

# An evidence rating scale for New Zealand

Understanding the effectiveness of  
interventions in the social sector

*Using Evidence for Impact*

MARCH 2017

## About Superu

---

Superu is a government agency that focuses on what works to improve the lives of families, children and whānau.

What we do:

- We generate evidence that helps decision-makers understand complex social issues and what works to address them.
- We share evidence about what works with the people who make decisions on social services.
- We support decision-makers to use evidence to make better decisions to improve social outcomes.

We also provide independent assurance by:

- developing standards of evidence and good practice guidelines
- supporting the use of evidence and good evaluation by others in the social sector.



© Crown Copyright



This work is licensed under the Creative Commons Attribution 4.0 International licence. In essence, you are free to copy, distribute and adapt the work, as long as you attribute the work to the Crown and abide by the other licence terms.

To view a copy of this licence, visit <https://creativecommons.org/licenses/by/4.0/international/>. Please note that no departmental or governmental emblem, logo or Coat of Arms may be used in any way which infringes any provision of the Flags, Emblems, and Names Protection Act 1981. Attribution to the Crown should be in written form and not by reproduction of any such emblem, logo or Coat of Arms.

---

**Superu**  
PO Box 2839  
Wellington 6140

**Telephone:** 04 917 7040  
**Email:** [enquiries@superu.govt.nz](mailto:enquiries@superu.govt.nz)  
**Website:** [superu.govt.nz](http://superu.govt.nz)

Follow us on Twitter: [@nzsfamilies](https://twitter.com/nzsfamilies)

Like us on Facebook: Social Policy Evaluation and Research Unit

ISBN 978-0-947489-70-0 (print)  
ISBN 978-0-947489-69-4 (online)

Learn more at: [superu.govt.nz](http://superu.govt.nz)



# Contents

---

<b>01 Overview</b>	<b>2</b>
1.1 Why are we doing this?	2
1.2 How might the standard be used?	2
1.3 What do we want to achieve?	3
1.4 What is the rating scale?	3
<b>02 Using the rating scale</b>	<b>4</b>
<b>03 The evidence rating scale</b>	<b>5</b>
3.1 What does the rating scale do?	5
3.2 The rating scale	5
3.3 What do we mean by evidence?	7
3.4 How do we define “robust”?	7
3.5 The effectiveness scale	8
3.6 The strength of evidence scale	10



# 01\_ Overview

The best decisions are made when supported by robust and relevant evidence. Understanding the quality of the evidence is a crucial part of decision making. This rating scale provides a standard against which the evidence for the effectiveness of social sector interventions can be assessed. This contributes to a better understanding of what works, improved programme design, and better evaluation and collection of information.

## 1.1\_ Why are we doing this?

There are three main reasons for developing a standard:

- the government wants to increase the use of evidence to support social investment and an evidence-informed approach<sup>1</sup>
- decision-makers wish to make evidence-informed decisions that improve outcomes for our most vulnerable citizens
- programme promoters (government, the community and voluntary sector and private organisations) need better information to run better programmes.

Superu, as an independent Crown entity, works where an impartial view is important and where expertise across the social sector is needed, especially with regard to family and whānau wellbeing. Part of its role is providing independent assurance, of which these standards of evidence are a part.

## 1.2\_ How might the standard be used?

The standard can be used by a range of stakeholders, from funders to service providers, to support programme design, to support funding allocation decisions across programmes, to assess the effectiveness of specific programmes and to help build capacity to collect and use evidence across the sector.



<sup>1</sup> Cabinet has defined social investment [CAB-15-MIN-0280 refers] as: “putting the needs of people who rely on public services at the centre of decisions on planning, programmes and resourcing, with systematically measuring the effectiveness of services, so we know what works well and for whom, and then feeding these learnings back into the decision-making process as an integral part of that.”



### 1.3\_ What do we want to achieve?

The use of a standard, consistent approach to the assessment of the effectiveness of programmes will enable:

- a consistent and transparent approach to assessing:
  - > whether a programme has good evidence for positive outcomes
  - > whether a programme has good evidence for no, or negative outcomes
  - > whether a programme has insufficient evidence about outcomes
- an explicit mechanism for evidence to influence these decisions
- increased visibility of evaluation as a decision-making tool
- better quality evaluation, more use of evidence in decision-making, and more sharing of findings about what programmes work.

### 1.4\_ What is the rating scale?

Rather than a binary standard (met/unmet), the rating scale provides a pathway to excellence, and a set of criteria against which programme evidence can be assessed.

