JUNE 2016 | Social Policy Evaluation and Research Unit

In Focus

Standards of evidence for understanding what works: International experiences and prospects for Aotearoa New Zealand

Good intentions for social interventions* are not always enough. Decision-makers need quality evidence to know whether the products or services they develop, invest in or deliver make a positive difference. Then we can avoid interventions such as Scared Straight[1], the crime prevention programme, which had no evidence base and caused harm to the young people it was trying to influence.

About *In Focus*

Superu's *In Focus* series is designed to inform and stimulate debate on specific social issues faced by New Zealanders. We draw on current policy, practice and research to fully explore all sides of the issue.

The topic of what works (and what does not) is not widely discussed in New Zealand. But this situation is changing, for a number of reasons. The Government has recently shifted from a social spending approach to one based on social investment. This change aims to improve outcomes for the most vulnerable and requires systematic measurement of social service effectiveness[2]. In New Zealand there is a need to build a learning system by strengthening the quality, use and sharing of evidence about social services, policy formation and evaluation[3,4].

Where do we begin to tackle these system-wide needs for quality evidence? How do we know which interventions are effective, promising or harmful? How can we make better evidence-based investments?

International jurisdictions have grappled with these issues and developed standards of evidence to assess whether interventions can be shown to be effective. Standards of evidence are tools that help decision-makers know how confident they can be that an intervention is responsible for its claimed outcomes. Standards help to directly feed evidence into the system in a rigorous and systematic way. They show people how to gather better evidence, increase accountability and share information on what works.

This *In Focus* examines a series of international and national standards of evidence. It provides a high-level synthesis of the different approaches to assessing intervention effectiveness.

We found key differences in the purpose and application of different standards of evidence. Some have a developmental approach where building evidence capability is a priority, while others have stricter criteria for demonstrating effectiveness. Most international standards take a Western perspective on the strength of evidence, but a few have been specifically developed to show what works from an indigenous perspective.

> "It is noteworthy that within the global conversation, there is growing recognition of the critical need to be more rigorous both in the employment of evidence for the development of policy, and in the assessment of its implementation" – Sir Peter Gluckman[4].

Based on our analysis of international standards of evidence and the need for understanding what works in New Zealand, we believe that a national standard should be developed and would:

- be based on a developmental approach to help build both capability and the evidence base
- consider Māori and Western perspectives to address what works for Māori and non-Māori from the outset
- require evidence of effectiveness and evidence that supports successful replication
- build towards the use of cost-benefit evidence to demonstrate value for money.

* We use the word 'intervention' to cover policies, programmes, and practices.

Our approach

Selected national and international peer-reviewed literature, government publications and grey literature on standards of evidence were reviewed. We searched academic databases, government and organisational websites using search terms such as 'evidence-based policy', 'evidence-based decision-making', 'programme effectiveness', 'evidence', 'quality of evidence', 'evidence criteria', 'evidence models', 'evidence standards' and 'measuring effectiveness'. There was no year restriction on the literature. It is important to note that this publication focuses on standards of evidence that assess interventions. There are other types of standards that guide research and evaluation practices, such as the Aotearoa New Zealand evaluation standards [5] but these are out of scope for this publication.

We would especially like to acknowledge the following people for their input to this publication:

Dr Nick Axford (Dartington Social Research Unit, UK), Nina Jetha (Public Health Agency of Canada), Susan Courage (Canadian Best Practice Initiative, Public Health Agency of Canada), Michael O'Donnell (Bond for International Development, UK), Steve Aos (Washington State Institute for Public Policy, US), Sharnee Moore (Australian Institute of Family Studies), Sue Holloway (Project Oracle, London, UK), Dr Fiona Cram (Centre for Social Impact), Dr Te Kani Kingi (Research Centre for Māori Health and Development, Massey University, NZ).

What are standards of evidence?

Standards of evidence can be used as a framework for grading interventions and specifying the level of evidence needed to reach each grade. Usually a highly graded intervention will have strong evidence for effectiveness, while an intervention with a lower grade will have no or only emerging evidence about effectiveness, or strong evidence demonstrating ineffectiveness or harm.

What do we mean by 'evidence' and 'strength of evidence'?

The Oxford Dictionaries define evidence as the available body of facts or information indicating whether a belief or proposition is true or valid[6]. Evidence can be quantitative or qualitative, and may come from various sources including performance monitoring, research, evaluation, statistics and information from experts or stakeholders.

However, different types of evidence have varying degrees of credibility. When we talk about strength of evidence we mean the level of confidence we can have that the findings are credible* and generalisable** to other situations[7].

These general principles are often used to judge the robustness of research evidence but are applied to a policy context in this *In Focus*. If a study concludes that an intervention is effective an assessment of the strength of evidence helps us to understand how confident we can be about that conclusion.

A standard can help to answer these questions:

> How strong is the evidence base for an intervention and what further evidence should we collect?

> Should we implement an overseas intervention in New Zealand?

> Should we roll out a New Zealand intervention more widely?

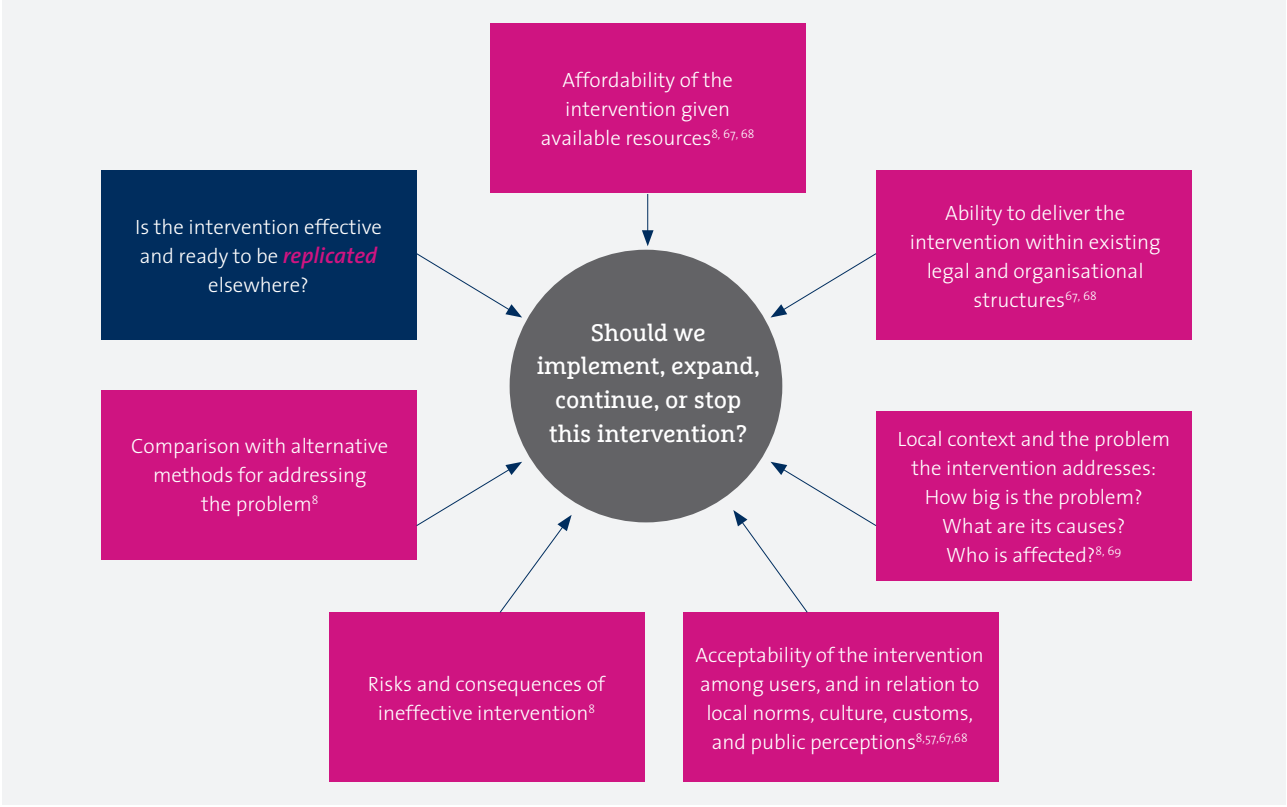
> Should we continue or stop an existing intervention?

Decisions about interventions will continue to be complex and politicised[8,9] and the use of a standard should not exclude expert advice or affected people from decision-making. Rather, it should help to integrate research evidence with other influences.

Credibility – Evidence has credibility when we can be confident in the conclusions presented because of the rigour of the analytic method used.
**Generalisability* – This refers to the inferences we can make from the evidence. For example, can we use the evidence in a different context or to answer a different question?

Figure 1_Assessment of an intervention informed by standards of evidence *(blue box)* as one of several inputs into the decision-making process

Other factors *(pink boxes)* also need to be considered



Replication refers to whether an intervention is suitable for scale-up or implementation in a new location. Successful replication requires the right combination of fidelity (keeping essential elements of the intervention the same), and adaptation (changing the adaptable elements to suit the new context).

Standards of evidence vary depending on their purpose

We examined eight case studies of standards used by international clearinghouses, including those in Australia, Canada, the United Kingdom and the United States[10–12], and two national standards. Clearinghouses have been developed in response to drivers from government and funding agencies to increase the use of evidence in evidence-informed decision-making[13]. Some clearinghouses grade interventions using their standards and publish the results for use. Selection of the eight case studies in the current *In Focus* was based on a larger sample of international websites that compile and assess evidence-informed interventions. These websites are listed in Superu's publication '*Finding and appraising evidence for what works*'[10]. Analysis of the case studies identified underlying dimensions along which different standards can be placed. We developed an organising framework for describing the dimensions, illustrated in Figure 2 below. Five major dimensions were identified (i.e. Levels, Entry Criteria, Includes Replication, Cost-benefit and Worldview).

Figure 2_Framework for describing standards of evidence

**Cost-benefit analysis (CBA)* compares the cost of an intervention with its outcomes assigning dollar values to costs and outcomes and calculating the net cost or benefit associated with the intervention[14].

Cost-effectiveness analysis (CEA) measures costs in monetary terms and outcomes in non-monetary quantitative units. Interventions can be compared when their outcomes are quantified in the same units[15].

