact is of little reach and significance. • 2: Recognised but modest reach in terms of reach and significance. • 3: High impact in term of reach and significance. • 4: Very high impact in terms of reach and significance for all of the intended outcomes. b) Standards of Evidence rating The quality of all materials is considered, and assessed against the following criteria: • 0: No evidence • 1: Weak evidence provided. • 2: Acceptable standard of evidence provided. • 3: Good evidence provided. • 4: Excellent standard of evidence provided. These two scores combined to provide an overall performance rating: 1. Poor 2. Fair 3. Average 4. Good 5. Excellent
Dartington Service Design Lab	Standards of Evidence	Yes, although the scale is not numbered. Within each of the four dimensions there are sub-categories which rank the intervention's evidence as 'good enough' or 'best'.

Organisation	Framework	Use of a numbered or ascending scale
Early Intervention Foundation	Evidence Standards	Yes. • Level 4 recognises programmes with evidence of a long-term positive impact through multiple high-quality evaluations. • Level 3 recognises programmes with evidence of a short-term positive impact from at least one high-quality evaluation. • Level 2 recognises programmes with preliminary evidence of improving a child outcome, but where an assumption of causal impact cannot be drawn. • NL2 (not level 2) distinguishes programmes whose most robust evaluation evidence does not meet the Level 2 threshold for a child outcome, so do not yet have direct evidence about the scale of impact of the programme at a 'preliminary' level. • NE (found not to be effective in at least one rigorously conducted study) is reserved for programmes where there is evidence from a high-quality evaluation of the programme that it did not provide significant benefits for children. This rating should not be interpreted to mean that the programme will never work, but it does suggest that the programme will need to adapt and improve its model, learning from the evaluation.
Education Endowment Foundation	Teaching and Learning Toolkit	Yes. There are 3 components: 1) Cost: scored on a 1 to 5 scale (1 being low cost). 2) Evidence strength: on a 1 to 5 scale (1 being evidence is weak). 3) Impact months: measured in number of months.
Education Endowment Foundation	Evaluation Padlock Rating	Yes. Between 0 to 5 'padlocks' are awarded, with 5 padlocks denoting the best kind of evidence that can be expected from a single study, to 0 padlocks which denotes a study that adds little or nothing to the evidence base. The ratings take no account of whether the intervention itself was successful.
HACT	Standards of Evidence	No. It is a seven-step process. 1) Describe; 2) Design; 3) Proceed; 4) Plan; 5) Protocol; 6) Study; 7) Findings
National Institute for Health and Care Excellence (NICE)	GRADE	Yes, although it is not numbered. The certainty of evidence is classified as high, moderate, low, or very low. • High: we are very confident that the true effect lies close to that of the estimate of the effect. • Moderate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. • Low: our confidence in the effect estimate is limited. The true effect may be substantially different from the estimate of the effect. • Very low: we have very little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of effects.

Organisation	Framework	Use of a numbered or ascending scale
Nesta	Standards of Evidence	Yes. A 1 to 5 scale. 1) You can give an account of impact. 2) You are gathering data that shows some change amongst those using or receiving your intervention. 3) You can demonstrate that your intervention is causing the impact, by showing less impact amongst those who don't recieve the product/service. 4) You can explain why and how your intervention is having the impact that you have observed and evidenced so far. An independent evaluation validates the impact. In addition, the intervention can deliver impact at a reasonable cost, suggesting that it could be replicated and purchased in multiple locations. 5) You can show that your intervention could be operated by someone else, somewhere else, whilst continuing to have positive and direct impact on the outcome, and whilst remaining a financially viable proposition.
Project Oracle	Standards of Evidence	Yes. A 1 to 5 scale. 1) We know what we want to achieve. 2) We have seen there is a change. 3) We believe the change is caused by us. 4) We know how and why it works – it works elsewhere. 5) We know how and why it works. It works everywhere.
Realising Ambition	The Confidence Review	No.
What Works Centre for Crime Reduction	EMMIE Framework	Yes. Each element of EMMIE is rated on a 1 to 4 scale. There is then a final score awarded on a 1 to 8 scale. 1) No evidence to suggest that the intervention has had a statistically significant impact on crime. 2) Statistical meta-analysis suggests that overall, the intervention has had a positive and statistically reliable effect on crime. 3) Overall, the intervention has not had a statistically significant effect on crime (or this was not tested), but there is evidence that it has had a statistically significant positive impact on crime in one or more studies, or under certain conditions. 4) Overall, the intervention has not had a statistically significant effect on crime (or this was not tested), but there is evidence from one or more individual studies that it has had either a statistically significant positive or negative impact on crime, depending upon the conditions. 5) Statistical meta-analysis suggests that overall, the intervention has had a statistically significant positive effect on crime, but it has also had a statistically significant negative effect on crime on one or more studies, or under certain conditions 6) Statistical meta-analysis suggests that overall, the intervention has had a negative and statistically reliable effect on crime. 7) Overall, the intervention has not had a statistically significant effect on crime (or this was not tested), but there is evidence that it has had a statistically significant negative impact on crime in one or more studies, or under certain conditions. 8) Statistical meta-analysis suggests that overall, the intervention has had a statistically significant negative effect on crime, but it has also had a statistically significant positive effect on crime in one or more studies, or under certain conditions.