The evidence assessment framework looks at the strength of the evidence, taking into account the maturity or otherwise of the programme. It would be expected that programmes would move along the spectrum as they accumulate experience and as outcomes are produced.

Strength of evidence alone is not sufficient, however. It is also important to understand the effectiveness identified – be that beneficial or harmful – to both the target population and to others who might be affected.

This gives an assessment matrix against which interventions can be measured. The full rating scale differentiates between New Zealand and overseas evidence, demanding more from overseas so that transferability and issues of scale are properly assessed.

Effectiveness	New Zealand					
	0	1	2	3	4	
Beneficial			✓	✓	✓	
Mixed	✓	✓	✓	Consider weight of evidence, risk, alternatives	Consider weight of evidence, risk, alternatives	
No effect	Strong theory of change with evidence-based logic	Too soon for effectiveness data, but processes and data suggest it is on track	Consider stopping	✗	✗	
Harmful			✗	✗	✗	

**Strength of evidence**

## 02 Using the rating scale

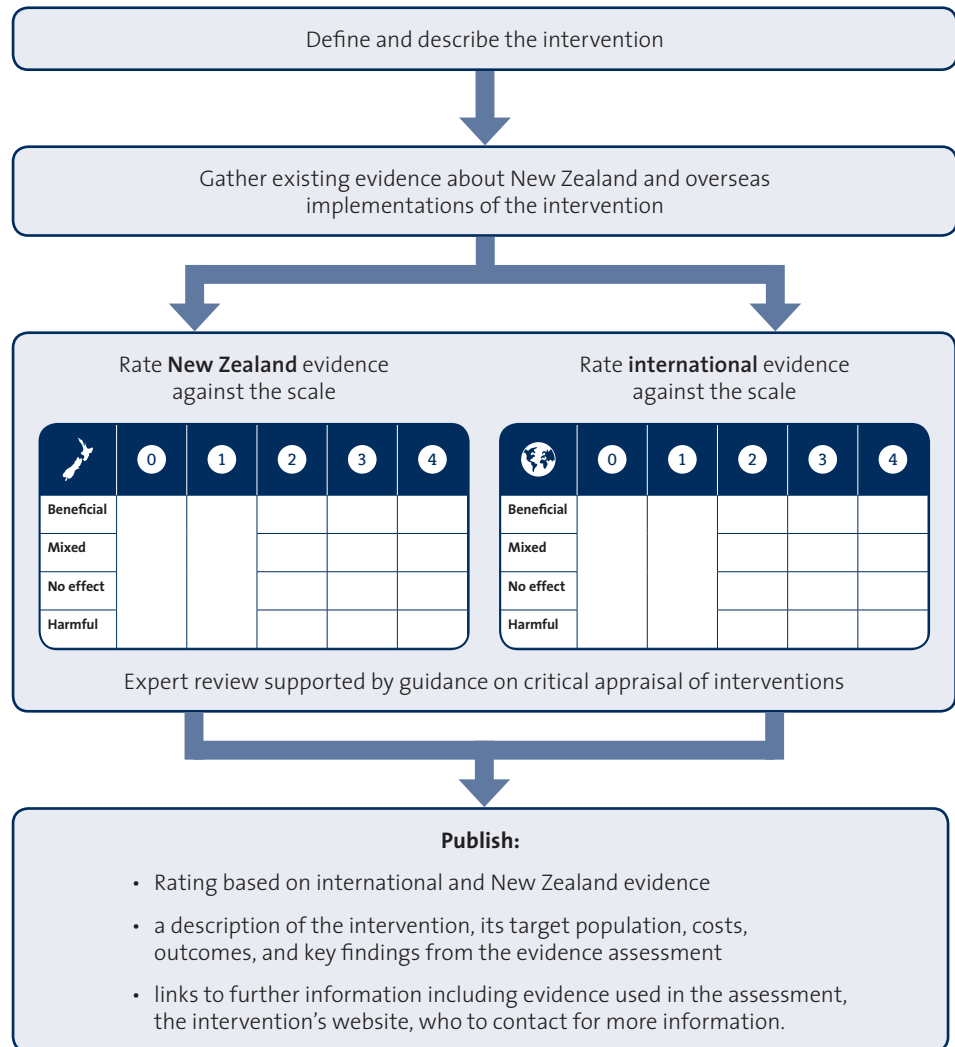
The framework can be used as a **self-assessment tool** by funders or providers, or as part of the design and budget process to help with decisions on resource allocation. However, initiatives may also request **independent assurance**, which can be provided by Superu through the use of expert review panels.