Table 1_ Five Dimensions for Describing Standards of Evidence

LEVELS *(Tiered versus single-level standards)*	Tiered standards rank interventions into multiple tiers, while single-level standards only include interventions deemed to have strong evidence for effectiveness and exclude the rest. Tiered standards are commonly comprised of 'positive tiers' for interventions with evidence for positive outcomes, 'negative tiers' for interventions with evidence for negative or harmful effects, 'null tiers' for interventions with evidence for no effect, and 'insufficient evidence tiers' for interventions without strong evidence to ascertain any kind of effect. While single-level standards are easier for users to interpret[16], tiered standards may be better at supporting decisions about whether to implement a new evidence-based intervention, because they provide more information about the relative advantages of one intervention over another[11].
ENTRY CRITERIA *(Developmental approach versus rigorous eligibility)*	There are concerns that some standards have set the bar too high, so that only a few interventions meet their criteria[17]. In this publication, standards that have more achievable entry criteria are regarded as having a developmental approach, while those that require very strong evidence are regarded as having rigorous eligibility criteria. For example, tiered standards with a developmental approach are different from other tiered standards. The lower tiers accept early-stage evidence that can be gathered as an intervention is being set up. As the intervention matures, higher tiers need stronger evidence. A tiered developmental approach encourages and guides progress through an evidence journey[8,18].
INCLUDES REPLICATION *(Inclusion, or not, of evidence to support successful replication)*	Only some standards require evidence to support replication of the intervention. Requirements can include: › evidence that the intervention has been successfully replicated in diverse contexts › evidence that there is support for replication with fidelity (for example, provision of manuals, training, or technical support) › evidence for how the intervention works, for whom and in what contexts, so as to enable adaptation.
COST-BENEFIT *(Inclusion, or not, of cost-benefit or cost-effectiveness evidence)*	Some clearinghouses publish cost-benefit or cost-effectiveness information for the interventions that they grade, and require evidence to support this analysis. CBA and CEA information helps decision-makers to compare interventions and understand which provide better value for money. Standards without CBA or CEA information focus only on the strength of evidence, and whether outcomes were positive. Standards with CBA or CEA information add extra information about how positive the outcomes were relative to cost. They allow interventions to be ranked against one another, helping decision-makers compare interventions. However, standards that require CBA evidence are limited in the number of interventions they can grade. Many interventions do not yet have rigorous quantitative evidence of outcomes, or do not have outcomes that can be quantified or monetised[19].
WORLDVIEW *(Indigenous versus Western standards)*	Most standards of evidence are grounded in Western scientific research, which values systematic and unbiased methods. Indigenous standards tend to value methods that involve communities, and that prioritise justice and action[20]. Validity in an indigenous context often means proving that the results accurately represent the knowledge, experiences, and needs of the communities involved. Consequently, indigenous research designs often do not meet Western standards for strength of evidence, and Western designs often do not meet indigenous standards[21].

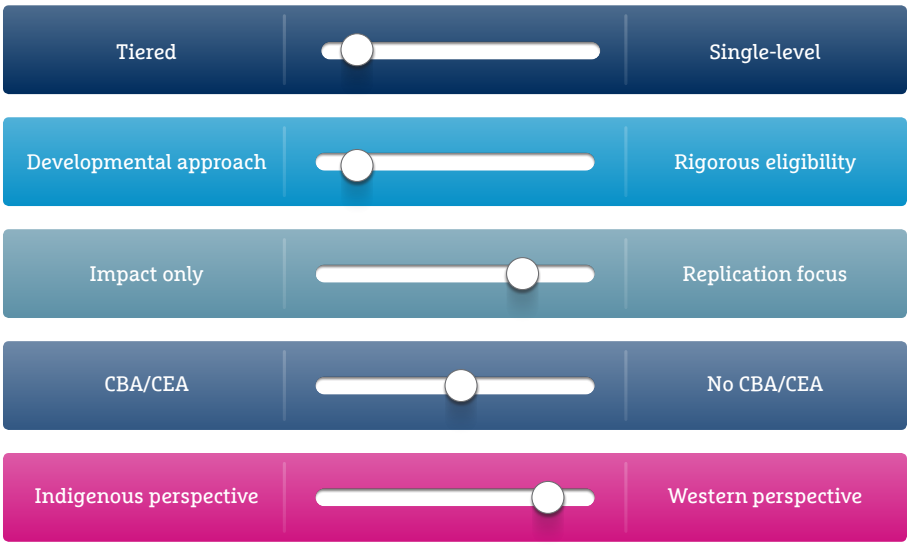
Case studies

In the examples below we have applied our organising framework to indicate where along the dimensions each standard is placed, for a quick snapshot of their purpose and focus.

1. Project Oracle (London, UK)

Project Oracle aims to improve outcomes for young people in London. It publishes information on interventions and provides evaluation support to providers[22]. Providers can apply to have their interventions validated against the Project Oracle standard of evidence, and validated interventions are listed in a searchable online database[23].

Figure 3_Features of the Project Oracle standard of evidence



The Project Oracle standard is tiered with a developmental approach. Lower tiers require early-stage evidence, while higher tiers require stronger evidence of impact and information to support replication. The Project Oracle standard has been used as a basis for the Nesta standard of evidence, which also has a developmental approach[24].

Table 2_Summary of Eligibility Criteria for Each Tier of the Project Oracle Standard of Evidence

1. Project model & evaluation plan	2. Indication of impact	3. Evidence of impact	4. Model ready	5. System ready
Theory of change Outline evaluation plan: › describe when and how you will measure impact	An evaluation report that: › Includes pre- and post- analysis › uses valid and reliable measurement tools that are appropriate for participants (comparison group not required)	At least one rigorous evaluation that: › uses a comparison group or other appropriate comparison data › ideally uses long term follow-up If the above is not possible, assessment considers the strength of underpinning theory and quality of data used to assess impact Resources to aid consistent implementation › manuals › staff training processes	At least two rigorous evaluations including: › an external evaluation › comparison data › rounded picture e.g. mixed methods, multiple outcomes, different timeframes › evidence of causal mechanism (how it works), dosage effects, impact on sub-groups, effective replication in new settings, consistent delivery as planned › cost-benefit analysis Support for replication › technical support › information on resources needed	Multiple rigorous evaluations including: › at least three independent evaluations covering at least five UK locations Support for large scale implementation and transfer to other agencies › systems that enable quality to be maintained and strong results to be consistently delivered

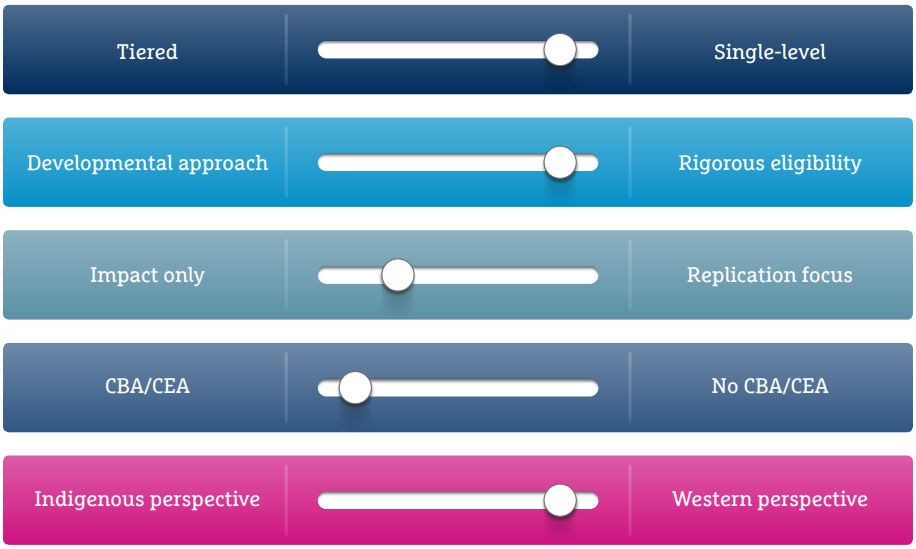
More information can be found on the Project Oracle website[25].

An evaluation of Project Oracle found that their standard of evidence had raised aspirations and helped providers to think about evaluation[22]. However, higher tiers were not understood as well and were felt to be unachievable. Higher tiers were revised, but even so, as of March 2016, none of the 291 validated interventions reached level 4 or 5, and only six interventions had reached level 3[23].

2. Washington State Institute for Public Policy (Washington, USA)

The Washington State Institute for Public Policy (WSIPP) is an independent research institute of the Washington State Legislature. Their main role is to provide unbiased information to the legislature on topics such as evidence-based initiatives[19].

Figure 4_Features of the WSIPP standard of evidence



The WSIPP publishes estimated costs and benefits of policy options for Washington State[26]. They identify interventions that have sufficient evidence to meet their standard, carry out *meta-analyses* to quantify outcomes, and then estimate the costs and benefits for Washington[27]. Table 3 below outlines the main standards-relevant aspects of this process. There are criteria for including or excluding evidence and issues that result in effect size adjustments, because they influence the strength of evidence.

A *meta-analysis* is a type of systematic literature review that uses statistical techniques to synthesise findings. A systematic literature review answers a research or evaluation question by collecting and summarising all of the evidence that fits a set of pre-specified eligibility criteria.

Table 3_ Aspects of the WSIPP Process That Address Standards of Evidence

Analytical step	Standards of evidence-relevant criteria*
Identification of evidence-based initiatives	Evidence can be included if it: › is peer reviewed or non-peer reviewed › uses a comparison group (*randomised controlled trials* are preferred but *quasi-experimental designs* are accepted if there is good comparability between the treatment and comparison groups) › uses an intent-to-treat sample (all participants are included, not just those who completed the programme) › has enough information to allow calculation of an effect size.
Conduct the meta-analysis and compute the economics	Effect sizes (and thereby the cost-benefit result) may be adjusted according to: › the credibility of the outcome measures › the relevance of the context of the study to real world settings › the strength of the research design (how prone it is to bias) › whether the researcher was involved in intervention implementation (researcher involvement tends to be associated with better outcomes than are seen in real world settings) › whether the comparison group received no treatment, or alternative treatments.

