

# 10 steps for evaluation success

March 2019

Kirsten Asmussen, Lucy Brims and Tom McBride

## Acknowledgments

We would like to thank everyone who provided advice and comments on this guidance, including EIF colleagues Jack Martin, Sarah Taylor and Donna Molloy, as well as Jason Bowden-King from Save the Children, Max Stanford from the Children's Commission, Jonathan Breckon from Nesta, and Dione Hills from the Tavistock Institute.

## About EIF

The Early Intervention Foundation (EIF) is an independent charity established in 2013 to champion and support the use of effective early intervention to improve the lives of children and young people at risk of experiencing poor outcomes.

Effective early intervention works to prevent problems occurring, or to tackle them head-on when they do, before problems get worse. It also helps to foster a whole set of personal strengths and skills that prepare a child for adult life.

EIF is a research charity, focused on promoting and enabling an evidence-based approach to early intervention. Our work focuses on the developmental issues that can arise during a child's life, from birth to the age of 18, including their physical, cognitive, behavioural and social and emotional development. As a result, our work covers a wide range of policy and service areas, including health, education, families and policing.

EIF IS PROUD TO BE A MEMBER OF THE WHAT WORKS NETWORK



**Early Intervention Foundation** 10 Salamanca Place London SE1 7HB

W: www.ElF.org.uk E: info@eif.org.uk T: @TheElFoundation P: +44 (0)20 3542 2481

This paper was first published in March 2019. © 2019

The aim of this report is to support policymakers, practitioners and commissioners to make informed choices. We have reviewed data from authoritative sources but this analysis must be seen as a supplement to, rather than a substitute for, professional judgment. The What Works Network is not responsible for, and cannot guarantee the accuracy of, any analysis produced or cited herein.

EIF is a registered charity (1152605) and a company limited by guarantee (8066785).

#### Download

This document is available to download as a free PDF at: https://www.eif.org.uk/resource/10-steps-for-evaluation-success

#### Permission to share

This document is published under a creative commons licence: Attribution-NonCommercial-NoDerivs 2.0 UK http://creativecommons.org/licenses/by-nc-nd/2.0/uk/



For commercial use, please contact info@eif.org.uk

## Contents

Summary: 10 steps for evaluation success	5
Introducing the 10 Steps framework	10
A commitment to children	10
What works for child development	11
The EIF evidence standards	11
How to use this guide	14
Step 1: Confirm your theory of change	15
There is nothing quite so practical as a <i>science-based</i> theory	15
Using science to specify why an intervention is important, necessary and will add value	16
How to develop a science-based intervention when you are not a scientist	20
Confirming a science-based theory of change through participatory methods	23
Preparing for the next step	24
Step 2: Develop your logic model	27
Turn your theory of change into a theory of action	27
The key components of a well-specified logic model	28
Confirming the logic model	
Preparing for the next step	
Step 3: Create an intervention blueprint	
Blueprints are not just for architects	
Confirming the intervention blueprint	40
Step 4: Conduct a feasibility study	
Can the intervention work? Verifying feasibility	43
Feasibility of delivery	44
Feasibility of recruitment, retention and reach	49
Tracking service use	51
Feasibility of cost	
Setting feasibility targets	54
Step 5: Pilot for outcomes	
The value of pilot studies	57
Selecting appropriate assessment measures	60
Recruiting a sufficiently large and representative sample	62
Measuring impact	65
Exceeding the minimum: The use of a comparison group and other pilot study alternatives	67
Step 6: Test for efficacy	71
Are you ready for Step 6?	71
Increasing validity through an efficacy trial	73
Creating and maintaining ideal study circumstances	78
Statistical analyses	79
Understanding impact	80
No effect	
Step 7: Test for effectiveness	
Efficacy versus effectiveness	
Three critical questions answered by effectiveness trials	87
Effectiveness trials observing no effect	

Step 8: Refine and monitor	90
A new set of steps	90
Outcome-based monitoring systems	92
Rapid cycle evaluations for testing an intervention's active ingredients	94
Determining workforce needs	96
Step 9: Adapt and transport	100
Managed adaptation	100
Adaptation for transportation	101
Step 10: Take to scale	106
Systems for scale-up	107
Readiness assessment	108
Licensing, purveyors and other forms of technical assistance	109
Independent replication	110
Appendix A: Strategies for searching the scientific evidence base	114
Appendix B: Targeting interventions on the basis of need	120
Universal support	120
Targeted selected interventions	121
Targeted indicated interventions	123
Appendix C: The EIF five-phase participatory model for confirming a science-based theory of change	124
Phase 1: Planning the exercise	124
Phase 2: Documenting the intervention's history	125
Phase 3: Gathering the evidence	126
Phase 4: Learning from each other	127
Phase 5: Confirming the theory of change	127
Appendix D: Statistical power, effect sizes and t-tests	130
Hypothesis testing: Accepting or rejecting the null hypothesis	130
Testing for significance with a t-test	131
Calculating the effect size	132
Power and sample size calculation	133
Appendix E: Common threats to internal validity	134
Appendix F: Readiness tool: Promoting Alternative Thinking Strategies (PATHS®)	138
Programme selection	138
Appendix G: General evaluation resources	140

## Summary 10 steps for evaluation success

Evaluation matters. It is important to know whether the services or interventions you provide are beneficial for the children and families who most need them. Evaluation methods often feel daunting, however, especially when they are unfamiliar or require technical expertise. It is therefore not surprising that those involved in the delivery of interventions often feel uncertain about how to evaluate them.

The EIF '10 steps for evaluation success' framework has been developed with these concerns in mind. Following on from the advice provided in *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them*,<sup>1</sup> this guide describes how evaluation evidence can be used to turn a good idea into an intervention that 'works', as well as developing quality assurance systems so that interventions remain effective when offered at scale. While not all 10 of these steps are necessary or practical for every children's service, we believe they are useful for making a good intervention even better, especially when followed in the order we set out here. These steps are also useful for more established interventions to reconfirm that their model is making a positive difference as it is implemented.

We believe that this guide is not only useful for those who develop interventions and services for children and their families, but also for those who commission and deliver these interventions on the ground.

- **Steps 1 through 7** are particularly useful for understanding the potential of early intervention activities to improve meaningful child development outcomes and establishing their evidence base.
- Steps 8 through 10 then describe how evaluation methods can be used to ensure that activities with established evidence remain effective as they are delivered in diverse contexts and populations.

<sup>&</sup>lt;sup>1</sup> See https://www.eif.org.uk/resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

#### **10 STEPS FOR EVALUATION SUCCESS**



Source: EIF

### Step 1: Confirm your theory of change

The best interventions are underpinned by clear theories that can explain why they are needed and how they will work. In this first step you will learn how to:

- develop a theory of change that not only explains what your intervention will do, but *why* your intervention is important for children's development
- consult the scientific evidence base to make sure that your theory of change is rooted in what is known about children's development
- use participatory methods to confirm your intervention's theory change and involve the scientific evidence as an active 'participant' in the process.

### Step 2: Develop your logic model

A logic model is a graphic representation of how an intervention's activities should support its intended child outcomes. In Step 2 you will learn how to:

- create a logic model that identifies how the intervention's resources, activities and participant outputs will support its intended child outcomes
- identify and explain the key assumptions underpinning the relationship between the intervention's resources, activities, outputs and outcomes specified in your logic model
- consider the external conditions necessary for your logic model to work.

## Step 3: Create a blueprint

An intervention blueprint identifies specific learning objectives for each of the intervention's core activities and then links them to the intervention's short-term outcomes. In Step 3, you will learn how to:

- create an intervention blueprint that links each intervention activity to a specific learning objective
- describe how each learning objective will lead to the intervention's intended short- and long-term child and family outcomes
- develop engaging intervention activities and learning materials that support a wide-range of learning styles and needs.

## Step 4: Conduct a feasibility study

A feasibility study (sometimes referred to as a process evaluation) tests whether the intervention can achieve its intended outputs. These outputs include the intervention's core activities, as well as its ability to recruit and retain its intended participants. In Step 4 you will learn how to:

- specify the intervention's core activities and identify the factors that support or interfere with their successful delivery
- use qualitative research methods to understand which factors contribute to the success of the intervention from the perspective of those delivering it
- use qualitative methods to understand how those receiving the intervention perceive the intervention's benefits and whether these perceptions are consistent with the intervention's original theory of change
- understand how best to recruit and retain participants
- develop systems for monitoring participant reach and core delivery targets
- apply methods for verifying user satisfaction
- track and document intervention costs.

## **Step 5: Pilot for outcomes**

Pilot studies are relatively inexpensive evaluations which investigate an intervention's potential for improving its intended child outcomes. Pilot studies are particularly useful for determining which measures are most appropriate for testing child outcomes, as well as how to best recruit and retain a sufficiently large and representative study sample. In Step 5, you will learn:

- the importance of validated measures and how to select and use them to measure preand post-intervention change
- methods for determining an adequate sample size based on the intervention's anticipated effects
- methods for recruiting and retaining participants from the intervention's target population
- analytic methods for determining whether changes in child outcomes are statistically significant
- how to interpret the findings from pilot studies and use them for designing more rigorous evaluations.