Standards of evidence are used in a number of ways overseas.<sup>2</sup> In New Zealand, the scale has been developed as a basis for evaluation guidance and funding support, and as a way to promote the development and uptake of evidence-based interventions.

The scale can be used across the range of stakeholders, from small community organisations wanting to provide a rationale for funding or to build their evidence of effectiveness, to central government agencies designing and funding large-scale investments in the social sector.

It can also go beyond this to be developed as a peer review-based external accreditation system, which would, among other things, provide support to funders in decision-making.

### Accrediting interventions against the rating scale?



<sup>2</sup> [www.superu.govt.nz/publication/in\\_focus\\_standards\\_of\\_evidence](http://www.superu.govt.nz/publication/in_focus_standards_of_evidence)



## 03 The evidence rating scale

---

The rating scale provides a standard of evidence that can be used to grade social interventions according to the strength of evidence for their effectiveness, and their suitability for scale-up or implementation in new locations. The standard uses a rating scale based on the level of evidence and the intervention maturity.

### 3.1\_ What does the rating scale do?

---

It can be used to rate:

- social interventions that are already operating in New Zealand
- innovative social interventions that are proposed for development in New Zealand
- social interventions that operate overseas and could be implemented in New Zealand.

The information about how interventions rate against the standard can be used to inform:

- decisions about implementing, continuing, stopping and scaling-up social interventions in New Zealand
- evaluation planning for interventions that have the potential to move up to higher levels of the scale.

The development of this scale has drawn on the findings of Superu's publication: *Standards of evidence for understanding what works: International experiences and prospects for Aotearoa New Zealand*. This paper states that a "New Zealand standard of evidence would help us to develop a more consistent and transparent mechanism for making evidence-based decisions about the future of an intervention."

The scale does not judge the quality of individual pieces of research, nor does it set out to judge the value of research on a specific sector, for which other tools are already available.

### 3.2\_ The rating scale


---

The scale consists of two dimensions: one that assesses the strength of evidence about an intervention (the strength of evidence scale), and one that assesses whether the intervention has had beneficial, or other effects on participants (the effectiveness scale). The scales can be combined into a two-dimensional array, and assessments can inform decisions on whether an intervention operating in New Zealand should be continued, scaled-up or stopped (see Panel A), or whether interventions operating overseas could be considered for implementation in New Zealand (see Panel B). To access continuing support, New Zealand interventions should have, or be working towards, good evidence that benefits participants. Overseas interventions should have very good evidence for benefits and information about how they work so that we can judge how good a fit they are likely to be for New Zealand<sup>3</sup> and what, if any, contextualising might be required for New Zealand conditions.

---


<sup>3</sup> For information on how this scale was arrived at, please see *Standards of evidence for understanding what works: International experiences and prospects for Aotearoa New Zealand*: [www.superu.govt.nz/publication/in\\_focus\\_standards\\_of\\_evidence](http://www.superu.govt.nz/publication/in_focus_standards_of_evidence)

**Panel A: Should we fund or continue this New Zealand intervention?**

Effectiveness		0	1	2	3	4
		Pilot initiative	Early stage, good in theory	Progressing, some evidence	Good evidence, sufficient for most interventions	Extra evidence for large or high risk interventions
	Beneficial			✓	✓	✓
	Mixed	✓	✓	✓	Consider weight of evidence, risk, alternatives	Consider weight of evidence, risk, alternatives
	No effect	Strong theory of change with evidence-based logic	Too soon for effectiveness data, but processes and data suggest it is on track	Consider stopping	✗	✗
Harmful			✗	✗	✗	

**Strength of evidence** →

**Panel B: Should we consider implementing this intervention in New Zealand?**

Effectiveness		0	1	2	3	4
		Pilot initiative	Early stage, good in theory	Progressing, some evidence	Good evidence	Well supported, large scale
	Beneficial			Maybe depending on alternatives	Maybe depending on alternatives	✓
	Mixed	✗	✗	✗	Consider weight of evidence, risk, alternatives	Consider weight of evidence, risk, alternatives
	No effect	No effectiveness data yet	No effectiveness data yet	✗	✗	✗
Harmful			✗	✗	✗	

**Strength of evidence** →





### 3.3\_ What do we mean by evidence?

---

Evidence in this context refers to information that helps to turn strategic priorities and other objectives into something concrete, manageable and achievable.<sup>4</sup> This definition emphasises the importance of the processes of turning a mass of information into an evidence base on which defensible judgements can be made. Evidence may be qualitative or quantitative and may come from various sources including performance monitoring, research, evaluation, statistics and information from experts or stakeholders.