*Detailed information can be found in Lee and Aos (2011) and Washington State Institute for Public Policy (2016)[19,26].

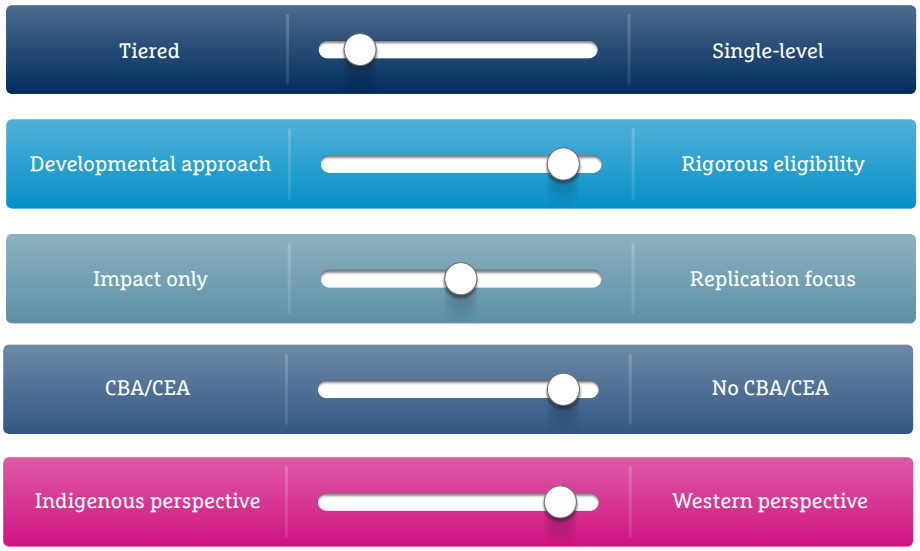
In *randomised controlled trials* (RCTs) eligible participants are randomly assigned to either the 'treatment group' who take part in the initiative, or the 'control group' who do not take part. Outcomes are compared between the two groups, and the effect of the intervention is calculated as the difference in outcomes between the two groups[28].

Quasi-experimental designs (QEDs) compare participants' outcomes to the outcomes of a comparison group of non-participants. But there is no random assignment, and the two groups may differ in more ways than just their participation or non-participation in the intervention. Statistical techniques are used to correct for differences between the two groups. There are a number of different types of quasi-experimental designs, and some are better than others at avoiding or compensating for selection bias[28].

3. California Evidence-Based Clearinghouse for Child Welfare (California, USA)

The California Evidence-Based Clearinghouse for Child Welfare's (CEBC) mission is to advance the effective implementation of evidence-based practices for children and families involved in the child welfare system[29]. Their website provides a searchable database of interventions that have been rated using The CEBC Scientific Rating Scale[30].

Figure 5_Features of the CEBC standard of evidence



This standard is a fairly typical example of a tiered approach that grades interventions using what is known as a methodological hierarchy. These hierarchies use study design as a key marker of the strength of evidence, usually placing RCTs in the top tier followed by QEDs. Other designs are either excluded or placed in lower tiers[8]. While they are used by many clearinghouses, methodological hierarchies have been subject to some criticism which is discussed later in this paper.

Table 4_Summary of Criteria for Each Tier of the CEBC Standard of Evidence

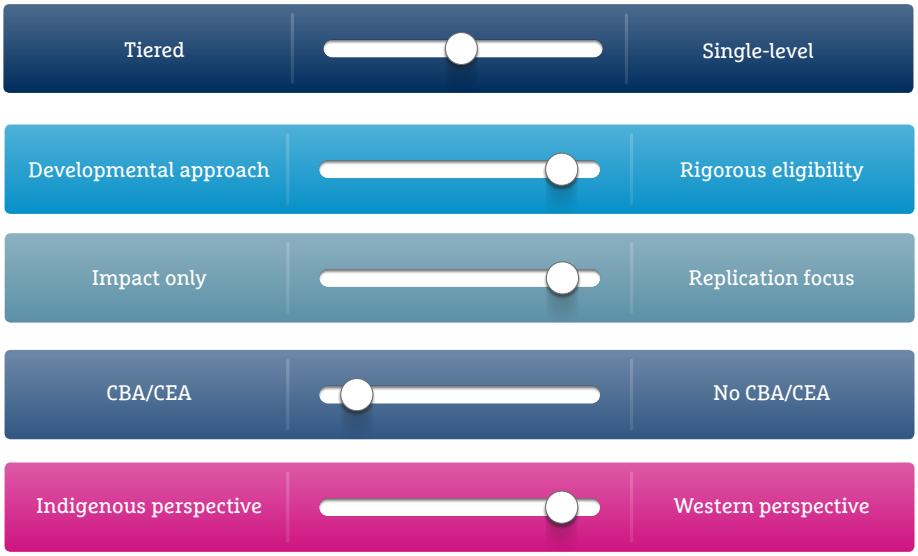
In addition to the criteria shown in Table 4 all evidence must have been peer-reviewed[30].

1. Well supported	2. Supported	3. Promising	4. No effect found	5. Concerning	6. Cannot be rated
≥2 rigorous RCTs in different settings show positive effect, using reliable, valid measures At least one of the RCTs shows sustained effect at least 1 year after treatment If multiple studies, overall weight of evidence supports benefit No case data, legal or empirical basis to suggest risk of harm There are practice manuals or other materials that support replication	≥1 rigorous RCT shows positive effect, using reliable, valid measures At least one of the RCTs shows sustained effect at least 6 months after treatment If multiple studies, overall weight of evidence supports benefit No case data, legal or empirical basis to suggest risk of harm There are practice manuals or other materials that support replication	≥1 study using some form of control (e.g. untreated group, matched wait list) shows positive effect If multiple studies, overall weight of evidence supports benefit No case data, legal or empirical basis to suggest risk of harm There are practice manuals or other materials that support replication	≥2 RCTs show no improvement in outcomes If multiple studies, overall weight of evidence does not support benefit No case data, legal or empirical basis to suggest risk of harm There are practice manuals or other materials that support replication	If multiple studies, overall weight of evidence suggests a negative effect and/or: › There is case data, a legal, or empirical basis suggesting that, compared to its likely benefits, there is a risk of harm › There are practice manuals or other materials that support replication	No published study using some form of control (e.g. untreated group, placebo group, matched wait list) Does not meet criteria for any other level on the CEBC Scientific Rating Scale No case data, legal or empirical basis to suggest risk of harm There are practice manuals or other materials that support replication

4. Investing in Children (UK)

Investing in Children is an initiative by the Dartington Social Research Unit (DSRU), which disseminates evidence about what works in improving children's outcomes. Interventions are assessed against a standard of evidence by a board of international experts[31]. In addition, cost-benefit analyses are conducted using the method developed by the WSIPP[32]. The results are published on the DSRU website.

Figure 6_Features of the Investing in Children standard of evidence



Their standard of evidence is summarised in Table 5 below. It has two tiers: a 'good enough' tier that sets the minimum standard that an intervention must meet to be deemed evidence-based and a 'best' tier with additional criteria[33]. Both tiers require strong evidence, so the eligibility criteria are rigorous. This standard explicitly divides its assessment criteria into four areas: these are intervention specificity, evaluation quality, impact, and system readiness. The system readiness dimension includes a strong focus on replication readiness.

Table 5_Investing in Children Standard of Evidence

Question	Good enough evidence criteria	Additional criteria for best evidence
Intervention specificity Is the intervention focused, practical, logical and designed based on the best available evidence about what types of factors affect child outcomes and what works in improving outcomes?	Intended population of focus is clearly defined Outcomes are clearly specified and reflect relevant key developmental outcomes for children The risk and protective factors that the intervention seeks to change are identified in the intervention's logic model or theory Clarity and documentation about what the intervention comprises	There is a research base summarising the prior empirical evidence to support the causal mechanisms that underlie the change in outcomes being sought
Evaluation quality Are the evaluation design and execution robust enough to permit confidence in the results?	≥1 RCT or ≥2 QEDs conducted in which plausible threats to validity are controlled for* Clear statement of the demographic characteristics of the population with whom the intervention was tested What participants received in the treatment and comparison conditions are documented No evidence of differential attrition between treatment and comparison groups Outcome measures: (a) are not dependent on the unique content of the intervention (b) reflect relevant developmental outcomes (c) are not rated solely by the people delivering the intervention	≥2 RCTs or 1 RCT and 1 QED conducted, in which plausible threats to validity are controlled for* Long-term follow-up (≥12 months after intervention completion) on at least one outcome measure Results indicate the extent to which fidelity of implementation affects impact Dose-response analysis is reported Where possible, analysis of the impact on sub-groups Verification of the theoretical rationale underpinning the intervention
Impact What do robust evaluations tell us about how much impact the intervention has on key developmental outcomes for children?	Positive impact on a relevant key developmental outcome A positive and statistically significant effect size* No adverse effects for intervention participants	Evidence of positive impact and an absence of adverse effects from a majority of the studies Evidence of a positive dose-response relationship
System readiness Can the intervention be implemented in the real world context of a public service system?	Explicit processes to ensure that the intervention gets to the right people Training materials and implementation procedures Manuals detailing the intervention Information on the financial and human resources required to deliver the intervention The intervention that was evaluated is still available	The intervention is being widely disseminated The intervention has been tested in real world conditions Technical support is available to help implement the intervention in new settings A fidelity protocol or assessment checklist accompanies the intervention

* More detail is provided in Dartington Social Research Unit (2013)[33].