## Step 6: Test for efficacy

An efficacy study is a rigorous evaluation designed to determine if an intervention works under ideal circumstances. Efficacy studies do this through research designs that systematically reduce potential sources of study bias, so that causality can confidently be attributed to the intervention model. In Step 6 you will learn:

- · how to determine whether an intervention is ready for an efficacy study
- the ways in which potential sources of biases can 'threaten' the validity of a study's findings
- how a comparison group and methods such as random assignment can be used to reduce potential sources of study bias
- strategies for reducing all sources of potential bias throughout the duration of the efficacy study
- strategies for increasing the likelihood that the study will take place under ideal circumstances
- · how to interpret findings from efficacy studies
- what to do when a rigorously conducted efficacy study fails to observe any positive effect on a child outcome of interest.

## Step 7: Test for effectiveness

An effectiveness study is a rigorous evaluation designed to determine if the positive child outcomes observed in the efficacy study can be replicated in real-world circumstances. From the perspective of EIF, it is also useful if an effectiveness study (or previous efficacy study) can consider whether the intervention can be confidently associated with child benefits that are sustainable for a year or longer. In Step 7, we describe:

- · how effectiveness studies can be conducted in real-world circumstances
- · methods for measuring change for a year or longer
- how effectiveness studies can be used to understand for whom and under what circumstances the intervention has its greatest impact
- how to interpret disappointing findings observed in effectiveness studies.

## Step 8: Refine & monitor

Once an intervention has confirmed that it can provide benefits for children that are meaningful from a public health perspective and are sustainable within real-world settings, further testing is required to develop quality assurance systems to ensure that these benefits remain replicable. In Step 8, you will learn:

- how evaluation methods can be incorporated into the running of an intervention to monitor its quality on an ongoing basis
- · how to monitor child outcomes on an ongoing basis
- how monitoring systems can be used to determine when an intervention is appropriate for an individual child's needs or when referral to other services may be necessary
- how to rapid cycle evaluations and micro-trials can be used to test and refine an intervention's active ingredients
- evaluation methodologies for testing an intervention's workforce requirements.

### Step 9: Adapt & transport

As interventions are taken to scale, the diversity of the contexts in which they will be offered will naturally increase. When interventions are 'transported' into new cultures, substantial changes are particularly necessary. In this step, you will learn how evaluation methods can be used to:

- determine the extent to which intervention contents are relevant within new cultures and countries
- determine whether the intervention's intended child outcomes are upheld through ongoing piloting
- make decisions about the extent to which interventions developed in one country are needed and will 'fit' within the context of another.

### Step 10: Take to scale

While taking an intervention to scale is the last step in our 10 Steps framework, it does not mean that the intervention's evaluation journey is over. Instead, it signifies that evaluation cycles have been successfully integrated into the intervention's delivery systems to verify that it will remain effective when offered at scale. In Step 10, we describe all the quality assurance processes necessary for offering interventions at scale, including those which help local systems determine if they are ready to offer an intervention in a way that will ensure that it remains effective. In Step 10, you will learn:

- methods for assessing local system readiness
- the role of the intervention provider for informing system readiness
- the ways in which technical support can be used to inform system readiness and install interventions within local systems
- methods for offering and using technical support, including licensing, purveyors and independent intermediaries.

## Introducing the 10 Steps framework

## A commitment to children

So, you are committed to improving the lives of children and families. One of the best ways to do this is to provide services and interventions that make a positive difference. While you are pretty sure that what you offer hits this target, you want to be absolutely certain. This may be because you want to *know* that the support you provide is effective; or you want to *understand* how to make it more effective. You may also want to be able to *prove* to others that your intervention or service provides benefits for children that are both measurable and meaningful. In short, you want to be able to verify that what you do 'works' in a way that is convincing to yourself and to others.

This guide explains how evaluation designs can be used to understand *how* an intervention or service works, and, more importantly, *whether* it actually makes a positive difference for children and families. Our *10 steps for evaluation success* framework breaks down EIF's evidence standards into 10 successive and achievable steps to follow to ensure that interventions and services provide their maximum benefits.

#### **FIGURE 1**

EIF 10 steps for evaluation success framework



#### Source: EIF

Each one of these 10 steps provides valuable information about *why* the intervention is important, *what* it can achieve, *who* will benefit from it the most, *how* it can best be implemented, and *whether* it can provide benefits for children that are meaningful and sustainable. We believe that this advice is not only useful for those who develop interventions and services for children and their families, but also for those who make interventions available through commissioning and delivery on the ground.

## What works for child development

The Early Intervention Foundation (EIF) was set up in 2013 as an independent UK What Works centre to provide information about child and family interventions that 'work'. This means activities that have been confirmed, through rigorous evaluation, to have made **a meaningful and measurable difference** in the lives of children.

Outcome domains that have been proven through research to be particularly meaningful for children's development include children's physical, cognitive, behavioural and social/ emotional development (figure 2), as well as substantial reductions in child maltreatment, risky sexual behaviour and adolescent substance misuse (figure 3).



Studies consistently show that improved outcomes within each of these domains significantly reduces the risk of physical, mental and employment problems throughout children's development. EIF therefore prioritises interventions with evidence of improving child outcomes within at least one of these domains for entry in its Guidebook.<sup>2</sup> The Guidebook is a searchable, web-based resource which describes interventions in terms of their primary child and family outcomes, costs, implementation requirements and strength of evaluation evidence.

## The EIF evidence standards

Information about an intervention's evaluation evidence is obtained through a rigorous assessment process that compares an intervention's most robust evaluation evidence against EIF's evidence standards. The standards were developed to provide a straightforward assessment of intervention activities in terms of their strength of evidence – in other words, the extent to which we can be confident that the intervention actually 'works'. Assessing evidence strength is not only useful for understanding an intervention's potential for improving child outcomes, but also for comparing interventions to one another.

<sup>&</sup>lt;sup>2</sup> See https://guidebook.eif.org.uk/

The EIF standards are expressed in terms of five discreet ratings, starting with 'Not level 2' (or NL2) and then running numerically from level 2 to 4. A fifth category, 'No effect', is reserved for interventions that have been shown through rigorous evaluation (see level 3 below) not to have measurably improved any of the child or parent outcomes that it intended to improve. Each one of these standards corresponds with the steps in our 10 Steps framework, as we describe here.

#### EIF evidence standards Level 4 Level 3 NE Level 2 NL2 Not level 2 **Preliminary evidence** Efficacy Effectiveness Key elements of the logic The programme has evidence The programme has evidence The programme has evidence model are being confirmed of improving a child outcome from at least one rigorously from at least two rigorously conducted evaluations (RCT/ and verified in relation to from a study involving at least conducted RCT/QED practice and the underpinning 20 participants, representing demonstrating a statistically QED) demonstrating positive 60% of the sample, using impacts across populations scientific evidence. Testing significant positive impact on of impact is underway but validated instruments. at least one child outcome. and environments lasting a evidence of impact at level 2 year or longer. The evidence is not yet achieved. may include significant No effect adaptations to meet the needs The programme has evidence from at least one rigorously of different target populations. Return to verify and confirm the logic model. conducted RCT/QED that is also the most rigorous impact evaluation demonstrating no effect on <u>child outcomes</u> Source: FIF

#### NL2: Not level 2

**FIGURE 4** 

The EIF evidence standards begin with NL2, which is shorthand for describing interventions that do not yet meet the EIF level 2 criteria. This means the interventions have not verified their intended child outcomes with objective and validated measures, or with a sample that is sufficiently representative of their target population.

NL2 represents the foundational work underpinning all subsequent evaluations. This work involves all the activities described in Steps 1 to 4 of our 10 Steps framework, including *confirming* the intervention's core assumptions (also referred to as its theory of change) and *verifying* the intervention's feasibility. Verifying feasibility includes understanding whether the intervention can successfully recruit and retain its target audience and be delivered to a consistently high standard.

We believe these activities lay the groundwork for the intervention's future development, so will be returned to again and again as the intervention undergoes more rigorous evaluations. A primary aim of this guide is to describe how Steps 1 to 4 prepare interventions for more rigorous evaluations designed to test their causal impact.

#### Level 2: Preliminary evidence

Level 2 corresponds with Step 5 in our 10 Steps framework, representing the point at which an intervention undertakes a pilot evaluation that tests the intervention's potential for improving child outcomes. The information gained through a Step 5 pilot must therefore be valid, reliable and involve participants representing the intervention's target population. A Step 5 study does not need to be large or complex, but it should ideally be able to test changes in participants' behaviour before the participants have received the intervention and then again thereafter. Such designs are often referred to as pre/post pilot evaluations or time series designs.

Positive findings from pilot evaluations are often viewed as a promising indication that the intervention has the potential to provide benefits to children that are meaningful from a public health perspective. However, pilot study findings are not indicative of *whether* an intervention works. This is because pilot studies are not sufficiently rigorous to verify whether any observed positive changes can be confidently associated with the intervention model, or whether they would have occurred in the absence of the intervention. Nevertheless, positive findings from a level 2 evaluation are viewed by many to indicate that the intervention is ready for more robust evaluation designs that can more rigorously test these comparisons.