The evidence base is dynamic – new research is produced, and existing data can be subject to new analysis or interpretation in the light of changed circumstances or priorities.

Evidence is also a necessary but not sufficient element of the decision making process. Wider social, cultural and political factors will always also play a role in shaping policies and interventions.

Different bodies, sectors, world views and research traditions have their own understandings on what constitutes “robust” evidence. In some instances, the rigour required relates to risks, such as potential life-threatening consequences as may be found in the health sector, or involving major financial investment. In contrast, a much lower standard may be the norm for lower financial or social risk initiatives or innovative/pilot implementations.

### 3.4\_ How do we define “robust”?

---

We need to distinguish between the quality of individual pieces of evidence, and the overall strength of a body of evidence. These proposed standards relate to the overall strength of a body of evidence, but clearly some judgement on the robustness of the individual contributions is also important.

There are again many frameworks that have been developed to judge the quality of research and evaluation. In this context, the judgement needs to include qualitative as well as quantitative evaluation approaches. It also needs to tell the difference between high quality and appropriate use of the approach and inappropriate or poor quality use – and indeed to use it in an appropriate manner in this context.

What many of the frameworks have in common, when looking at individual pieces of evidence is an assessment of reliability, robustness, transparency, validity and rigour – expressed in a variety of forms, and subject to much debate on interpretation.

Examples include:

- *Quality in Qualitative Evaluation: A framework for assessing research evidence*, from the UK Government Chief Social Researcher’s Office (in turn based on a review of 29 qualitative assessment frameworks)
- OECD-DAC *Quality Standards for Development Evaluation*
- *Guidance and Protocols*, from the Campbell and Cochrane collaboration (more geared towards systematic reviews than evaluation).

---

<sup>4</sup> Shaxson, L. (2005). ‘Is your evidence robust enough? Questions for policy makers and practitioners.’ *Evidence & Policy*, 1(1).

Different audiences also have different concerns with regard to quality – evaluation practitioners and commissioners may be more interested in technical aspects, whereas policy makers or funders may approach it from a need to understand the quality of the programme. There is also an assumption that the administrative/financial convenience of a “programme” is what matters to end users, rather than an outcome involving a range of related interventions.

A recent study for the UK Department for International Development<sup>5</sup> suggests a quality framework for individual studies should consider:

- how the study/evaluation has been conducted
- the technical quality of designs and methods
- the normative, ethical and institutional context.

A similar approach could be adopted to take account of the specific situation in New Zealand.

### 3.5\_ The effectiveness scale

---

While the strength of evidence scale grades interventions according to the strength of their supporting evidence, it does not specify what the evidence says about the intervention. An intervention may be supported by evidence that meets level 2 to level 4 criteria in the strength of evidence scale (section 3.6), but the evidence may show that the intervention has had beneficial effects on participants, or it may show no effect, mixed effects, or harmful effects. The effectiveness scale (opposite) can be used to provide more information on interventions in levels 2 to 4, specifying whether they are beneficial, ineffective, harmful, or have mixed effects.

The effectiveness scale is only applicable to interventions that have some evidence about effectiveness. Interventions at level 0 or 1 of the strength of evidence scale can only be assigned to the ‘not applicable’ category opposite. This includes interventions that are at level 1 because they do not yet have any evidence about effectiveness, or because they have evidence, but it does not meet level 2 criteria.

---

<sup>5</sup> Stern, E. et al. (2012). *Broadening the range of designs and methods for impact evaluations*, Department for International Development Working Paper 38.