Organisation	Framework	Use of a numbered or ascending scale
What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	Yes. Level 1: Either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups or periods. Level 2: Use of adequate control variables and either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. In (a), control variables or matching techniques used to account for cross-sectional differences between treated and control groups. In (b), control variables are used to account for before-and-after changes in macro-level factors. Level 3: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference). Justification given to choice of comparator group that is argued to be similar to the treatment group. Evidence presented on comparability of treatment and control groups. Techniques such as regression and (propensity score matching may be used to adjust for difference between treated and untreated groups, but there are likely to be important unobserved differences remaining. Level 4: Quasi-randomness in treatment is exploited, so that it can be credibly held that treatment and control groups differ only in their exposure to the random allocation of treatment. This often entails the use of an instrument or discontinuity in treatment, the suitability of which should be adequately demonstrated and defended. Level 5: Reserved for research designs that involve explicit randomisation into treatment and control groups, with Randomised Control Trials (RCTs) providing the definitive example. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance. There should be limited or, ideally, no occurrence of 'contamination' of the control group with the treatment.
What Works Centre for Wellbeing	GRADE	Yes, although it is not numbered. The evidence is graded as: • High quality: Further research is very unlikely to change our confidence in the estimate of effect. • Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. • Low quality: further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. • Very low quality: any estimate of effect is very uncertain.
What Works Centre for Wellbeing	CERQual	Yes, the What Works Centre for Wellbeing uses the same scale as GRADE, which is not numbered. • High quality: Further research is very unlikely to change our confidence in the estimate of effect. • Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. • Low quality: further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. • Very low quality: any estimate of effect is very uncertain.

Endnotes

1. https://www.bond.org.uk/data/files/Effectiveness_Programme/120828Full_Bond_checklist_and_guide.pdf
2. http://www.pewtrusts.org/en/multimedia/data-visualizations/2015/results-first-clearinghouse-database
3. By social policy we mean the activities that affect human quality of life, and includes education, health, social care, and criminal justice, to name a few areas.
4. In some documents OFFA refer to these as the Standards of Evaluation Practice and as Standards of Evidence. We assume that these names refer to the same framework.
5. Mulgan, G. and Puttick, R. (2014) 'From Good Intentions to Real Impact: Rethinking the role of evidence in education businesses.' London: Nesta. Available online: https://www.nesta.org.uk/sites/default/files/from_good_intentions_to_real_impact_wv.pdf
6. Standards of Evidence for London evolved into Project Oracle's Standards of Evidence. For details see Ilic, M. and Bediako, S. (2011) 'Project Oracle'. In Puttick, R. (ed) 'Evidence for Social Policy and Practice Perspectives on how research and evidence can influence decision making in public services.' London: Nesta.
7. Puttick, R. (2011) 'Using Evidence to Improve Social Policy and Practice Perspectives on how research and evidence can influence decision making.' London: Nesta. Available online: https://www.nesta.org.uk/sites/default/files/using_evidence_to_improve_social_policy_and_practice.pdf
8. For example, we take for granted how institutionalised evidence is in health and medicine, but it has not always been this way. (For more details on the debates in the UK, see Timmins, N., Rawlins, M. and Appleby, J. (2016) 'A Terrible Beauty: A Short History of NICE'.)
9. Crawford, C., Dytham, S. and Naylor, R. (2017) 'The Evaluation of the Impact of Outreach Proposed Standards of Evaluation Practice and Associated Guidance.' Office for Fair Access. Available online: https://www.offa.org.uk/wp-content/uploads/2017/06/Standards-of-Evaluation-Practice-and-Associated-Guidance-FINAL.pdf
10. For more details on RAMESES see www.ramesesproject.org
11. Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J. and Pawson, R. (2013) RAMESES publication standards: Realist syntheses. 'BMC Medicine.' 11, 21.
12. Nuffield Department of Primary Care Life Sciences (2016) 'The RAMESES Projects.' Oxford: University of Oxford. Available online: https://www.phc.ox.ac.uk/research/interdisciplinary-research-in-health-sciences/research-studies/rameses
13. NICE (2012) The guidelines manual: Process and methods [PMG6] Published date: November 2012. Available online: https://www.nice.org.uk/process/pmg6/chapter/reviewing-the-evidence
14. Anttila,S., Johannes, P., Vareman, P. and Nils-Eric, S. (2016) Is GRADE Confusing Statistical Terms and Should They Use the Term 'Conclusiveness' to Inform Decision-Making? 'Journal of Clinical Epidemiology.' Volume 75, July 2016, Pages 1-5 c.
15. Catch 22 (2017) 'Dartington Social Research Unit Standards of Evidence.' Available online: https://www.catch-22.org.uk/services/realising-ambition/learning-far/dartington-social-research-unit-standards-of-evidence/
16. Project Oracle (2018) Validation against the Standards. Available online: https://project-oracle.com/about-us/
17. Sutton Trust (2014) 'Teaching and learning Toolkit.' Available online: https://www.suttontrust.com/about-us/education-endowment-foundation/teaching-learning-toolkit/
18. Puttick, R. and Ludlow, J. (2013) 'Nesta Standards of Evidence.' London: Nesta. Available online: https://www.nesta.org.uk/publications/nesta-standards-evidence
19. Bond (2012) 'An introduction to the principles for assessing the quality of evidence.' Available online: https://www.bond.org.uk/data/files/Effectiveness_Programme/120828Full_Bond_checklist_and_guide.pdf
20. Education Endowment Foundation (2016) 'Classification of the security of findings from EEF evaluations.' Available online: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/2016_Classifying_the_security_of_EEF_findings.pdf
21. EIF Evidence Standards, available online: http://www.eif.org.uk/eif-evidence-standards/
22. Centre for Analysis of Youth Transitions (2015) 'Evidence.' Available online: http://cayt.mentor-adepis.org/evidence/
23. Centre for Analysis on Youth Transitions (2017) 'CAYT Database of Evidence-based Programmes: Scoring Application Form.' Available online: http://cayt.mentor-adepis.org/wp-content/uploads/2015/10/CAYT-Scoring-Application-Form-2017.pdf
24. Text provided by the What Works Centre for Local Economic Growth via email, April 2018.
25. Text provided by the What Works Centre for Local Economic Growth via email, April 2018.