This guide provides advice on how Step 5 can be completed through a fit-for-purpose pre/ post study which meets the threshold criteria for an EIF level 2 rating. This advice covers methods for selecting validated instruments, determining an appropriate sample size, recruiting and retaining participants, collecting data and comparing change. We also provide advice on how findings from pilot studies can be used to inform decisions on whether to invest in further, more rigorous studies.

#### Level 3: Determining efficacy

The requirements of Step 6 are consistent with the EIF criteria for a level 3 rating. Completing Step 6 represents a significant step in an intervention's development, as it is the point at which judgments of causality can first be made about whether an intervention actually 'works' when it is delivered under ideal circumstances. These judgments are permitted through evaluation designs – also referred to as efficacy studies – that allow researchers to consider whether the outcomes observed are actually associated with the intervention model, as opposed to chance, business as usual or the characteristics of the intervention's participants. Step 6 is also the point at which secure judgments can be made about an intervention's potential lack of effect.

We view randomised controlled trials (RCTs) or equivalently rigorous designs (usually referred to as quasi-experimental designs or QEDs) as the most robust method of attributing causality to an intervention model. While this guide does not provide technical advice on how to design and conduct rigorous efficacy studies, **Step 6 does describe how findings from a well-designed efficacy trial can be used to understand an intervention's impact to inform its ongoing development.** 

#### Level 4: Understanding effectiveness

While positive findings from a single, rigorously conducted RCT or QED can provide insight as to whether an intervention has improved child outcomes on one occasion, it provides little information about whether the intervention will provide these benefits again. This is because the ideal circumstances of efficacy studies may not be replicable in the real world. In this respect, the first efficacy study does not fully differentiate between positive outcomes that may be associated with the intervention model from those associated with the circumstances under which the intervention was delivered.

Additional rigorous studies are therefore necessary to understand the external circumstances required to replicate positive outcomes with diverse populations in diverse situations. Replication studies (often referred to as effectiveness trials) are also useful for understanding how to adapt interventions, as well as determine the quality assurance mechanisms necessary to ensure that an intervention achieves its maximum benefits when delivered at scale.

Steps 7 to 10 of this guide provide information about how the general principles underpinning effectiveness trials can be used to monitor interventions as they are delivered, as well as the ways in which they can be used to revise and adapt interventions so that they are feasible with different populations and settings.

## How to use this guide

We recognise that many of these steps are not easy and that climbing all 10 will be impractical or unnecessary for many child and family services. Nevertheless, we believe that it is important for interventions that aim to support children's development to be rigorously evaluated – both as a means of verifying their impact and as a form of quality assuring their delivery. We therefore believe that climbing these steps is important for increasing the effectiveness of activities aimed at improving child and family outcomes, and that costly mistakes can be avoided when they are followed in our proposed order.

This guide provides advice on how each of the 10 steps can be completed in a way that is practical, efficient and useful. It also provides a clear rationale as to why following these steps in a specific order is useful for avoiding many of the pitfalls that often compromise the integrity of an evaluation's findings. This advice is provided in 10 brief chapters describing each of the steps in terms of:

- why the step is important for understanding an intervention's impact and improving its quality
- how the step can be accomplished in a way that is achievable and efficient, through the use of case study examples
- what successful completion of the step looks like, through checklists that can be used by developers and commissioners
- resources for finding further information on how to complete each step.<sup>3</sup>

Additional information for completing each step (including online templates and brief, 'how-to' guides) is also provided in the appendices and through links to other EIF resources.

<sup>&</sup>lt;sup>3</sup> We have drawn on some of the sources listed in the further resources sections in producing the guidance for each of the 10 steps.

## Step 1 Confirm your theory of change

#### 1. CONFIRM YOUR THEORY OF CHANGE

#### The best interventions are underpinned by clear theories that can explain why they are needed and how they will work.

In this first step you will learn how to:

- develop a theory of change that not only explains *what* your intervention will do, but *why* your intervention is important for children's development and *who* will benefit from it the most
- consult the scientific evidence base to make sure that your theory of change is rooted in what is known about children's development
- use participatory methods to confirm your intervention's theory of change and involve the scientific evidence, as represented by a researcher, as an active 'participant'.

# There is nothing quite so practical as a science-based theory

Kurt Lewin, regarded by many to be the founder of social psychology, once observed that 'there is nothing quite so practical as a good theory'. By this, Lewin meant that good theories provide a practical basis for understanding how various activities and conditions are causally linked to positive changes in human behaviour. However, his statement also more wryly implied that good theories were hard to come by. This view was based on Lewin's own observation that intervention theories were often unspecified or based on assumptions that were either untested or unfounded.

Fundamentally, most child and family interventions are informed by implicit or explicit theories that causally link intervention activities to improved child and caregiver outcomes. These theories, also referred to as *theories of change* (ToC), are based on assumptions about children's development and the ways in which families, communities and services support various developmental processes. However, consistent with Lewin's original observations, these theories are often informed by assumptions or conceptions about children's development that are not fully supported by scientific evidence.

It goes without saying that interventions that are rooted in findings from research have a greater chance of providing benefits to children and families. Understanding this evidence is therefore essential for ensuring that interventions achieve their maximum impact. A clear understanding of what is known through child development research is also useful for confirming the intervention's primary outcomes, who the intervention is for and what the content will be.

Time and again, evaluations have failed to verify an intervention's benefits either because they have investigated the wrong outcomes or because the intervention's core assumptions were essentially wrong. This chapter therefore provides clear advice on how to avoid these failures by using scientific evidence to confirm a theory of change that is both practical and firmly rooted in what is known about children's development. This advice includes methods for consulting the scientific literature on a regular basis and strategies for using this information to make science-based decisions about the intervention's content, target population and primary child outcomes. In our view, a science-based ToC is fundamental to interventions that intend to improve outcomes for children. This means that the science underpinning interventions should be continually revisited throughout an intervention's development.

# Using science to specify why an intervention is important, necessary and will add value

It is not uncommon for people to use the terms 'theory of change' and 'logic model' interchangeably. This is because both terms explain processes that create a theoretical link between an intervention's activities and its intended short- and long-term outcomes. For example, Wikipedia defines a theory of change as identifying the *causal linkages in an initiative*, *i.e.*, *its shorter-term*, *intermediate*, *and longer-term outcomes*. A logic model, by contrast, is described as a hypothesised description of the chain of causes and effects leading to an outcome of interest.

Thus, ToCs and logic models are both useful for identifying *what* an intervention is trying to achieve. However, they differ in one important way: while logic models (which we describe in Step 2) are primarily concerned with *how* an intervention will achieve its primary outcomes, ToCs identify *why* these outcomes are important in the first place.

We believe that being able to specify *why* the intervention's outcome is important should be the core purpose of any ToC. In fact, we believe that ToCs should be able to answer three related 'why' questions when articulating the fundamental purpose of the intervention.

- What is the intervention's primary intended outcome and more fundamentally, why is the outcome important? When it comes to interventions for children and parents, we feel that it is particularly important that this question should address why the intervention is important from the perspective of children's development. The answer to these first two questions will then inform all further questions regarding what the intervention will do.
- Why is the intervention needed? Answers to this question should consider the developmental processes the intervention is addressing, as well as individual differences in the development of these processes. The answer to this question will then inform further specification of who the intervention is for.
- Why will the intervention add value over what is currently available for children and families? Another way to phrase this is to ask what difference the intervention will make. The answer to this question should consider how the intervention will contribute to children's development in comparison with the services which are currently available. Answering this question should also help developers further specify what the intervention activities will be and how much of the intervention children and families will receive.

Intervention developers frequently rely on their own experiences to answer these three 'why' questions. After all, their beliefs about why the intervention was needed likely contributed to its initial development. However, we believe that explanations about why an intervention is important and necessary should also be informed by what is known about children's development and the ways in which various intervention models can add value.

Many of the interventions included in the EIF Guidebook can successfully answer these three 'why' questions by explicitly linking their ToCs to what is known about children's development from the scientific evidence base. This research includes knowledge gained from experimental designs (for example, RCTs and QEDs as described in the introduction), observational studies and longitudinal research that considers the ways in which risk and protective factors impact children's development over time.

Examples of science-based theories informed by longitudinal research include attachment theory, social learning theory, ecological theory, the Collaborative for Academic, Social and Emotional Learning (CASEL) model, and the family stress model. Information about these science-based frameworks and the evidence underpinning them can be found in the further resources section at the end of this chapter. In case study 1, we also use the history of coercion theory to exemplify the ways in which scientific findings have been used to develop and refine a science-based ToC.

#### Case study 1: Building a science-based theory of change

All theories of child development are fundamentally theories of change. Coercion theory is an example of a science-based ToC developed by Gerald Patterson and his colleagues at the Oregon Social Learning Center (OSLC), informed by decades of longitudinal research observing child and adult interaction. Patterson and his colleagues used these findings to develop and test theories of childhood aggression and coercive family processes, as well as develop evidencebased interventions for reducing antisocial behaviour and improving family functioning.