## The effectiveness scale

Effectiveness	Criteria
<b>Beneficial</b>	<p>Evidence that meets the criteria of this level of the strength of evidence scale:</p> <ul style="list-style-type: none"> <li>• demonstrates positive effects on desired outcomes, and</li> <li>• shows no substantive harmful effects of the intervention.</li> </ul> <p>If there were multiple evaluations at this level, and the overall weight of evidence supports beneficial effects even though some evaluations suggest no effect on any desired outcomes.<sup>6</sup></p>
<b>Mixed effects</b>	<p>There has been more than one evaluation that meets the criteria of this level of the strength of evidence scale, and:</p> <ul style="list-style-type: none"> <li>• some evaluations demonstrate positive effects while others demonstrate no effects on any desired outcomes</li> <li>• there has been no demonstration of substantive harmful effects of the intervention.</li> </ul> <p>The overall weight of evidence does not clearly support either beneficial effects or no effect.</p>
<b>No effect</b>	<p>Evidence that meets the criteria of this level of the strength of evidence scale:</p> <ul style="list-style-type: none"> <li>• demonstrates no positive effects on any desired outcomes</li> <li>• shows no substantive harmful effects of the intervention.</li> </ul> <p>If there were multiple evaluations at this level, and the overall weight of evidence supports no effect even though some evaluations suggest a positive effect on desired outcomes.</p>
<b>Harmful</b>	<p>There is evidence that meets the criteria of this level of the strength of evidence scale that:</p> <ul style="list-style-type: none"> <li>• shows substantive harmful effects of the intervention on the target public, or</li> <li>• has harmful effects on others that outweigh the benefits to the target group.</li> </ul>
<b>Not applicable</b>	<p>Interventions at level 1 of the strength of evidence scale do not yet have sufficient evidence with which to assess effectiveness. Depending on the nature and scale of pilot interventions, this may also be the case here.</p>



<sup>6</sup> A weight of evidence judgement takes into account the number of evaluations, the relative strength of evidence underpinning the evaluations, and the relative importance and reliability of different outcome measures.

### 3.6\_ The strength of evidence scale

---

The strength of evidence scale consists of five levels. The levels correspond not just to ascending rankings for strength of evidence, but also to expectations about the type of evidence that can and should be generated about an intervention as it matures and grows. Level 1 is appropriate for new interventions that are as-yet untested, but have a good theoretical basis and an evaluation plan. Level 4 is appropriate for mature, large-scale interventions with a strong evidence base. Levels 2-3 guide an “evidence journey”, describing the intermediate steps between level 1 and level 4. There is also a level 0 which applies specifically to pilot initiatives, where there may be an appetite for higher risk, but also a requirement for robust evidence gathering as part of the intervention.

Level 4 of the scale requires large-scale implementation and a very strong evidence base. Interventions that reach this level will be rare, especially among those that operate only in New Zealand, but we think it is reasonable to expect New Zealand-only interventions to develop their evidence base over time, reaching level 3 within three to 10 years. Providers and funders of very large or high-risk New Zealand interventions might wish to put in extra effort to reach level 4.

The strength of evidence scale has been designed to be inclusive of different evaluation traditions. It should be able to be used to judge the strength of evidence from any evaluation approach, so long as the evaluation has:

- addressed questions about efficiency, effectiveness, and impact
- used recognised methods
- been rigorously carried out.

In particular, both western and Māori approaches to evaluation can be used to meet the criteria.

Although indicative timescales are suggested, these need to take into account the nature of the programme, and the point at which outcomes are actually expected.





## The strength of evidence scale

Level	Description	Criteria
0	Intervention is a pilot of a new initiative.	<p><b>Effectiveness</b></p> <ul style="list-style-type: none"> <li>Has a plausible and evidence-based logic model or theory of change that describes what the intervention is, what change it hopes to achieve and for whom, and how the intervention is supposed to work (how its activities will cause change). Assumptions and risks should be clearly documented.</li> <li>There is an evaluation plan for the intervention that describes how the intervention's efficiency and emerging information on effectiveness, and impact will be measured, including a plan for how attribution of impact to the intervention will be assessed. The plan specifically looks at issues of scale and transferability.</li> </ul>
1	Intervention is in its early stages of implementation, or planned but not yet implemented. This intervention's evidence base will be built over time.	<p><b>Effectiveness</b></p> <ul style="list-style-type: none"> <li>Has a plausible and evidence-based logic model or theory of change that describes what the intervention is, what change it hopes to achieve and for whom, and how the intervention is supposed to work (how its activities will cause change).</li> <li>There is an evaluation plan for the intervention that describes how and when the intervention's efficiency, effectiveness, and impact will be measured, including a plan for how attribution of impact to the intervention will be assessed.</li> </ul>
2	Typically, this intervention has been in operation for around one to three years. It has met all level 1 criteria and has been evaluated at least once. The evaluation indicates some effect, but it may not yet be possible to directly attribute outcomes to it. This intervention's evidence base will continue to be built over time.	<p><b>Effectiveness</b></p> <ul style="list-style-type: none"> <li>There is information about the efficiency of the intervention, (the delivery of outputs in relation to inputs).</li> <li>An evaluation that assesses effectiveness has been completed and reported on.</li> <li>The evaluation used a convincing method to measure change, such as pre- and post-analysis, or a recognised qualitative method. It is not required to have used a comparison group.</li> <li>The evaluation used valid and reliable methods and measurement tools that are appropriate for participants and relevant to what the intervention is trying to achieve.</li> <li>The evaluation has analysed data appropriately and its conclusions are supported by the findings.</li> </ul> <p><b>Intervention consistency and documentation</b></p> <ul style="list-style-type: none"> <li>There is clarity and documentation about what the intervention comprises.</li> <li>Procedures are in place to ensure consistent implementation of the intervention, for example manuals and staff training processes.</li> <li>Information is available on the resources (money and people) required to deliver the intervention.</li> </ul>