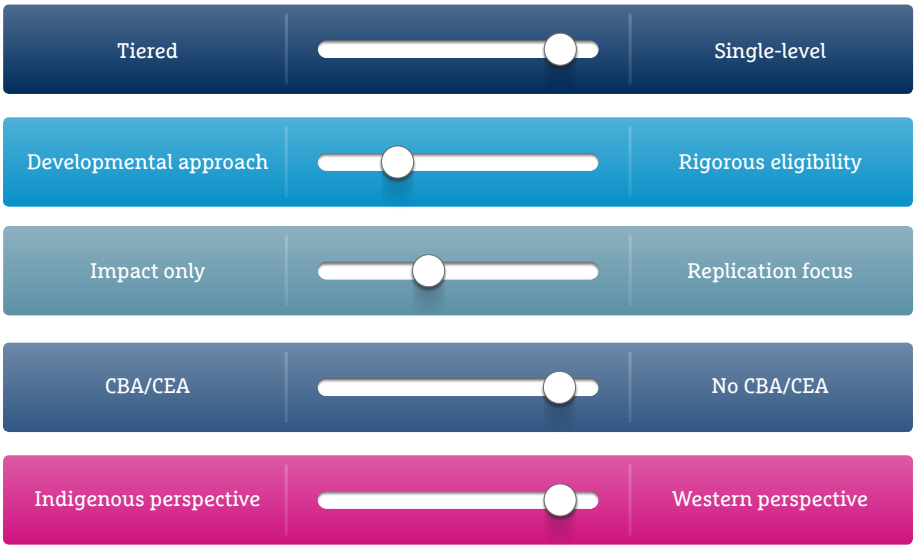
5. Child Family Community Australia (Australia)

The Child Family Community Australia (CFCA) sits within the Australian Institute of Family Studies and is an information exchange for people working with children, families and communities. They have two standards of evidence: one that supports the selection of ***Children Facilitating Partners*** evidence-based programmes and another that supports the ***Knowledge Circle Practice Profiles***.

5a. Children Facilitating Partners (Australia)

The Australian Department for Social Services requires that Communities for Children Facilitating Partners organisations use 30 percent of their funding for high-quality, evidence-based services[34]. The standard described in this section is used to determine which interventions are eligible for this funding.

Figure 7_Features of the Children Facilitating Partners standard of evidence



This standard is not tiered. Only interventions that meet all of the criteria in Table 6 below are listed on the CFCA website and are eligible for funding[35]. The criteria are relatively easy to meet and a range of research designs are accepted as long as there were at least 20 participants. This may have been a pragmatic choice as more stringent criteria may limit the number of interventions that can be funded.

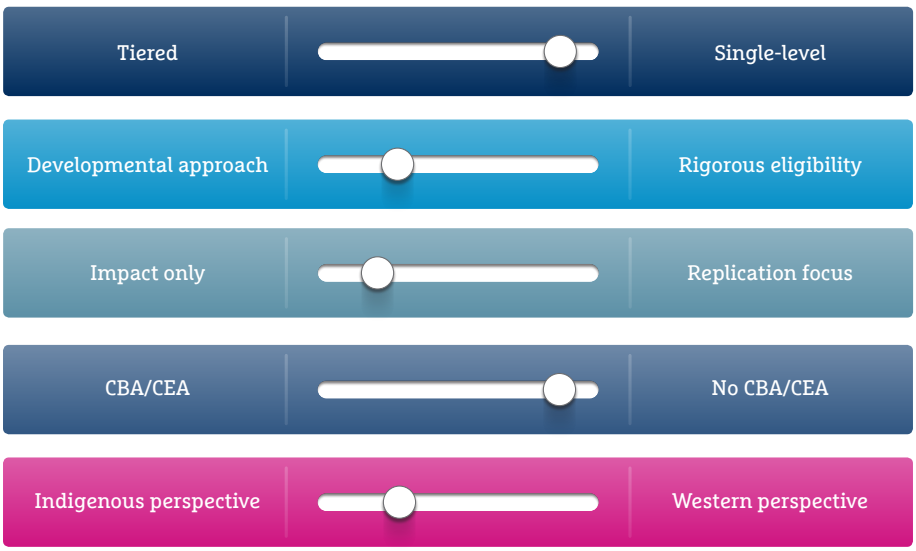
Table 6_Children Facilitating Partners – Evidence-based Programme Profile Standard

Criteria for inclusion in the CFCA evidence-based programme profiles*†
The programme must have documented the following: › theoretical and/or research background › programme logic › target group and activities.
The programme has a training manual and has been replicated or shows potential for replication.
At least one evaluation was conducted, that: › shows positive impacts on desired outcomes and finds no negative effects and › uses a randomised controlled trial, quasi-experimental design, or pre- and post-test, with n≥20 in control and treatment groups › either uses a high-quality or a qualitative method with n≥20 (quality considers participant selection processes, sample representativeness, data collection processes, and independence) › or is a high quality combination of the above (mixed methods).

5b. Knowledge Circle Practice Profiles (Australia)

CFCA publishes the Knowledge Circle Practice Profiles, which list interventions that deliver outcomes for Aboriginal and Torres Strait Islander children, families and communities[37]. The Profiles' purpose is to share experience about what works. They are not linked to any funding incentives.

Figure 8_Features of the Knowledge Circle Practice Profiles standard of evidence



* All requirements must be met; † More information can be found on the CFCA website[36].

Interventions can be included in the Profiles if they meet the criteria shown in Table 7 below. These criteria were developed from the results of a review of aspects of service delivery that are effective for vulnerable children and families[38]. They require consultative, participatory, and culturally appropriate approaches along with a strong evaluative component. However, there is little published detail on what constitutes a culturally appropriate approach and a strong evaluative component.

Table 7_Knowledge Circle Practice Profile Standard

Criteria for inclusion in the Knowledge Circle Practice profiles*†
The programme uses culturally appropriate approaches, including: › a consultative process to identify needs › participation and involvement of the Aboriginal and Torres Strait Islander communities in decisions about planning, delivering and evaluating the program › culturally relevant tools in delivering services.
The programme is informed by research or theory, with a strong evaluation component. › Evaluation shows that desired outcomes for Aboriginal and Torres Strait Islanders have occurred in accordance with programme objectives › There are ongoing feedback and evaluation processes that improve programme delivery.

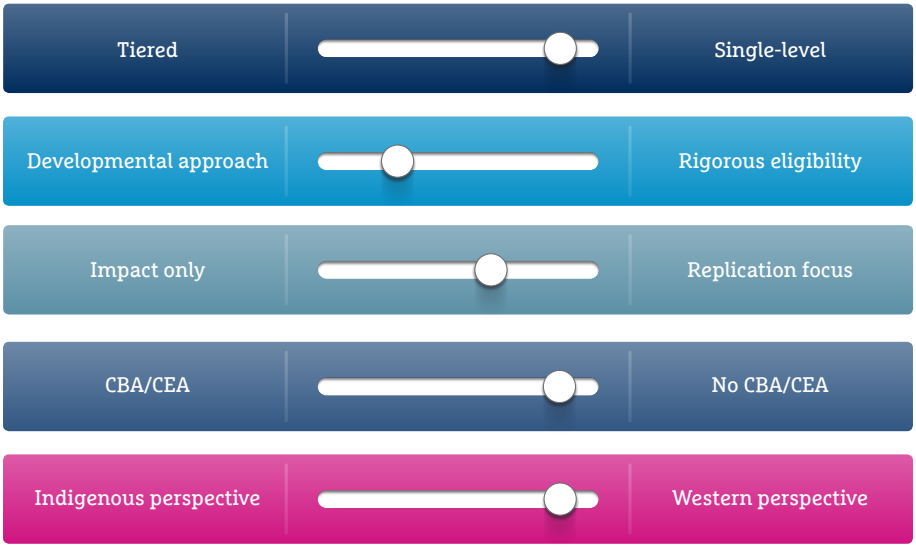
* All requirements must be met; † More information can be found on the CFCA website[39].

6. Canadian Best Practices Portal (Canada)

The Public Health Agency of Canada's Best Practices Portal lists evidence-based interventions in health promotion and chronic disease prevention[40]. The goal is to help practitioners and decision-makers identify interventions that they could implement. At the time of writing this paper the Portal used two standards of evidence. The ***Best Practice standard***, and the ***Aboriginal Ways Tried and True standard***.

6a. Canadian Practices Portal - Best Practice (Canada)

Figure 9_Features of the Best Practice standard of evidence



The Best Practice standard is not tiered. Only interventions that meet the criteria listed in Table 8 below are included in the Portal. The standard does not require particular study designs, but there is a review of whether the evidence meets quality and rigour criteria appropriate to the design. Consistent with the Portal's goal of encouraging replication, the standard requires interventions to have been replicated at least once, and to have documentation that supports implementation fidelity.

Table 8_Best Practice Standard

Criteria for inclusion in the Best Practice Standard*
The intervention must: › have been evaluated with results described in a report or peer-reviewed journal article › demonstrate effectiveness in producing a positive effect on health-related outcomes › be beyond the pilot stage and have been replicated at least once › be run by an authoritative/credible source with contact information available › have been developed free of commercial interests that could compromise integrity › be fully documented online (e.g. with a manual, resources, training materials, information on measurement of outcomes and processes).
Further assessment by the Public Health Agency of Canada checks that the evaluation meets quality and rigour criteria appropriate to the study design.

*More information can be found on the Canadian Best Practices Portal website[41,42].

6b. Aboriginal Ways Tried and True (Canada)

In 2013, the Public Health Agency of Canada found that only 23 out of the 374 interventions in the Best Practices Portal were aboriginal-specific interventions, or adaptations of mainstream interventions in aboriginal contexts[21]. Attempts to bolster this number were not very successful due to differences in research and evaluation values, different concepts of best practice, and a lack of evidence meeting the Best Practice standard. In response, the Public Health Agency worked with aboriginal communities, leaders and academics to develop a standard grounded in an aboriginal worldview[21].