Patterson's work began in the early 1960s through observations of children interacting with their peers in preschool settings. Patterson and his colleagues noted that all preschool children engaged in relatively high levels of aggression and that aggressive behaviour increased when it was reinforced by attention, or verbally aggressive responses from adults. In particular, Patterson observed that adults who were 'coerced' into giving into an aggressive child's demands or responded to the child in a verbally or physically aggressive manner, inadvertently encouraged the child to continue to behave aggressively. In addition, adults who responded to their children aggressively often did not view their behaviour as contributing to their children's aggression, or as otherwise inappropriate. By contrast, preschool children whose caregivers or teachers either ignored aggressive behaviours or sanctioned them through non-aggressive means (by removing a privilege or enforcing a time-out) engaged in less-aggressive behaviour as they developed.

In a series of subsequent studies, Patterson and the OSLC team observed that high levels of aggressive child behaviours in preschool predicted continued aggressive behaviours in primary school. Children who were more aggressive in school, in turn, were more likely to be rejected by their non-aggressive peers and struggle academically. Once in secondary school, aggressive children were then at further risk of affiliating with other aggressive children, who in turn reinforced each other's deviant and risky adolescent behaviours. Examples of common behaviours encouraged by deviant peer groups during adolescence included drug and alcohol misuse, youth violence and criminal offending.

Once these longitudinal associations were confirmed, Patterson and his colleagues had the building blocks of a scientifically supported ToC of childhood aggression (figure 1.1 below).

Patterson's theories of aggression and coercive family interactions alsoprovided the starting point for answering the three important 'why' questions identified at the beginning of this chapter.

• Why was the intervention outcome important? Patterson's team confirmed that early childhood aggression was an important outcome to target. Not only did it predict later childhood aggression, it was associated with a wide variety of other negative child outcomes as children grew older.

- Why was the intervention needed? Certain caregiver responses inadvertently reinforced and increased the frequency of childhood aggression. Interventions were therefore necessary to help parents identify ineffective parental responses to childhood aggression and replace them with more effective and appropriate responses.
- Why would the intervention add value over what was currently available for children and families? At the time coercion theory was developed, parenting interventions were a relatively new concept and few programmes provided advice for reducing aggressive child behaviour. Thus, Patterson's team could safely assume that any parenting advice had the potential to add value over what was currently available.

#### **FIGURE 1.1**

A theory of childhood aggression supported by scientific observations of adult child interaction



By using the findings from longitudinal studies to answer these three questions, Patterson and the OSLC team were then able to confirm their hypothesised answers to two additional 'what' questions – *what* the intervention would do (in other words, what were its change mechanisms) and *what* the intervention would achieve (in other words, what were its specific short- and long-term outcomes) – as illustrated in figure 1.2.

#### FIGURE 1.2

A theory of change for improving children's behaviour through intervention's targeting coercive parenting behaviours



Patterson and his colleagues were thus well on their way to developing an intervention model that was fully informed by scientific evidence. However, two more important questions still needed to be answered: *who* would benefit from these interventions, and *how much* of the intervention was needed.

The common-sense answers to these questions were that all caregivers would want and immediately understand simple advice about coercive family processes. In other words, caregivers just needed to be told that coercive family interactions contributed to aggressive child behaviours and they would want and know how to stop engaging in them immediately.

However, these assumptions proved to be wrong, and determining who would benefit from parenting interventions and how much advice was need ended up requiring years of further research. Fundamentally, not all caregivers required advice on how to respond to aggressive child behaviour and families varied considerably in their ability to accept and act on it.

- Many caregivers already responded appropriately to their children's aggressive behaviour, and therefore advice about coercive family interactions provided no measurable value.
- Families were less likely to understand or act on advice that was not specific to their child's age or family circumstances.
- Some parents actively resisted the advice, believing it to be unhelpful or unnecessary.
- Some parents were struggling with other challenges that made it difficult for them to benefit from the advice.

In the end, Patterson and his colleagues confirmed that these important variations in family characteristics significantly influenced the effectiveness of interventions that aimed to reverse parent-child coercive family cycles. They also suggested that interventions would be more effective if they were specifically needed and included components that helped parents overcome their initial resistance or other challenges that may interfere with their ability to benefit. In other words, interventions needed to target families on the basis of children's age and family need, and also needed to be sufficiently intensive to meet the families' needs.

In order to determine family need, Patterson and his team returned to their findings (and the scientific literature more broadly) to identify the risk factors associated with coercive family interactions. This exercise further confirmed that some parents were at far greater risk of experiencing difficulties with aggressive child behaviour than were others. Risk factors that significantly increased the likelihood of aggressive child behaviour included low income, single parenthood, high levels of family stress, a history of family violence and parental incarceration. The OSLC team also observed that children whose parents had offended were at particular risk of behaving aggressively.

This knowledge allowed Patterson's team to more deeply specify their ToC by identifying who would most benefit from the intervention (figure 1.3). Knowing that those who required the intervention were also highly vulnerable additionally helped the developers consider how much of the intervention was required, although ongoing testing was still necessary for confirming the specifics and duration of the content.

#### FIGURE 1.3

A theory of change based on coercion theory Linking why the intervention is necessary to who will receive it, what they will receive and what the intervention will hopefully achieve



There are now countless family interventions that are based in Patterson's original theory of change involving coercion theory. Many were developed by members of the OSLC team, although developers throughout the world external to the OSLC have also developed effective interventions based on coercion theory. These interventions range from relatively light-touch advice offered through books, videos and online resources, to highly intensive family interventions lasting a year or longer. These interventions also vary in terms of their impact, with findings from their ongoing evaluations suggesting that while many reduce family coercion and child aggression in the short-term, more work is required to understand the extent to which they prevent conduct problems from occurring in the long term. Interventions are also continually being refined and developed to ensure the interventions are acceptable and used by the families who most need them.

# How to develop a science-based intervention when you are not a scientist

Although Patterson and the research team described in case study 1 were developmental psychologists with significant training in research design and statistical methods, being a scientist is not a prerequisite for developing an effective, science-based intervention. In fact, many effective interventions have been developed by professionals who were not scientists – at least not initially. The majority of these developers nevertheless consulted the scientific literature to better understand why their intervention was needed, what it should do and who would most benefit from it.

We recommend that intervention developers and those designing children's services similarly consult the scientific evidence base (see appendix A for advice on how to do this) to answer the seven essential questions we describe here, to ensure that their intervention's ToC is underpinned by sound scientific evidence. *We also recommend that these questions be answered in the order we provide below, even though this will not be their final order when the intervention's ToC is ultimately confirmed*.

#### 1. What is the intervention's primary intended child outcome?

This might appear to be a fairly straightforward question, but it is not uncommon for interventions to target outcomes that are either overly ambitious or poorly specified (for example, all children will 'flourish' or children will 'thrive'). While such outcomes may be appropriate for an organisational mission statement or strapline, they are not sufficiently specific for an intervention's ToC.

We therefore strongly recommend that those designing interventions and services restrict themselves to no more than one or two primary outcomes that are fairly specific, especially when initially confirming their intervention's ToC. The outcomes should fall within one of the seven developmental domains described in the introduction, and should identify short- and long-term improvements that are consistent with what is known about children's development.

We should also stress that interventions aiming to support children's development identify specific child outcomes, such as those described in the introductory chapter, even if the intervention is delivered to caregivers or teachers. Evaluations of early interventions repeatedly show that it is not sufficient to assume that children will automatically benefit from interventions that may be beneficial for their caregivers. A ToC should therefore be able to specify child outcomes that are developmentally appropriate, even when their caregivers are the primary recipients.

#### 2. Why is the primary intended child outcome important?

The answer to this question should consider the importance of the outcome from the perspective of children's development. For example, the OSLC team in case study 1 identified aggressive child behaviour as an important child outcome on the basis of longitudinal research findings showing that aggressive child behaviour in preschool increased the likelihood of aggressive behaviour in primary and secondary school. Intervention developers should similarly be able to justify why their intervention's primary intended outcome is important for children's long-term development by mapping short- and long-term outcomes to scientifically supported theories of child development.

Ideally, short-term outcomes should be consistent with what is known about normal child development at a specific age. For example, the ToC in figure 1.3 identifies age-appropriate child behaviour as a short-term goal. A variety of resources (as described in appendix A and the end of this chapter) provide comprehensive and up-to-date information about age-appropriate milestones in children's development. We also provide some information about key milestones in the EIF Foundations for Life<sup>4</sup> and Key competencies in early cognitive development<sup>5</sup> reports.

#### 3. Why is the intervention necessary?

Most interventions are developed to fulfil a need. A science-based ToC should therefore be able to justify the need for an intervention from the perspective of what is known about the developmental processes that contribute to the intervention's intended outcomes. For example, many ToCs identify children's language as an important intervention outcome to target, despite the fact that the majority of children learn how to talk with seemingly little instruction. Nevertheless, children differ in their ability to talk and these differences are scientifically linked to developmental processes occurring between the child and his or her environment that support the acquisition of language. As described in case study 1, a science-based ToC should be able to specify how the intervention will address developmentally relevant processes by 1) identifying the specific processes it is trying to improve and 2) justifying why these developmental processes are important on the basis of the scientific evidence.