26. What Works Centre for Local Economic Growth (2015) 'The Maryland Scientific Methods Scale (SMS).' Available online: http://www.whatworksgrowth.org/resources/the-scientific-maryland-scale/
27. The Confidence Framework (2016) 'What is the Confidence Framework?' Available online: https://www.theconfidenceframework.org.uk/
28. HACT (2016) 'Standard for Producing Evidence – Effectiveness of Interventions – Part 1: Specification.' HACT, UK. Available online: http://www.hact.org.uk/sites/default/files/StEv2-1-2016%20Effectiveness-Specification.pdf
29. What Works Centre for Crime Reduction (2015) 'About the Crime Reduction Toolkit and EMMIE.' Available online: http://whatworks.college.police.uk/toolkit/About-the-Crime-Reduction-Toolkit/Pages/About.aspx
30. Snape, D. (2017) 'A guide to our Evidence Review Methods.' London: What Works Centre for Wellbeing. Available online: https://whatworkswellbeing.files.wordpress.com/2017/04/wwcw-methods-guide-mar-2017.pdf
31. Snape, D. (2017) 'A guide to our Evidence Review Methods.' London: What Works Centre for Wellbeing. Available online: https://whatworkswellbeing.files.wordpress.com/2017/04/wwcw-methods-guide-mar-2017.pdf
32. Project Oracle (2018) 'Validation Support.' Available online: https://project-oracle.com/what-we-offer/impact-pioneers-programme/validation-support/
33. Project Oracle Evidence Hub available online: https://project-oracle.com/
34. What Works Centre for Local Economic Growth (2015) How to use our evidence reviews. Available online: http://www.whatworksgrowth.org/resources/how-to-use-the-evidence-reviews/
35. Data correct in February 2018.
36. Discussed in an interview during March 2018.
37. This does not include analysis for Toolkits.
38. Data correct in April 2018.
39. Data correct in March 2018.
40. https://www.nesta.org.uk/blog/we-dont-need-what-works-we-need-know-what-working
41. The Confidence Framework (2016) 'Frequently Asked Questions.' Available online: https://www.theconfidenceframework.org.uk/faqs/
42. Project Oracle is currently reviewing how projects stay on the standards indefinitely, so this 'snap shot' of impact may change in future.
43. Centre for Analysis on Youth Transitions (2017) 'CAYT Database of Evidence-based Programmes Scoring Application Form.' Available online: http://cayt.mentor-adepis.org/wp-content/uploads/2015/10/CAYT-Scoring-Application-Form-2017.pdf
44. The answer given is no, but it is worth noting that the answer would come out through the theoretical framework before the GRADE process.
45. This is asked separately to the GRADE process.
46. The questions addressed will depend on the review question, as defined in the protocol.
47. Centre for Analysis on Youth Transitions (2017) CAYT Database of Evidence-based Programmes Scoring Application Form. Available online: http://cayt.mentor-adepis.org/wp-content/uploads/2015/10/CAYT-Scoring-Application-Form-2017.pdf

nesta

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500
information@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk
www.nesta.org.uk