Figure 10_Features of the Aboriginal Ways Tried and True standard of evidence

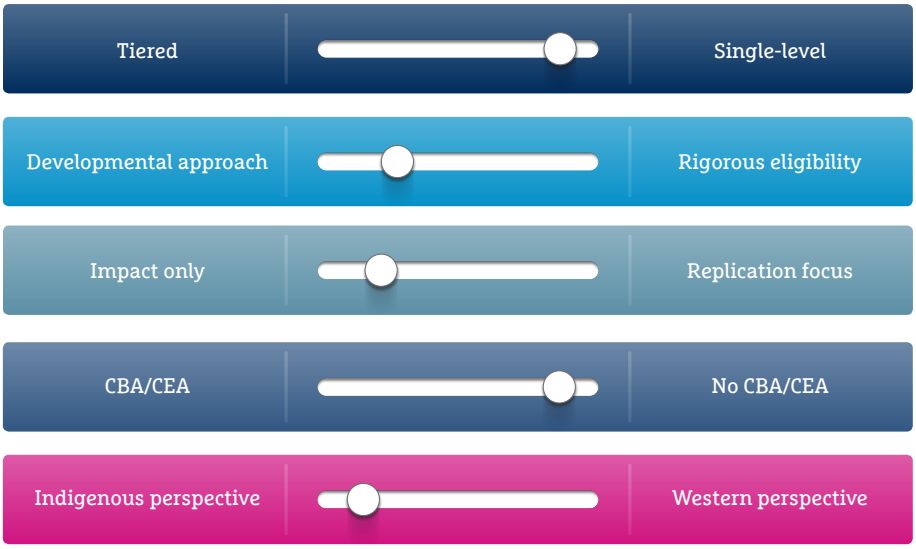


Table 9 below summarises the Aboriginal Ways Tried and True standard. Compared to standards that are grounded in Western perspectives, it focuses more on working with communities, and using collaborative and holistic approaches. These elements are thought to play an important role in intervention success, and in the quality of evidence[21]. Of note is that the aboriginal perspective argues against the notion that any one intervention will work for all communities. Instead interventions are valued when they are specific to, and developed by, the communities that they serve.

Table 9_Aboriginal Ways Tried and True Standard

Aboriginal Ways Tried and True*
The intervention must: › be community-based (with indigenous people involved in its planning, design, delivery, adaptation, and evaluation) › be holistic (addressing multiple issues, wellness, the implementation environment, the nature of target group and involving cross-sector departments) › integrate indigenous cultural knowledge (addressing and incorporating the values, culture, experiences and principles of the community in which it operates) › build on community strengths and needs (recognising community capacity or readiness, building on strengths, filling gaps) › use partnership and collaboration (using collaborative approaches to address needs, and involving other organisations inside and outside the community) › be effective (demonstrating substantive or statistically significant positive outcomes in target groups).

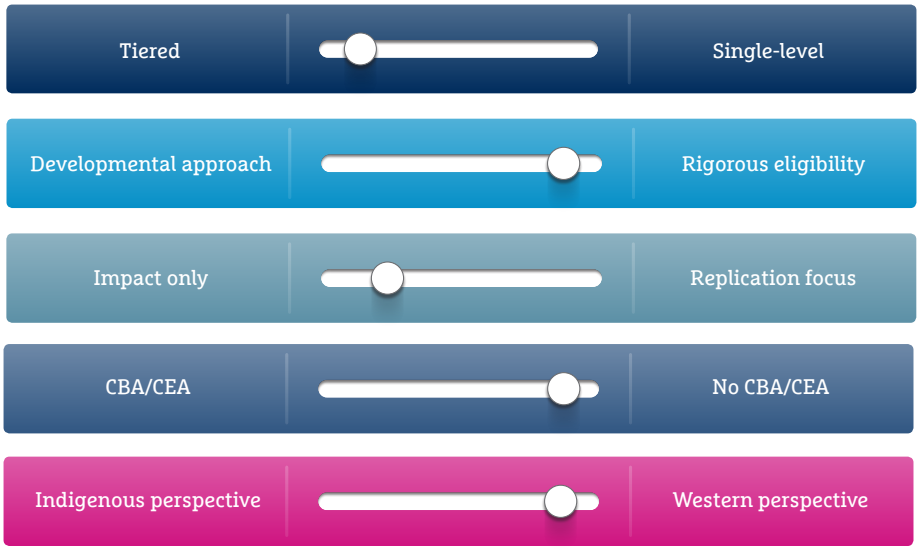
*More information can be found on the Public Health Agency of Canada website [42,43] and in Public Health Agency of Canada (2015)[21].

7. Ministry of Justice standard of evidence (New Zealand)

In New Zealand, the Ministry of Justice uses a standard of evidence to assess the robustness of evidence supporting interventions that aim to reduce crime. Assessments feed into their investment brief papers, which inform Ministry investment decisions[44].

The standard guides two separate assessments, one that grades international evidence, and one that grades New Zealand evidence. The grades are then combined to produce a six-tiered scale in which interventions are assigned to levels ranging from "Dubious" to "Very strong". More weight is given to New Zealand evidence because of concerns about the applicability of overseas evidence to New Zealand. As discussed later in this paper, even interventions that are strongly evidence-based sometimes do not produce expected outcomes when they are implemented in a new country.

Figure 11_Features of the Ministry of Justice standard of evidence



Tables 10 and 11 below show the Ministry's two-dimensional standard of evidence, and the characteristics of the interventions in the six tiers, respectively.

In Table 10 levels three, four and five of the standard for New Zealand studies use the What Works Centre for Local Economic Growth interpretation of the Scientific Maryland Scale[45]. Level 3 includes studies that compare outcomes before and after an intervention using a comparison group. Statistical adjustment may be made for differences between the treated and comparison groups, but there are likely to be important differences. Level 4 includes strong QEDs where it can be credibly assumed that treatment and comparison groups differ only in their exposure to the intervention. Level 5 is reserved for RCTs only.

Table 10_Ministry of Justice Standard of Evidence

		New Zealand studies				
		≥1 level 4 or 5 study finds statistically significant negative impact No conflicting level 4+ studies	Studies show conflicting results OR no impact OR no level 3+ study exists	≥1 level 3 study finds statistically significant positive impact No conflicting level 3+ studies	≥1 level 4 study finds statistically significant positive impact No conflicting level 4+ studies	≥1 level 5 study finds statistically significant positive impact No conflicting level 5 studies
International studies	Meta-analysis or systematic review of ≥5 studies finds significant positive impact, no conflicting results	Fair (Promising)	Very Promising	Strong	Strong	Very Strong
	Meta-analysis or systematic review with <5 studies finds positive impact OR No meta-analysis or systematic review exists and RCTs or strong QEDs find a positive impact	Speculative	Fair (Promising)	Fair (Promising)	Very Promising	Strong
	Meta-analysis or systematic review finds conflicting results	Speculative	Speculative	Fair (Promising)	Very Promising	Strong
	Meta-analysis or systematic review shows no impact OR No meta-analysis or systematic review exists	Dubious	Speculative	Fair (Promising)	Very Promising	Strong
	Meta-analysis or systematic review shows negative impact, no conflicting results	Dubious	Dubious	Speculative	Fair (Promising)	Strong

Table 11_Characteristics of Interventions in the Ministry of Justice Standard of Evidence

Tier	Interpretation
Very Strong	Very robust international and local evidence that the intervention tends to reduce crime Likely to generate a return if implemented well Simple monitoring approach should confirm the investment is providing a positive return Little additional evaluation required
Strong	Robust international and local evidence that the intervention tends to reduce crime Likely to generate a return if implemented well Could benefit from additional evaluation to confirm the intervention is delivering a positive return and to support fine-tuning of the design
Very Promising	Robust international or local evidence that the intervention tends to reduce crime May well generate a return if implemented well Further evaluation is desirable to confirm the intervention is delivering a positive return and to support fine-tuning of the design
Fair (Promising)	Some evidence that the intervention can reduce crime Uncertain whether it will generate return even if implemented well May be unproven in New Zealand or be subject to conflicting research May benefit from trial approaches with a research and development focus Robust evaluation needed to confirm the investment is delivering a positive return and to aid detailed service design
Speculative	Little or conflicting evidence that the intervention can reduce crime Highly uncertain whether it will generate return even if implemented well Primarily suited to trial approaches with a strong research and development focus Full rollout should be subject to high-quality evaluation to ensure the intervention is delivering a positive return, and to deliver insights into detailed service design questions
Dubious	Robust evidence that the intervention does not reduce crime or that it increases crime Should be priority for divestment

8. Ministry of Social Development standard of evidence (New Zealand)

The Ministry of Social Development has developed a standard of evidence for the purpose of assessing the effectiveness of larger scale community investment programmes and services[46]. This standard uses a methodological hierarchy approach. The standard can be used to identify programmes that offer little value, programmes that work well elsewhere and are worth future investment, programmes that are good candidates for robust evaluations and also ways to build continuous improvement into programmes and services. The standard has six tiers and the criteria for each tier are shown in Table 12 below.

Figure 12_Features of the Ministry of Development standard of evidence

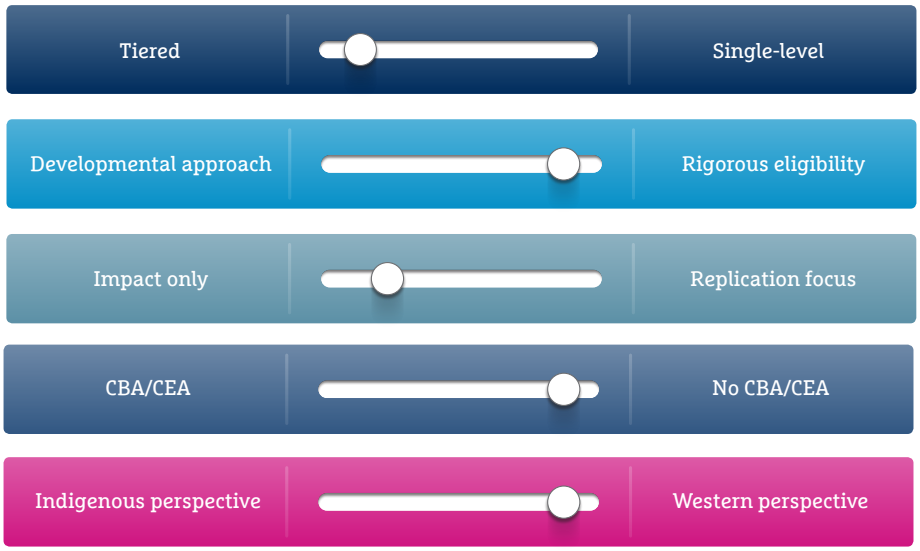


Table 12_Ministry of Social Development Standard of Evidence[46].