#### 4. Why will the intervention add value?

In order for an intervention to have an impact, it needs to provide measurable value over what is currently available. In other words, the intervention needs to fill a gap. Quite often, gaps are identified on the basis of needs identified by families living in the community or by practitioners who work with these families. Addressing these needs is an excellent reason for setting up an intervention and determining how it could add value over what is currently available.

Scientific findings also provide an excellent way of identifying the need for an intervention. Findings from systematic reviews not only identify *where* gaps exist, but *why* gaps exist. For example, gaps in the availability of effective interventions may exist because interventions have not yet been developed or rigorously tested. Alternatively, interventions may have been rigorously tested, but have not yet been shown to be effective.

Understanding why evidence gaps exist is essential so that developers do not inadvertently replicate the circumstances that may have led to the gaps in the first place. A well-developed ToC should therefore not only articulate why the intervention might add value, but how it might avoid the pitfalls that have caused interventions to fail in the past.

<sup>&</sup>lt;sup>4</sup> See: https://www.eif.org.uk/report/foundations-for-life-what-works-to-support-parent-child-interaction-in-the-early-years

<sup>&</sup>lt;sup>5</sup> See: https://www.eif.org.uk/report/key-competencies-in-early-cognitive-development-things-people-numbers-and-words

#### 5. Who is the intervention for?

While developers often assume their intervention will provide advice that is useful for everyone, in reality this is rarely the case. The impact of an intervention will therefore be the greatest if it is delivered to the children and families who most need it. For many interventions, there will be families who may be too vulnerable to benefit. This is likely to be particularly true of families where there are multiple, complex issues, including child protection concerns. Likewise, there will always be children and caregivers who do not require any additional support.

Developers should therefore use scientific findings to determine who is most likely to benefit from the intervention. Figure 1.4 provides a helpful starting point for making these distinctions. The pyramid on the left is useful for considering *who* the participants will be, while the inverted pyramid on the right is helpful for considering *how much* of the intervention they will receive.

#### FIGURE 1.4



Source: EIF

These pyramids are informed by public health frameworks which commonly identify population needs within three categories – universal, targeted selected and targeted indicated – as defined in the box below. Further information about offering interventions at the universal, targeted selected and targeted indicated levels is provided in appendix B.

### A continuum of population needs

**Universal** interventions and services are those that are made available to all children and families with the aim of preventing problems from occurring in the first place.

**Targeted selected** interventions are those offered to children or families based on demographic risks, such as low family income, single parenthood or adolescent parenthood. Although children growing up in these circumstances may not be experiencing any specific problems, interventions offered to children or families because of these risks have the potential to prevent more serious problems from arising.

**Targeted indicated** interventions are for children or families identified or assessed by practitioners as having a specific or diagnosed problem which requires more intensive support. In this case, early intervention has the potential to address these problems and stop them getting worse.

#### 6. What will the intervention do?

No ToC is complete without specifying what the intervention will do. However, the initial confirmation of the ToC does not need to do this in any great detail, as this is the primary objective of Steps 2 and 3. Nevertheless, it is useful for developers to consult the scientific evidence base to find out more about the relative impact of various intervention activities to determine the content of their own interventions. Content-related factors found to contribute to an intervention's impact include:

- the intervention's learning activities and materials
- the specificity of the intervention or service's content is it specific to the child's age and the needs of the target population?
- opportunities for participants to practise and gain feedback
- the frequency of the intervention
- the format (for example, group vs individual)
- the settings in which the intervention has been successful
- the characteristics of the individuals delivering the intervention, including their professional qualifications
- the intervention's training and quality assurance systems.

As described in appendix A, many of the details of effective interventions are provided in their published evaluations.

#### 7. How much of the intervention is required?

Decisions about how much of the intervention is required should also consider the needs of the target population. As described previously and in appendix B, families with more complex needs often require more intensive interventions, although the details about the frequency and intensity of an intervention are rarely confirmed until Steps 2 and 3. At the initial stages of confirming a ToC, it is only necessary to identify whether the intervention will be high or low intensity and suitable for meeting the needs of the target population. Further information about intervention intensity will need to be determined through ongoing testing and evaluation, as described in subsequent steps.

# Confirming a science-based theory of change through participatory methods

While many are likely to see the value of a science-based theory of change, many will also question whether it is necessary for interventions that are developed and in operation. While this question is understandable, the answer remains clear: **science-based theories of change are highly relevant for established interventions, although the methods used for specifying them will be different to those used by developers who are starting from scratch.** 

We recommend that the providers of established interventions use participatory methods as we describe in appendix C (sometimes referred to as 'co-creation' or 'co-production') to confirm the intervention's ToC. Participatory methods are useful because they ensure that the views of all the stakeholders involved in the intervention's design, set-up and delivery are considered and taken forward when developing the intervention. These stakeholders may include:

- The intervention's developer, who may be an individual or team of individuals who developed the intervention's initial content and quality assurance systems.
- A host organisation, also referred to as a 'purveyor' (see Step 10) who is responsible for marketing the intervention to multiple sites. Purveyor responsibilities include the coordination of the intervention's training activities, as well as the provision of implementation and quality assurance support.
- Local authority commissioners, who are responsible for purchasing the intervention and ensuring that it is relevant for the needs of the children and families in their communities.
- The host agency, who is responsible for the intervention's day-to-day delivery by practitioners and service managers.
- Evaluators, who may be independent, or work directly for the purveyor, commissioner or host agency. Although evaluators may be hired to test an intervention's efficacy, more often than not they are hired to determine whether the intervention is feasible and is reaching the families it was intended to reach (we describe this in greater detail in Step 4).
- If the intervention has been running for some time, recipients (including teachers, caregivers or children) may also be included in the ToC exercise, although this is optional if the sole purpose is to verify the ToC. Involving participants is essential, however, when conducting a comprehensive feasibility study, as we describe in Step 4. Further advice on how to involve recipients in ToC exercises is also provided in appendix C.

All of these stakeholders will likely have views about why the intervention is important and necessary, who the intervention is ideally for and its potential for providing value over other available activities. However, each of these stakeholder's views may or may not be consistent with the views of other stakeholders or the scientific evidence base. A participatory ToC exercise is therefore useful for helping the various stakeholders gain a shared understanding about what the intervention is trying to achieve and the extent to which its goals and activities are supported by scientific evidence.

A wide variety of methods exist for conducting a participatory ToC exercise, although relatively few include strategies for involving the scientific evidence-base as an additional stakeholder. In appendix C, we introduce the EIF five-phase participatory model which offers a step-by-step approach for including scientific evidence base as an active ToC 'participant', as represented by a researcher. Our model also provides strategies and templates for systematically answering the seven essential ToC questions identified in this chapter to produce a dynamic ToC document that can be updated on a regular basis as developers and other intervention stakeholders continue to take their intervention forward.

## Preparing for the next step

Confirming a science-based theory of change is a major milestone in the development of any intervention, whether it occurs during the initial phases of the intervention's development or at later points in its life cycle. There is no question that a science-based theory of change which answers the seven questions identified in this chapter will provide a solid foundation for further specifying the intervention's logic model and core activities, as described in the following chapters. A well-articulated theory of change which provides clear links to the scientific evidence base is also a powerful tool for establishing credibility with commissioners and others who will be involved in the intervention's set-up and delivery. **We therefore strongly recommend that intervention developers view the theory of change process as the cornerstone of their intervention's development and implement processes to ensure that it is revisited and revised on a regular basis.** 

## Step 1 checklist

- The intervention has developed a ToC which identifies short- and long-term goals that are meaningful for children's development. This is justified by clear and rational links to the scientific literature.
- This is a clear explanation as to why the intervention is necessary for children's development from the perspective of the scientific evidence base.
- There is a clear explanation as to how the intervention will provide value over current provision. This explanation is rooted in the scientific evidence and a clear case has been made as to why the intervention may be better than what is currently available.
- The intervention's ToC has clearly articulated who the intervention is for, both in terms of its recipients' level of need and the target age of the children.
- There is a clear reason why the target population might need and benefit from the intervention and this reason is clearly and rationally supported by scientific evidence.
- There is a clear explanation about what the intervention will do. The explanation is supported by scientific evidence showing that these activities are appropriate for the target population and the intervention's intended outcomes.
- The ToC has specified the intensity of the intervention and there is evidence showing that this intensity is sufficient for the needs of the target population and improving the intervention's primary intended outcomes.

#### **Further resources**

#### General

Asmussen, K. (2011). The evidence-based parenting practitioner's handbook. Routledge.

#### Developing a theory of change

- Ghate, D. (2018). Developing theories of change for social programmes: co-producing evidence-supported quality improvement. *Palgrave Communications*, 4(1), 90.
- Lewin, K. (1944). Constructs in psychology and psychological ecology. *University of Iowa Studies in Child Welfare*, 20, 1–29.
- Van de Ven, A. H. (1989). Nothing is quite so practical as a good theory. *Academy of management Review*, 14(4), 486–489.
- Weiss, C. H. (1998). Methods for studying programs and policies. Prentice Hall.
- Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. *New approaches to evaluating community initiatives: Concepts, methods, and contexts, 1,* 65–92.