Level	Description	Criteria
<p style="text-align: center;"><b>3</b></p>	<p>Typically, this intervention has been in operation for around three to 10 years. It has an established design which is consistently implemented, and quality assurance procedures are in place. It has met all the level 2 criteria, plus it has at least one evaluation that provides evidence about impact. It also has some information available that will help with implementation in new contexts.</p>	<p><b>Effectiveness</b></p> <ul style="list-style-type: none"> <li>• At least one evaluation that assesses impact has been completed and reported on.</li> <li>• Evaluation has measured change using pre- and post-analysis of outcomes.</li> <li>• Evaluation has investigated attribution of impacts to the intervention using a comparison group or other appropriate comparison data, ideally with long-term follow-up. <ul style="list-style-type: none"> <li>&gt; If it is not possible or extremely difficult to obtain an appropriate comparison group or comparison data, the evaluation has used other methods to examine causation, such as checking whether the evidence is consistent with the causal relationships in the theory of change, and investigating alternative explanations for the observed outcomes.</li> <li>&gt; If it is not possible or extremely difficult to carry out long-term follow-up, there is good evidence that the intermediate outcomes that have been measured predict long term outcomes.</li> </ul> </li> <li>• Evaluation has used valid and reliable methods and measurement tools that are appropriate for participants and relevant to what the intervention is trying to achieve.</li> <li>• Evaluation has analysed data appropriately and conclusions are supported by the findings.</li> <li>• There has been an assessment of the cost of the intervention relative to the impacts it generates.</li> <li>• There is evidence that supports the causal mechanism, indicating how the intervention leads to outcomes.</li> </ul> <p><b>Intervention consistency and documentation</b></p> <ul style="list-style-type: none"> <li>• There is clarity and documentation about what the intervention comprises.</li> <li>• There is regular review of procedures, manuals and staff training processes.</li> <li>• Information is available on the resources (money and people) required to deliver the intervention.</li> </ul>





Level	Description	Criteria
4	<p>Typically, this intervention has been in operation for around eight years or longer and is large scale or high risk, justifying extra evaluation effort. It has met all the level 3 criteria, plus it has been replicated at least once. It has been evaluated at least twice and the evaluations provide strong evidence about effectiveness and impact, insights into how the intervention causes change, what works well or less well for different participants, and cost-benefit. There is support for implementation in new contexts.</p>	<p><b>Effectiveness</b></p> <ul style="list-style-type: none"> <li>At least two evaluations that assess impact have been undertaken, covering at least two implementations (e.g. the original implementation plus at least one replication). These evaluations have measured change using pre- and post-analysis, have used valid and reliable methods and measurement tools that are appropriate for participants and relevant to what the intervention is trying to achieve, have analysed data appropriately, and their conclusions are supported by the findings. The evaluations have used comparison groups or other appropriate comparison data, ideally with long-term follow-up, but if this was not possible or extremely difficult, other methods of examining causation or intermediate outcome measures that are good predictors of long-term outcomes were used.</li> <li>There has been at least one cost-benefit analysis, using methods that meet established standards.</li> <li>There is evidence that supports the causal mechanism, indicating how the intervention leads to outcomes.</li> <li>There is evidence about which elements of the intervention are necessary to implement with fidelity, and which can be adapted.</li> <li>There is evidence of the impact of the intervention on different sub-groups in the target population, for example, outcomes for different age groups, ethnicities, genders.</li> <li>There is evidence that the intervention is consistently delivered as planned and reaches its target groups.</li> </ul> <p><b>Intervention consistency and documentation</b></p> <ul style="list-style-type: none"> <li>There is clarity and documentation about what the intervention comprises.</li> <li>There is regular review of procedures, manuals and staff training processes.</li> <li>Information is available on the resources (money and people) required to deliver the intervention.</li> <li>Technical support is available to help implement the intervention in new settings.</li> </ul>