Level	Type of evidence
Well supported	Evidence of a positive impact on desired outcomes from ≥2 RCTs and no evidence from a well-executed study of harm
Moderately supported	Evidence of a positive impact on desired outcomes from 1 RCT and no evidence from a well-executed study of harm
Promising	Evidence of a positive impact on desired outcomes from ≥1 well-designed (non-randomised) controlled or quasi-experimental study and no evidence from a well-executed study of harm
Not effective	Evidence for the absence of any impact on desired outcomes from ≥2 RCTs and no evidence from any well-executed study of harm
Harmful	Evidence for adverse impact on any desired outcome or any other clinically significant outcome from any well-executed study
Unknown	Insufficient evidence to meet any of the above criteria

What is the current debate about standards of evidence?

There has been debate about standards of evidence, with criticism that some standards unfairly exclude certain types of interventions and do not focus enough on factors that support replication success. Recent developments have addressed some of these criticisms.

Determining whether an intervention caused an outcome is at the centre of the debate

RCTs are generally acknowledged as the best method of assessing causation because random allocation of participants to treatment and comparison groups minimises the risk that differences other than the intervention might account for outcomes[28,47]. Likewise, some types of QEDs are good at selecting treatment and comparison groups that are similar in all respects except for their participation in the intervention[28]. This has led to the development of methodological hierarchies which place interventions evaluated using RCTs in the top tier followed by those evaluated with QEDs and either excluding other interventions or placing them in lower tiers.

While few people would argue that all methods produce equally good evidence, it is generally accepted that different methods are better at answering different questions, and are appropriate for different situations[50–53]. A standard of evidence that only allows an intervention to reach the top tier if it has been evaluated using an RCT or QED design will grade interventions that are not agreeable to those methods poorly, regardless of effectiveness[8,16,49].

Furthermore, there are several types of interventions that can be difficult to evaluate credibly with RCT and QED designs. These include policies that are implemented across the whole country at once with no domestic comparison group, interventions that change and adapt making it difficult to make a 'clean' comparison, and interventions with small populations where the sample sizes are too small to have statistical power[49].

Aside from RCTs and QEDs there are now emerging methods such as theory-based approaches that can be used to understand causation[48,54,55]. These approaches aim to establish how the intervention worked and whether the outcomes were caused by the intervention, or due to other factors. Theory-based approaches can be feasible when RCT and QED designs are not. However, they cannot usually quantify the amount of change in outcomes to the intervention, limiting their ability to support CBA or CEA.

There are different views about the merits of different approaches to determining whether the intervention caused an outcome[54]. Some advice suggests that RCT or QED designs should be used wherever possible[28] while other advice suggests choosing an approach that is appropriate for the type of intervention and the available resources[52,54]. Some advice suggests combining different approaches as they are complementary[56]. A New Zealand standard will have to address this issue and reach a decision about the approach to determining cause.

What other factors affect strength of evidence?

When standards of evidence focus on methods for assessing whether an outcome is caused by an intervention, they address some aspects of strength of evidence but not others. Many problems can reduce the credibility of an evaluation including the following[7,51,55]:

- Outcome measures may not accurately reflect phenomena of interest. For example, survey respondents may interpret questions differently to what was expected.
- The analytical techniques used to interpret the results may have been applied inappropriately.
- Conflicts of interest, or a desire to confirm preconceptions, may affect how researchers collect data, interpret findings, and report results.
- The study may not reflect real world conditions, or current conditions, so we cannot generalise from it.

Standards of evidence have been criticised for their poor coverage of the range of factors that affect strength of evidence. While many standards look at how causation was determined, they are less consistent in their treatment of other factors[16]. If other factors are not considered it could result in the endorsement of interventions with limited evidence[57].

Ideally a standard would cover every factor that can affect the strength of evidence but standards need to be understood by non-specialists. The review processes that underlie standards may consider a wider range of issues than those that are described by the standards. In many cases however there is no published information about these processes so it is difficult to know what criteria are used or how consistently they are applied.

Social interventions are difficult to replicate successfully

Even interventions that are strongly evidence-based sometimes do not produce positive outcomes when they are implemented in a new context. For example, in the United States, the Nurse-Family Partnership uses trained nurses to visit low income teenage mothers to help them achieve stability and a successful start for their children. Many evaluations across the United States have found it is successful[58]. But when it was implemented in the United Kingdom it did not demonstrate any benefit over and above existing services[59].

Possible reasons for inconsistent results when an intervention is replicated elsewhere include the following[60]:

- The intervention may not have been replicated with fidelity. Key components may have been delivered differently (or not at all), or participants might be different, for example if the intervention was less tightly targeted to people in need.
- The intervention may not have been adequately adapted to the new context. Interventions normally need to be modified for new situations, addressing issues such as language, cultural acceptability, and accessibility. A lack of adequate adaptation may reduce effectiveness.
- The intervention may be less effective due to socio-demographic or cultural factors, or a different service delivery environment. For example, the Nurse-Family Partnership may have been less effective in the United Kingdom because teenage mothers there can access many statutory health and social services already. The Nurse-Family Partnership may not have provided any additional advantage[59].
- Studies of interventions in different contexts can differ in how they are carried out, with different designs or methods used. This can affect results.

Standards vary in their evidence for replication readiness

Some standards require support for replication fidelity such as manuals, training materials, or technical support.

Other standards require evidence of successful replication in multiple real world contexts. The rationale is that an intervention is more likely to be successful in a new context if it has already proved successful in diverse contexts.

Some standards require evidence of how the intervention works, who it has had beneficial effects for and under what circumstances. This can help organisations to understand whether the intervention is likely to work in their context, features that must be delivered with fidelity and features that can be adapted [61].

Indigenous standards tend to reject the notion that any one intervention will work for all communities.

The strongest evidence about an intervention is a mix of high-quality studies

Many sources of guidance now state that the strongest evidence about how effective an intervention is comes from a mixed portfolio of high-quality studies[13,50,51].

An ideal standard would have criteria for judging the strength of evidence for different methods including RCTs, QEDs, CBA, qualitative studies and others. This would allow people to use a standard to judge all forms of evidence about an intervention rather than being limited to having criteria for only some methods e.g. RCTs. One standard that addresses this issue to some extent is the Project Oracle standard. Higher tiers of this standard require a mixed portfolio of evidence including evidence of impact, how the intervention worked, effects on sub-groups, replication and cost-benefit.

The criteria for judging the strength of evidence, however, are method-specific and there are a number of challenges to be met in developing and implementing an all-encompassing standard.

How do we incentivise the use of standards of evidence?

To be useful, a standard must be accompanied by:

- a process to grade interventions against the standard
- a strategy to encourage generation of evidence that meets the standard
- a way to encourage organisations to implement interventions that meet the standard and to improve or discontinue interventions that do not.

Resources required for the process of grading interventions should not be underestimated

A standard of evidence only becomes useful when it is used to grade interventions. Broadly speaking there are two different approaches to grading interventions:

1. Some clearinghouses actively search for interventions that meet their criteria, gather the relevant information, grade interventions and publish the results. For example, research topics for the WSIPP are first selected by the Washington State Legislature. This is followed by a search for interventions and assessment of evidence by the WSIPP[62].
2. Some clearinghouses only accept nominated interventions for review. Providers submit information about their interventions. The clearinghouse reviewers then assess and validate the information against their standard.

Both processes require considerable resource and expertise. Clearinghouse managers have reported that review processes are labour intensive and resource constraints are a challenge[11]. Project Oracle reports that it takes four to seven days to validate an intervention in addition to the effort by the provider[63].

Funding, legislative and policy strategies can use standards to push for better evidence

In some cases, funding for evaluation incorporates a requirement that evaluations meet strength of evidence criteria. This approach is used in some of the federal evidence-based funding initiatives in the United States. Interventions that are promising but need more evidence are given support for evaluation as part of their funding with a requirement that the evaluation meets specified criteria[64].

Legislation or policy can be used to require or encourage organisations to apply standards of evidence to their work. The United States Federal Government requires agencies to establish procedures to ensure the objectivity, utility, and integrity of information provided to the public[13], and several agencies have developed policies requiring that evaluations meet specified criteria[65].

Funding and knowledge translation approaches encourage implementation of interventions that meet high standards of evidence

Two main approaches have been used to encourage the implementation of interventions that meet standards of evidence:

1. Funding incentive approaches assign a proportion of funding to interventions that rate well against a standard of evidence. In Australia, Communities for Children Facilitating Partners organisations must put 30 percent of their funding towards interventions that meet standards. In the United States, tiered funding initiatives allocate most of their funds to interventions that have strong evidence for effectiveness and are ready for expansion. They reserve a smaller pool of funding and support for evaluation for promising interventions with less evidence[64,66].
2. Clearinghouses use a knowledge translation approach that aims to promote the replication of evidence-based interventions by providing easily accessible, user-friendly and policy-relevant information on interventions.

Pathways forward for Aotearoa New Zealand

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. There are a few key issues related to the purpose and use of a New Zealand standard that need particular attention:

1. To really add value a standard in New Zealand needs to be accompanied by an assessment process in which reviewers with appropriate expertise grade interventions against the standard. Resourcing the grading process should be given considerable thought including which agency or agencies would be the best placed to take on this role.

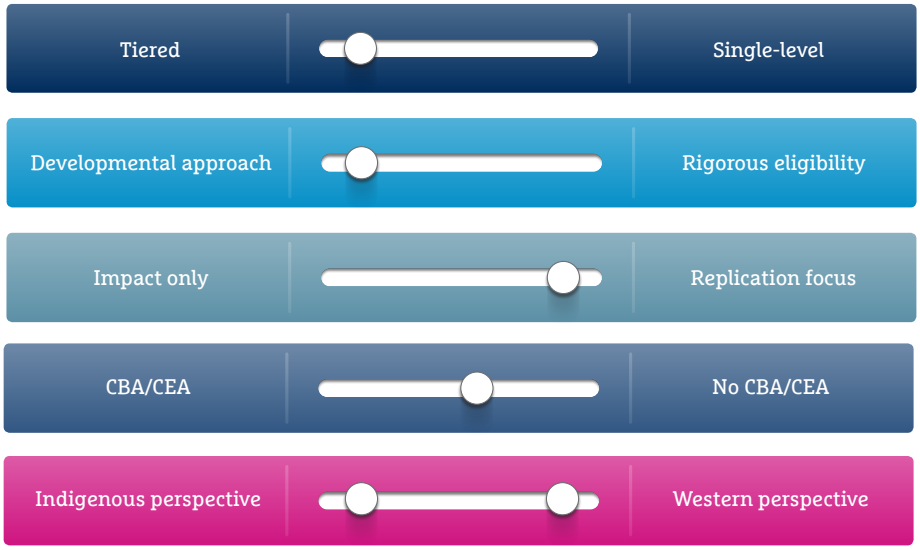
2. Crucially, we know that generating evidence to meet the standard will be a challenge for many service providers. Gaps in this area have been identified[63] and a standard of evidence will not have the desired effect of raising the quality of evidence if evaluation capability for non-government organisations is not addressed. As part of Superu's *Using Evidence for Impact* work programme, tools to support good evaluation practices such as the *Evaluation Standards for Aotearoa New Zealand*[5] and *Evaluation planning for funding applicants*[10] have been published. Further resources for evaluation capacity building are also in development. Evaluation capability is therefore a key resourcing consideration when developing a New Zealand standard.