#### Participatory evaluation

Aubel, J. (1999). Participatory program evaluation manual. Claverton, MD: The Child Survival Technical Support Project. https://www.alnap.org/system/files/content/resource/files/main/participatory-program-evaluationmanual.pdf

Patton, M. Q. (1997). Utlization-focused evaluation: The new century text, 3rd edn, Sage Publications.

#### Examples of science-based theories of change

#### Attachment theory

Bowlby, E. J. M. (2008). Attachment: Volume One of the Attachment and Loss Trilogy. Random House.

Ainsworth, M. D. S. (1978). The Bowlby-Ainsworth attachment theory. *Behavioral and brain sciences*, 1(3), 436–438.

Ecological systems model

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes.

Sameroff, A. (2009). The transactional model. American Psychological Association.

#### Family stress model

Conger, K. J., Rueter, M. A., & Conger, R. D. (2000). The role of economic pressure in the lives of parents and their adolescents: The Family Stress Model. In L. J. Crockett & R. K. Silbereisen (eds), *Negotiating adolescence in times of social change* (pp. 201–223). Cambridge University Press.

Social learning theory (also referred to as coercion theory)

Dishion, T. J., & Patterson, G. R. (2015). The development and ecology of antisocial behavior in children and adolescents. *Developmental psychopathology: Volume three: Risk, disorder, and adaptation*, 503–541.

#### Social and emotional learning

Graczyk, P., Matjasko, J., Weissberg, R., Greenberg, M., Elias, M., & Zins, J. (2000). The role of the Collaborative to Advance Social and Emotional Learning (CASEL) in supporting the implementation of quality school-based prevention programs. *Journal of Educational and Psychological Consultation*, *11*(1), 3–6.

#### Scaffolding theory

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2), 89–100.

#### **Case study 1 references**

Patterson, G. R., Littman, R. A., & Bricker, W. (1967). Assertive behavior in children: A step toward a theory of aggression. *Monographs of the society for research in child development*, 32(5), iii–43.

Patterson, G. R. (1976). The aggressive child: Victim and architect of a coercive system. *Behavior modification and families*, 1, 267–316.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (2017). A developmental perspective on antisocial behavior. In *Developmental and Life-course Criminological Theories* (pp. 29–35). Routledge.

Scott, S., & Gardner, F. (2008). Parenting programs. Rutter's child and adolescent psychiatry, 5, 1046–1061.

# Step 2 Develop your logic model



## A logic model is a graphic representation of how an intervention's activities should support its intended child outcomes.

In Step 2 you will learn how to:

- create a logic model that identifies how the intervention's resources, activities and participant outputs will support its intended child outcomes
- identify and explain the key assumptions underpinning the relationship between the intervention's resources, activities, outputs and outcomes specified in your logic model
- consider the external conditions necessary for your intervention or service to work as your logic model describes.

## Turn your theory of change into a theory of action

As described in Step 1, ToCs and logic models serve two separate, but important purposes. A ToC specifies *why* an intervention's outcomes are important for children's development. A logic model graphically represents *what* the intervention will do to achieve these outcomes. Some therefore describe logic models as *theories of action* or action maps. A logic model specifies action by describing in more detail how the resources required to implement an intervention will lead to specific outputs, which should in turn contribute to the short- and long-term outcomes initially identified in the intervention's ToC.

A wide variety of resources already provide guidance on building a logic model and there is no single correct method. However, we favour methods, such as those proposed by University of Wisconsin Extension (UWEX),<sup>6</sup> that encourage intervention stakeholders to identify the ways in which external circumstances might also impact an intervention's success. External circumstances include the features of the wider system in which the intervention will operate, as well as the availability of a suitably qualified workforce to deliver it. Once these features have been specified, logic models not only provide a useful starting point for designing evaluations, they are also a powerful tool for monitoring intervention quality as they are implemented.

<sup>&</sup>lt;sup>6</sup> See: https://fyi.extension.wisc.edu/programdevelopment/files/2016/03/lmcourseall.pdf

## The key components of a well-specified logic model

Figure 2.1 sets out the basic elements of the UWEX logic model. These include the intervention's inputs (resources), outputs (activities), and short-, medium- and long-term outcomes. Developers can start building their model with its intended outcomes (from right to left) or with the available resources (from left to right). Interventions that have already specified their outcomes through their ToC are in a good position to start building their logic model from the right and then work their way backwards. However, we must stress that there is no right or wrong order when confirming a logic model. Although the end result will be essentially linear, the process of confirming a logic model is typically iterative.

#### Starting with outcomes

As figure 2.1 indicates, developers can begin building their logic model by identifying short-, medium- and long-term outcomes. Some of this work will already have been done when confirming the ToC, although logic models often specify outcomes in greater detail.

- **Short-term outcomes**, also described as *proximal outcomes*, are the immediate result of the intervention. For example, the ToC in figure 1.3 in Step 1 identifies changes in parenting behaviours (also described as the intervention's learning goals) as the most immediate and proximal short-term outcomes of the intervention.
- **Medium-term outcomes** are identified as more *distal* intervention outcomes. For example, the ToC in figure 1.3 of Step 1 identifies changes in children's behaviour as a more distal goal that will hopefully buffer children from negative outcomes over time.
- Longer-term outcomes describe the wider benefits of the intervention for children's overall development. These will likely already have been considered when the ToC was confirmed. In the example in figure 1.3, a reduced likelihood of antisocial behaviour is identified as the longest-term and most distal benefit of the intervention.

It should be kept in mind that the short-, medium- and long-term outcomes do not need to be specified in any great detail when the logic model is first considered. These outcomes will be revisited in Step 3, when the intervention's blueprint is developed, and then again in all subsequent steps when specific child outcomes are evaluated. Developers should, however, continually consider the feasibility of their outcomes by asking themselves the following four questions on an ongoing basis.

- Are the outcomes important? As mentioned in the introductory chapter, not all child outcomes are equally meaningful to children's development. Developers will need to continually ask themselves if the outcomes they are attempting to achieve are sufficiently meaningful from a public health standpoint to justify the time and expense of the intervention.
- 2. Are the outcomes reasonable? Fundamentally, logic models are built around plausible 'if-then' assumptions. Developers therefore continually need to ask themselves if their 'if-then' assumptions follow in the order they assume. For example, is it reasonable to assume that a proximal goal of improved parenting could lead to a more distal goal involving child behaviour? Or, could it be the other way around? Could short-term improvements in skills, in fact, lead to longer-term increases in self-confidence?
- **3.** Are the outcomes realistic? In other words, is it reasonable to assume that the intervention's activities will be sufficient to bring about the intervention's intended short-term outcomes? Is it also sensible to assume that the short-term changes are sufficient to bring about the intervention's intended medium- and longer-term changes? Examples of realistic outcomes are provided in case study 2.

4. What is the potential for unintended or negative outcomes? Even the most carefully designed interventions have the potential to bring about unintended, or negative consequences. For example, the Cambridge-Somerville Youth project involving mentoring and home visiting support to delinquent youths resulted in a greater number of arrests and early deaths in comparison to children who did not receive the treatment. Although such negative and severe results are rare, developers should continually ask themselves whether there are aspects of their model that may be misleading or unhelpful for the processes they are trying to improve.

#### FIGURE 2.1

The key elements of a well-specified logic model



#### **Identifying outputs**

Once the short-, medium- and long-term outcomes have been confirmed, developers can start to consider the specific outputs required to achieve them. The UWEX model identifies outputs as the intervention's key deliverables – in other words, what the intervention will do and who it will reach.

As we describe in Step 4, it is vital that an intervention's activities and recruitment targets are fully specified and tested. Common sense dictates that interventions will not deliver results if their core activities do not take place or if children and caregivers do not attend the intervention. The logic model is the time when these details should first be considered, so that their feasibility can be confirmed in subsequent steps.

Figure 2.2 illustrates how these outputs have been specified in our case study 2 example involving the Generation Parent Management Training Oregon (GenerationPMTO) programme. In this instance, PMTO labelled its intervention activities as 'implementation' and their recruitment targets as 'outputs'. What is particularly helpful about this example is that the developers identified a series of delivery and participation milestones that needed to be achieved in order for the intervention to be delivered successfully. This knowledge then allowed the evaluators to systematically examine the barriers to successful delivery and recruitment, including factors independent of the intervention's delivery.

It is worth keeping in mind that many of an intervention's output details will not be fully confirmed until Steps 3 and 4, once the intervention's blueprint has been specified and some feasibility testing has taken place. Nevertheless, making an educated guess about what the key delivery outputs might be is useful for determining the resources required to deliver the intervention, as well as understanding the contextual factors that might influence its success.

## Case study 2: Using a logic model to specify the core outputs of an intervention

Generation Parent Management Training Oregon (GenerationPMTO) is one of several evidencebased family interventions developed by Gerry Patterson and his team at OLSC to reduce problematic child behaviour by reversing coercive family cycles (see case study 1). This case study describes how the detailed specification of the intervention's delivery outputs in its logic model (see figure 2.2) helped users consider the feasibility of the intervention for increasing the placement stability of looked-after children with serious emotional disturbance.

#### FIGURE 2.2

Key elements of the GenerationPMTO logic model

Inputs	Outputs		Outcomes - Impact		
	Activities	Participation	Short term	Medium term	Long term
• Workforce • Venue • Costs	<ul> <li>Child assessment for SED</li> <li>Parent engagement early in case</li> <li>Family assessment</li> <li>Parenting training/PMTO</li> <li>Parent-child visits</li> <li>Concrete services</li> <li>Robust referrals &amp; service coordination</li> </ul>	<ul> <li>#/% of children assessed for SED</li> <li>#/% of parents who have comprehensive assessment</li> <li>#/% of weekly parent visits</li> <li>#/% of parents who complete PMTO services</li> <li>#/% of practitioners PMTO certified</li> <li>#/% of practitioners rated at adequate or better on fidelity score</li> </ul>	<ul> <li>Increased positive parenting behaviours</li> <li>Reduced coercive parenting behaviours</li> <li>Increased use of community resources &amp; supports</li> <li>Reduced parent mental health &amp; substance abuse problems</li> <li>Improved readiness for reunification</li> <li>Reduced child's problematic behaviour</li> <li>Improved child's pro-social skills</li> <li>Improved child functioning</li> </ul>	<ul> <li>Increased reunification</li> <li>Reduced long-term foster care</li> <li>Increased stable permanency</li> <li>Increased child safety</li> </ul>	<ul> <li>Improved school achievement</li> <li>Reduced offending</li> <li>Improved emotional security</li> </ul>

Source: Derived from Akin, B. A., Bryson, S. A., Testa, M. F., Blase, K. A., McDonald, T., & Melz, H. (2013). Usability testing, initial implementation, and formative evaluation of an evidence-based intervention: Lessons from a demonstration project to reduce long-term foster care. *Evaluation and Program Planning*, *41*, 19–30.

#### Situation

The Kansas Intensive Permanency Project (KIPP: a state-wide partnership between the Kansas Department for Social Services, the Kansas Department of Children and Families, and four fostering agencies) was experiencing high levels of placement instability among looked-after children placed in long-term foster care. Data mining techniques observed that longer-term placements and placement instability was particularly high among looked-after children prone to behavioural problems related to serious emotional disturbances (SED).

#### **Priorities**

KIPP identified the need to improve the social and emotional outcomes of these children as a short-term service priority, with the aim of increasing their placement stability or improving family reunification rates. GenerationPMTO was selected as an intervention with potential for increasing placement stability on the basis of strong evidence from two RCTs showing improvements in children's SEDs and increased family harmony. However, the intervention had only been used as a form of early help while children were still living with their parents. KIPP was therefore interested in modifying the model to understand its potential for increasing the placement stability of looked-after children exhibiting severe emotional and behavioural problems.

#### Inputs

GenerationPMTO is delivered by master's-level social workers or psychologists to families on an individual basis (frequently through home visiting) for a period of 19 to 30 weeks, depending on the individual family's needs. US estimates for the annual costs of the set-up and implementation of GenerationPMTO during its first year is \$1,170,000 for 16 practitioners at half time. These costs include initial training fees, supervision and licensing requirements. Unit costs vary depending on the individual social services team's case load requirements.

KIPP delivered GenerationPMTO to families just at the time children were being placed. The team theorised that improvements would be more likely if families received this support at the beginning of placement before child emotional and behavioural problems became intrenched. However, this required some modification to the service's referral and delivery pathways, as well as the hiring and redeployment of staff.

#### **Outputs: Implementation targets**

Figure 2.2 lists the core implementation activities, which KIPP divided into outputs pertaining to practitioner training (which involved technical support from the developer) and implementation outputs (which were the primary responsibility of the agencies implementing the intervention). These implementation outputs included all assessment and referral procedures, family engagement processes and intervention delivery. KIPP then carefully documented how these targets were being met when the intervention was first set up and delivered.

#### **Outputs: Participation targets**

Figure 2.2 also lists a number of participation targets involving the intervention's reach. These targets not only include the number of children initially assessed, but also the characteristics of families successfully recruited and retained in the intervention. Targets were initially set at 70% as a threshold for determining whether the intervention was feasible. Tracking progress towards these targets was then useful for understanding how the intervention was reaching its target audience at each stage of its implementation.

#### Outcomes

Figure 2.2 retained GenerationPMTO's original short-term outcomes of improved parenting and child behaviour, which were consistent with the intervention's original ToC. However, the KIPP team also identified increased placement stability and reunification rates as an additional longer-term intervention goal.

#### Assumptions

When KIPP first identified GenerationPMTO as a potential intervention, they also tested a number of key assumptions about the wider system that would likely impact the intervention's overall success. These assumptions included expectations about the local leadership, the relationships between the agencies delivering the intervention, current referral protocols and the availability of a suitably qualified workforce to deliver the intervention. The KIPP team thus permitted a 10-month exploration period to test these assumptions prior to the set-up and delivery of the intervention.

#### **External factors**

During the exploration period, KIPP also considered a variety of external conditions that could potentially influence the quality of the intervention's implementation and ability to meet its intended outcomes. These conditions included an uncoordinated and fragmented service system, ongoing budget cuts, the recent privatisation of foster care and adoption services and ongoing problems with high caseloads and staff turnover.

#### Implementation outcomes

Through the identification of specific implementation targets, KIPP was able to assess GenerationPMTO's feasibility for achieving its intended outcomes and factors which impacted its successful delivery. This analysis revealed that the social service teams were able to successfully recruit 68% of the families into the intervention within two weeks of placement and retain 98% of the participants through to at least seven weeks of treatment. Surprisingly, one of the factors contributing to this retention was the privatisation of the foster care services, which increased the agencies' accountability to the pre-agreed targets.

The analysis further allowed the department to understand the characteristics of the families who they were not able to successfully engage. These parents were more likely to be angry with the social services team or had renounced involvement with their out-of-home child. This information was highly useful for refining subsequent recruitment and referral efforts.

#### Intervention outcomes

An RCT was set up alongside the implementation study to determine the extent to which the intervention was also able to achieve its intended outcomes with this highly vulnerable population. While the sample size was not large enough to permit firm conclusions, the study observed statistically significant improvements in the behaviour and social and emotional wellbeing of the GenerationPMTO children, with findings suggesting that improvements were associated with increases in children's placement stability.

#### Specifying inputs (resources)

Inputs refer to the resources required to implement the intervention. These resources include the practitioners who will deliver the intervention, their training and qualification needs, the time required to deliver the intervention, as well as management and supervision time. Inputs also include the expense involved in producing or purchasing materials, and any equipment, technological, transportation or venue costs.

Developers should carefully estimate what these resources are from the standpoint of what is required for the intervention to achieve its expected outputs and outcomes. All too often interventions fail because developers have underestimated the time and resources required to recruit families and deliver the intervention to a high standard.

It is also not uncommon for stakeholders to reduce their output activities on the basis of the resources that are available to deliver the intervention, rather than meeting all of the resource requirements. We strongly caution against this, as this will severely limit the intervention's ability to achieve any of its intended outcomes and ultimately result in money being wasted.

#### Describing the situation and identifying priorities

It is useful for logic models to provide a brief description of the situation, which should include some explanation for why the intervention is needed in the first place and could add value to current provision. The nature of this description will vary, however, depending on the purposes of the logic model. For example, in case study 2, the GenerationPMTO intervention was chosen to address a locally determined problem. If the logic model is being developed for more general purposes, the explanation for why the intervention is needed first identified in the ToC should be sufficient. The ToC explanation for why the intervention for why the intervention is needed first identified in the ToC should be sufficient. The ToC explanation for identifying intervention priorities.

#### Assumptions

Assumptions are expectations about how the intervention will be impacted by external processes occurring in the wider system. These assumptions include expectations about the ongoing availability of various resources, including funding, venue space and an appropriately qualified workforce. Assumptions also include expectations about what is required from local leaders, the way in which the intervention should be aligned with local drivers and priorities, and the quality of interagency relationships and referral systems.

It is not uncommon for many of these assumptions to be implicit. This creates a potential for failure or disappointing results if implicit assumptions are not upheld. Some intervention providers therefore aim to make all assumptions explicit through processes that consider 'system readiness.' As we describe in Step 10, developers ideally will have specified these requirements as part of developing the intervention for scale.

Those hosting the intervention may also be encouraged to reflect their own system readiness. For example, the KIPP team permitted a 10-month period to examine the suitability of their wider system for delivering GenerationPMTO by the intervention's developers before they began to install it. This helped ensure that some of the key assumptions were met, while helping to optimise interagency working and quality assurance mechanisms as the intervention was implemented.