3. Finally, most of the standards that are used by overseas clearinghouses focus on specific areas within the social sector, such as crime reduction, school level educational achievement or child welfare. These clearinghouses grade a limited range of interventions and often require demonstration of particular types of outcomes. In principle, a standard that covers the range of intervention types in the social sector is possible but it will need to be tested for its applicability to the wide range of topics.

We recommend developing a New Zealand standard of evidence

Based on our analysis of the international evidence and initial agency consultation Figure 13 below outlines the recommendation of what a future New Zealand standard of evidence might look like and why. It shows where a New Zealand standard should be placed on each of the dimensions identified in this paper.

Figure 13_Features of the recommended New Zealand standard of evidence



A tiered standard not single-level

A New Zealand standard of evidence should be tiered and not single-level because of the need to build up the evidence in the New Zealand system. A tiered standard would allow for a building-block approach to evidence gathering and the assessment of more interventions than using a single-level approach.

Use a developmental approach

A New Zealand standard should adopt a developmental approach. Its lower tiers would accept emerging evidence of the type that can be generated early in an intervention's life, while higher tiers would require more evidence, and stronger evidence. This would have the following benefits:

- It would be easier to accept interventions with emerging evidence into lower tiers. This is important given the limitations of the current evidence base in New Zealand.
- Higher tiers would require stronger evidence for effectiveness and good support for replication among more mature interventions. So the standard could be used to distinguish mature interventions that are supported by strong evidence from early stage interventions with only emerging evidence.
- It could be used to describe an evidence journey that could help to guide evaluation progress and raise aspirations around evidence.

Consider evidence that supports replication as well as impact

A New Zealand standard of evidence should require evidence to support replication or consistent implementation in addition to evidence for impact. This would be consistent with best practice and it would provide better support for decisions on whether to replicate or scale up interventions in New Zealand.

Three types of information that can support replication could be required by the standard:

- evidence that the intervention has been replicated in multiple real world contexts
- evidence that there are documents and procedures that are available to assist others to replicate the intervention with fidelity
- evidence about how the intervention works: its mechanism of action and how well different aspects work, for what people, and under what circumstances.

Further work will need to consider which requirements would be appropriate at different tiers of the standard.

Build toward cost-benefit or cost-effectiveness evidence in higher tiers

A future New Zealand standard should encourage the use of accessible methods of demonstrating value for money at the lower tiers. In higher tiers of the standard fuller CBA/CEA would be required. Very few New Zealand interventions have been subjected to robust CBA or CEA and some have outcomes that cannot be fully quantified or monetised. Insisting on full CBA/CEA evidence at entry point on the standard would limit the number of interventions that could be graded. Therefore we believe that evidence of cost-benefit should also take a developmental approach similar to the evidence journey described in this paper.

Furthermore, the expectation should be that only large-scale social interventions are required to provide full CBA/CEA evidence. For smaller scale interventions, providing more limited information on value for money would be considered good enough.

Use both Māori and Western perspectives to develop standards

A New Zealand standard of evidence should incorporate both Māori and Western approaches to evidence from conception to implementation. Based on the international evidence reviewed in this paper and after some initial consultation there are two possible approaches for New Zealand:

- to develop two separate standards, one based on Western perspectives and another based on indigenous knowledge
- to develop a single overarching standard that incorporates criteria that can be interpreted using both Western and indigenous approaches.

We have found Canadian and Australian examples of separate Western and indigenous standards, but no examples where these two perspectives are explicitly incorporated into a single standard. That is not to say a single standard will not work and should not be developed in New Zealand, but rather that further consultation is needed to ensure that any single standard speaks effectively to the needs of Māori and non-Māori.

WHAT NEXT?

This *In Focus* provides the basis for development of a Standard of Evidence Framework for New Zealand. For more information see www.superu.govt.nz

References

1. **Royster M.** The Success and Failure of Scared Straight: A Reassessment of Juvenile Delinquency Deterrent Methods and their Measurements. *International Journal of Interdisciplinary Social Sciences*. 2012;6(8):145–151. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=91821605&site=ehost-live.

2. **English B.** Speech to the Treasury Guest Lecture Series on Social Investment. 2015. Available at: https://www.beehive.govt.nz/speech/speech-treasury-guest-lecture-series-social-investment.

3. **New Zealand Productivity Commission.** *More effective social services*.; 2015. Available at: http://www.productivity.govt.nz/inquiry-content/2032?stage=4.

4. **Gluckman P.** *The Role of Evidence in Policy Formation and Implementation*.; 2013. Available at: http://www.pmcsa.org.nz/wp-content/uploads/The-role-of-evidence-in-policy-formation-and-implementation-report.pdf.

5. **Superu (Social Policy Research and Evaluation Unit), ANZEA (Aotearoa New Zealand Evaluation Association).** *Evaluation standards for Aotearoa New Zealand*.; 2015. Available at: http://www.superu.govt.nz/sites/default/files/Superu_Evaluation_standards.pdf.

6. **Oxford Dictionaries.** Definition of evidence in English. *Oxford Dictionaries*. Available at: http://www.oxforddictionaries.com/definition/english/evidence. Accessed March 15, 2016.

7. **Shaxson L.** Is your evidence robust enough? Questions for policy makers and practitioners. *The Policy Press*. 2005;1(1):101–112.

8. **Nutley S, Powell A, Davies H.** *What counts as good evidence? Provocation paper for the alliance for useful evidence*.; 2013. Available at: http://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf.

9. **Head BW.** Reconsidering evidence-based policy: Key issues and challenges. *Policy and Society*. 2010;29(2):77–94. doi:10.1016/j.polsoc.2010.03.001.

10. **Superu (Social Policy Research and Evaluation Unit)**. *Finding and appraising evidence for what works*.; 2016. Available at: http://www.superu.govt.nz/finding-and-appraising-evidence-using-evidence-impact.

11. **Burkhardt JT, Schröter DC, Magura S, Means SN, Coryn CLS.** An overview of evidence-based program registers (EBPRs) for behavioral health. *Evaluation and program planning*. 2015;48:92–9. doi:10.1016/j.evalprogplan.2014.09.006.

12. **Soydan H, Mullen EJ, Alexandra L, Rehnman J, Li Y-P.** Evidence-Based Clearinghouses in Social Work. *Research on Social Work Practice*. 2010;20(6):690–700. doi:10.1177/1049731510367436.

13. **Office of Management and Budget.** *Chapter 7. Building the capacity to produce and use evidence. In: Analytical Perspectives, Budget of the United States Government, Fiscal Year 2017*. Washington D.C.; 2016. Available at: https://www.whitehouse.gov/sites/default/files/omb/budget/fy2017/assets/ap_7_evidence.pdf.

14. **Kaplan J, Montain A.** Cost Benefit Analysis. *Better Evaluation*. Available at: http://betterevaluation.org/evaluation-options/CostBenefitAnalysis. Accessed March 24, 2016.

15. **Kaplan J.** Cost Effectiveness Analysis. *Better Evaluation*. Available at: http://betterevaluation.org/evaluation-options/CostEffectivenessAnalysis. Accessed March 24, 2016.

16. **Means SN, Magura S, Burkhardt JT, Schröter DC, Coryn CLS.** Comparing rating paradigms for evidence-based program registers in behavioral health: evidentiary criteria and implications for assessing programs. *Evaluation and program planning*. 2015;48:100–16. doi:10.1016/j.evalprogplan.2014.09.007.

17. **Sharples J.** *Evidence for the Frontline a Report for the Alliance for Useful Evidence*.; 2013. Available at: http://www.alliance4usefulevidence.org/assets/EVIDENCE-FOR-THE-FRONTLINE-FINAL-5-June-2013.pdf.

18. **Axford N, Morpeth L.** Evidence-based programs in children's services: A critical appraisal. *Children and Youth Services Review*. 2013;35(2):192–201. doi:10.1016/j.childyouth.2012.10.017.

19. **Lee S, Aos S.** Using Cost-Benefit Analysis to Understand the Value of Social Interventions. *Research on Social Work Practice*. 2011;21(6):682–688. doi:10.1177/1049731511410551.

20. **National Collaborating Centre for Aboriginal Health.** Aboriginal Research Designs - Valuing Knowledge in Context. Available at: http://www.nccah-ccnsa.ca/393/Aboriginal_research_Designs.nccah. Accessed March 24, 2016.

21. **Public Health Agency of Canada.** *Ways Tried and True Aboriginal Methodological Framework for the Canadian Best Practices Initiative*.; 2015. Available at: http://publications.gc.ca/collections/collection_2015/aspc-phac/HP35-59-2015-eng.pdf.

22. **Gloster R, Aston J, Foley B. Evaluation of Project Oracle.**; 2014. Available at: http://www.nesta.org.uk/publications/evaluation-project-oracle.

23. **Project Oracle. Project Oracle - Projects.** Available at: http://project-oracle.com/projects/. Accessed March 29, 2016.

24. **Puttick R, Ludlow J.** *Standards of Evidence : an Approach That Balances the Need for Evidence With Innovation*.; 2013. Available at: http://www.nesta.org.uk/publications/nesta-standards-evidence.

25. **Project Oracle.** Validation against the Standards. Available at: http://project-oracle.com/support/for-youth-service-providers/validation-against-the-standards/. Accessed February 25, 2016.

26. **Washington State Institute for Public Policy.** Benefit-Cost Results. Available at: http://www.wsipp.wa.gov/BenefitCost. Accessed March 24, 2016.

27. **Washington State Institute for Public Policy.** *Benefit-Cost Technical Documentation*.; 2015. Available at: http://www.wsipp.wa.gov/TechnicalDocumentation/WsippBenefitCostTechnicalDocumentation.pdf.

28. **Campbell S, Harper G.** *Quality in policy impact evaluation: understanding the effects of policy from other influences (supplementary Magenta Book guidance)*.; 2012. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190984/Magenta_Book_quality_in_policy_impact_evaluation__QPIE_.pdf.

29. **The California Evidence-Based Clearinghouse for Child Welfare.** Welcome to the CEBC: California Evidence-Based Clearinghouse for Child Welfare. Available at: http://www.cebc4cw.org/. Accessed March 24, 2016.

30. **The California Evidence-Based Clearinghouse for Child Welfare.** Scientific rating Scale. Available at: http://www.cebc4cw.org/ratings/scientific-rating-scale/. Accessed March 24, 2016.

31. **The Social Research Unit at Dartington.** *Investing in Children: An Overview*.; 2013. Available at: http://investinginchildren.eu/sites/default/files/Investing in Children - An Overview (Version 1.0 September 2013)_1.pdf.

32. **The Social Research Unit at Dartington.** *Investing in Children: Technical Report*.; 2013. Available at: http://investinginchildren.eu/sites/default/files/Investing in Children - Technical Report %28September 2013%29.pdf.

33. **The Social Research Unit at Dartington.** *The " What Works " Standards of Evidence*.; 2013. Available at: http://investinginchildren.eu/sites/default/files/Investing in.

34. **Department of Social Services.** *Communities for Children Facilitating Partner Operational Guidelines*.; 2014. Available at: https://www.dss.gov.au/our-responsibilities/families-and-children/programs-services/family-support-program/communities-for-children-facilitating-partner-operational-guidelines.

35. **Child Family Community Australia.** Communities for Children Facilitating Partners Evidence-based programme profiles. Available at: https://apps.aifs.gov.au/cfca/guidebook/programs. Accessed March 29, 2016.

36. **Child Family Community Australia.** Evidence-based programme profiles. Available at: https://aifs.gov.au/cfca/expert-panel-project/information-service-providers/frequently-asked-questions-communities-children-facilitating-partners#evidence-based. Accessed March 29, 2016.

37. **Child Family Community Australia.** A-Z listing: Knowledge Circle Practice Profiles. Available at: https://apps.aifs.gov.au/ippregister/projects/list. Accessed March 29, 2016.

38. **Child Family Community Australia.** Evidence used to develop the Knowledge Circle Practice Profiles. Available at: https://www2.aifs.gov.au/cfca/knowledgecircle/evidence-used-develop-knowledge-circle-practice-profiles. Accessed March 29, 2016.

39. **Child Family Community Australia.** Knowledge Circle Practice Profiles. Available at: https://www2.aifs.gov.au/cfca/knowledgecircle/knowledge-circle-practice-profiles. Accessed March 29, 2016.

40. **Public Health Agency of Canada.** Canadian Best Practices Portal - About Best Practices. Available at: http://cbpp-pcpe.phac-aspc.gc.ca/interventions/about-best-practices/. Accessed March 30, 2016.

41. **Public Health Agency of Canada.** Canadian Best Practices Portal - our process. Available at: http://cbpp-pcpe.phac-aspc.gc.ca/our-process/. Accessed March 30, 2016.

42. **Public Health Agency of Canada.** Canadian Best Practices Portal - Recommend an Intervention. Available at: http://cbpp-pcpe.phac-aspc.gc.ca/interventions/recommend-intervention/. Accessed March 30, 2016.

43. **Public Health Agency of Canada.** Aboriginal Ways tried and True. Available at: http://cbpp-pcpe.phac-aspc.gc.ca/aboriginalwtt/aboriginal-ways-true/. Accessed March 30, 2016.

44. **Slyuzberg M.** *Personal Communication*. April(2016).

45. **What Works Centre for Local Economic Growth.** The Scientific Maryland Scale. Available at: http://www.whatworksgrowth.org/resources/the-scientific-maryland-scale/. Accessed April 19, 2016.

46. **Mackay R.** *Personal Communication*. April(2016).

47. **Donaldson SI.** Examining the Backbone of Contemporary Evaluation Practice. In: *Credible and Actionable Evidence*. SAGE Publications Inc.; 2015:3–26.

48. **Stern E, Stame N, Mayne J, Forss K, Davies R, Befani B.** *Broadening the range of designs and methods for impact evaluations. Report of a study commissioned by the Department for International Development*.; 2012. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/67427/design-method-impact-eval.pdf.

49. **Davidson JE.** The RCTs-Only Doctrine: Brakes on the Acquisition of Knowledge? *Journal of MultiDisciplinary Evaluation*. 2006;3(6):ii–v.

50. **Breckon J (Nesta), Roberts I (Nesta)**. Using Research Evidence: *A Practice Guide*.; 2016. Available at: http://www.alliance4usefulevidence.org/assets/Using-Research-Evidence-for-Success-A-Practice-Guide-v6-web.pdf.

51. **Department for International Development.** Assessing the Strength of Evidence.; 2014. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291982/HTN-strength-evidence-march2014.pdf.

52. **UNEG Impact Evaluation Task Force.** *Impact evaluation in UN Agency Evaluation Systems: Guidance on Selection Planning and Management*.; 2013. Available at: http://www.uneval.org/document/detail/1433.

53. **Berriet-Solliec M, Labarthe P, Laurent C.** Goals of evaluation and types of evidence. *Evaluation*. 2014;20(2):195–213. doi:10.1177/1356389014529836.

54. **Rogers P.** *Overview: Strategies for Causal Attribution*.; 2014. Available at: http://devinfolive.info/impact_evaluation/ie/img/downloads/Overview_Strategies_for_Causal_Attribution_ENG.pdf.

55. **White H, Phillips D.** *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework*.; 2012. Available at: http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf.

56. **White H.** *Theory-Based Impact Evaluation: Principles and Practice*.; 2009. Available at: http://www.3ieimpact.org/media/filer_public/2012/05/07/Working_Paper_3.pdf.

57. **Littell J, Shlonsky A.** Toward Evidence-Informed Policy and Practice in Child Welfare. *Research on Social Work Practice*. 2010;20(6):723–725. doi:10.1177/1049731509347886.

58. **Washington State Institute for Public Policy.** Nurse Family Partnership for low-income families. Benefit-cost estimates updated December 2015. Literature review updated April 2012. *Benefit-Cost results*. 2015. Available at: http://www.wsipp.wa.gov/BenefitCost/Program/35.

59. **Robling M, Bekkers MJ, Bell K, et al.** Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): A pragmatic randomised controlled trial. *The Lancet*. 2015;387. doi:10.1016/S0140-6736(15)00392-X.

60. **Sundell K, Ferrer-Wreder L.** The transportability of empirically supported interventions. In: Shlonsky A, Benbenishty R, eds. *From Evidence to Outcomes in Child Welfare : An International Reader*. Oxford University Press; 2014:41–58.

61. **Porter S.** Week 22: Using evaluation in programme design – a funder's perspective. *52 Weeks of Better Evaluation Blog*. 2014. Available at: http://betterevaluation.org/blog/funders-perspective-on-eval-in-design. Accessed March 15, 2016.

62. **Lee S, Aos S.** Using Cost-Benefit Analysis to Understand the Value of Social Interventions. *Research on Social Work Practice*. 2011;21(6):682–688. doi:10.1177/1049731511410551.

63. **Ilic M, Bediako S.** Project Oracle. Understanding and sharing what really works. In: *Using Evidence to Improve Social Policy and Practice. Perspectives on how research and evidence can influence decision-making*.; 2011:52–91.

64. **Results for America.** *Federal Evidence-Based Innovation Programs Tiered-Evidence Approach The Social Innovation Fund Investing in Innovation Fund*.; 2015. Available at: http://results4america.org/policy-hub/invest-works-fact-sheet-federal-evidence-based-innovation-programs/.

65. **Results for America**. *Invest in What Works Federal Index (March 2015).*; 2015. Available at: http://results4america.org/wp-content/uploads/2015/03/2015-March-Federal-Index-v11.pdf.

66. **Office of Management and Budget.** Executive office of the President. Memorandum to the heads of departments and agencies. 2015:1–14.

67. **Flitcroft K, Gillespie J, Carter S, Salkeld G, Trevena L.** Incorporating evidence and politics in health policy: can institutionalising evidence review make a difference? *Evidence & Policy: A Journal of Research, Debate and Practice*. 2014;10(3):439–455. doi:10.1332/174426514X672399.

68. **Fisher M.** The Social Care Institute for Excellence and Evidence-Based Policy and Practice. *British Journal of Social Work*. 2014;1–16. doi:10.1093/bjsw/bcu143.

69. **Head BW.** Three lenses of evidence-based policy. *Australian Journal of Public Administration*. 2008;67(1):1–11. doi:10.1111/j.1467-8500.2007.00564.x.

Our purpose

To increase the use of evidence by people across the social sector so that they can make better decisions – about funding, policies or services – to improve the lives of New Zealanders, New Zealand's communities, families and whānau.

What we do

We work across the wider social sector to:

- **promote** informed debate on the key social issues for New Zealand, its families and whānau, and increase awareness about what works
- **grow** the quality, relevance and quantity of the evidence base in priority areas
- **facilitate** the use of evidence by sharing it and supporting its use in decision-making.

For more information about the work of Superu contact enquiries@superu.govt.nz

Superu
Level 7, 110 Featherston Street
PO Box 2839,Wellington 6140

P: 04 917 7040
W: superu.govt.nz

Follow us

facebook.com/SuperuNZ

twitter.com/nzfamilies

linkedin.com/
families-commission

Families Commission Kōmihana ā Whānau

The Families Commission operates under the name Social Policy Evaluation and Research Unit (Superu)

ISBN 978-0-478-36920-5 (print)
ISBN 978-0-478-36921-2 (online)