#### **External factors**

External factors are similar to assumptions, as they pertain to circumstances that are independent of the intervention that may nevertheless impact its success. However, assumptions are beliefs about what should be in place for the intervention to be successful, whereas external factors describe circumstances that currently exist. Examples of external factors include local priorities and infrastructures, seasonal considerations (including school cycles) and other factors that may negatively or positively affect an intervention's longer-term outcomes. For example, an intervention aiming to increase young people's access to employment is unlikely to be successful if few employment opportunities actually exist.

## Confirming the logic model

An intervention's logic model could be considered initially confirmed when information is provided in all of the boxes illustrated in figure 2.1. This information does not need to be complete, particularly when it comes to details regarding the intervention's content and short-term outcomes, as these will undergo further refinement when the blueprint has been confirmed and some feasibility testing has taken place. However, the logic model should represent a logical extension of what was originally specified in the ToC, including what the intervention will do, who its recipients will be, and what its intended short-term benefits are. There should also be some explanation about the intervention's required inputs, why it is needed and the priorities it might address. A logic model should not be considered completed unless it has also identified assumptions that are core to its successful operation and the ways in which external circumstances might affect its delivery.

A variety of methods can be used to confirm an intervention's logic model, although we recommend that it is useful for the intervention's primary stakeholders to be involved in this process to ensure that the model is relevant and used. In case study 2, participatory methods (see appendix C) were used to reconfirm the GenerationPMTO logic model with looked-after children, and to agree implementation targets to test its feasibility. In some instances, it may be possible to confirm an intervention's logic model during a workshop that takes place subsequent to the ToC workshop(s). However, we advise that there is suitable

time between the two workshops, so that those preparing the logic model can gather information to ensure that the logic model accurately identifies costs and core assumptions.

The template provided in figure 2.1 provides a sufficient starting point for identifying the logic model's core components. However, this template is often not sufficient for specifying the individual components of all of the intervention's delivery and recruitment activities. Although the relationships between inputs (resources), outputs and outcomes will remain fundamentally linear, the processes within these categories will often involve feedback loops and cyclical process. We therefore recommend that additional templates be used to further specify relationships between individual output and outcome components – a blank example is provided on the following page.



## Preparing for the next step

Once developers have confirmed their logic model, they are in a good position to further confirm their intervention's content and test their intervention's feasibility. The logic model can also be used to explain the intervention's core activities and purpose to others.

Nevertheless, developers should remember that an intervention's logic model is never fully finalised. Instead, it should sit alongside the intervention's ToC and be continually revisited, tested and revised throughout the intervention's development. This does not mean, however, that greater specification from the outset is not beneficial. When a logic model is carefully considered and agreed from the beginning, it provides the basis for assuring quality and accountability as the intervention is implemented. From this perspective, better initial thinking will always lead to better ongoing results.

## Step 2 checklist

- The intervention has developed a logic model that identifies inputs, activity outputs, participation outputs, and short-, medium- and long-term outcomes.
- The outcomes identified in the logic model are important from a public health and child development perspective.
- The intervention's intended child outcomes are sufficiently important to justify the intervention's resources and core activities.
- The assumed relationship between the intervention's activities and identified shortterm outcomes is realistic and reasonable. The activities appear sufficiently intensive to bring about the desired change in the target population.
- The relationship between the intervention's identified short-term outcomes and medium- and longer-term outcomes appears realistic and reasonable.
- The intervention's core activities appear to be sufficiently specified.
- The logic model specifies the intervention's intended reach and strategies for recruiting and retaining its target population.
- The logic model has considered the wider context in which the intervention will operate.
- The logic model identifies external factors that might positively or negatively influence the intervention.
- The logic model is well-presented and easy to follow.
#### **Further resources**

#### Logic model building

- Project Oracle: https://project-oracle.com/
- New Philanthropy Capital: https://www.thinknpc.org/
- Nesta: https://www.nesta.org.uk/
- Centre on Theory of Change: https://www.theoryofchange.org/
- Better Evaluation: https://www.betterevaluation.org/en
- Inspiring Impact: https://www.inspiringimpact.org/
- My Community: https://mycommunity.org.uk/
- NCVO Charities Evaluation Services: https://www.ncvo.org.uk/
- RAND: https://www.rand.org/
- Tavistock Institute: http://www.tavinstitute.org/
- University of Wisconsin: https://fyi.extension.wisc.edu/programdevelopment/files/2016/03/Imcourseall.pdf

#### The Cambridge-Somerville Youth Study

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, 54(9), 755.

#### **Case study 2: GenerationPMTO**

Akin, B. A. (2011). Predictors of foster care exits to permanency: A competing risks analysis of reunification, guardianship, and adoption. *Children and Youth Services Review*, 33(6), 999–1011.

- Akin, B. A., Bryson, S. A., McDonald, T., & Walker, S. (2012). Defining a target population at high risk of long-term foster care: barriers to permanency for families of children with serious emotional disturbances. *Child Welfare*, *91*(6).
- Akin, B. A., Bryson, S. A., Testa, M. F., Blase, K. A., McDonald, T., & Melz, H. (2013). Usability testing, initial implementation, and formative evaluation of an evidence-based intervention: Lessons from a demonstration project to reduce long-term foster care. *Evaluation and Program Planning*, *41*, 19–30.
- Bryson, S. A., Akin, B. A., Blase, K. A., McDonald, T., & Walker, S. (2014). Selecting an EBP to reduce long-term foster care: Lessons from a university-child welfare agency partnership. *Journal of Evidence-Based Social Work*, 11(1–2), 208–221.

# Step 3 Create an intervention blueprint



# An intervention blueprint identifies specific learning objectives for each of the intervention's core activities and then links them to the intervention's short-term outcomes.

In Step 3, you will learn how to:

- create an intervention blueprint that links each intervention activity to a specific learning objective
- describe how each learning objective will lead to the intervention's intended short- and long-term child and family outcomes
- develop engaging intervention activities and learning materials that support a wide range of learning styles and needs.

# Blueprints are not just for architects

Once the intervention's ToC and logic model have been confirmed, developers are in a good position to specify their intervention's blueprint. A blueprint is a plan or technical drawing that illustrates how something might be achieved. Developers might therefore rightly assume that a logic model accomplishes this function. However, a blueprint, as we describe it, further specifies the logic model by identifying objectives for each of the intervention's core activities and then links these objectives to the intervention's short-term outcomes. A well-specified blueprint should also describe how each objective is supported by the intervention's content, learning activities and learning materials.

Intervention blueprints serve a variety of useful purposes. First, and perhaps most importantly, an intervention blueprint allows developers to carefully consider whether the intensity of their intervention is sufficient to achieve its intended outcomes. Developers should therefore make sure that their intervention provides adequate opportunities for recipients to understand, practise and apply new skills.

Second, blueprints help developers consider the nature of their intervention's activities. Are they appropriately varied, engaging and easy to do? If not, are there strategies in place to encourage participants to engage in activities they find difficult? For example, in case study 2, the developers discovered that some of the participants took much longer to engage than originally anticipated because of ongoing resentments towards the social services system. Activities were therefore necessary to help practitioners gain the participants' trust and

reduce their resentment, so that they were motivated to work through issues contributing to their child being taken into care.

Third, blueprints help developers consider whether there is an obvious and logical connection between their intervention's activities, core objectives and intended outcomes. While these connections may be clear to the developer – they may not always be clear to others. For example, drama interventions are commonly used to promote health issues within schools. However, the connection between the drama activities and the information children are intended to learn is not always clear. The intervention blueprint provides an important opportunity to clarify connections that may be otherwise ambiguous.

Finally, an intervention blueprint is highly useful for training and quality assurance purposes. Once the blueprint has been specified, developers are then in a good position to start confirming the intervention's lesson plans and training materials. Intervention blueprints are also highly useful for helping those delivering the intervention monitor the quality of implementation. We will provide further details about how this can be done in Step 4.

# **Case study 3: Specifying an intervention blueprint**

The Promoting Alternative Thinking Strategies (PATHS) intervention is a four-year schoolbased curriculum that aims to help primary school children develop positive social and emotional learning (SEL) skills. Its ToC is informed by scientific evidence linking improved social and emotional understanding during primary school to increases in children's school achievement and positive citizenship during adolescence. In the short term, PATHS aims to improve children's self-control, emotional understanding, interpersonal problem-solving and peer relationships. In the long term, PATHS aims to increase children's commitment to school and school achievement, as well as reduce their involvement with harmful substances, crime and risky sexual behaviour.

The PATHS curriculum is designed to be integrated into a standard primary school curriculum, consisting of four year-long units which correspond with each of its four short-term outcomes. Year 1 is devoted to helping children increase their self-control, whereas years 2, 3 and 4 respectively cover children's emotional understanding, interpersonal problem-solving skills and peer relationships. Decisions about this order and length were based on consistent evidence from the scientific literature indicating that the ability to control one's behaviour and understand emotions precedes interpersonal problem-solving skills and positive peer relations. The four-year curriculum is also based on evidence suggesting that comprehensive, school-wide strategies are necessary for social and emotional skills to be fully learned and applied in diverse circumstances.

Table 3.1 summarises the lessons and activities for the first 10 weeks of the PATHS Year 2 curriculum, linking them to their core learning objectives and intended short-term outcomes. The PATHS activities were chosen specifically to help children integrate their emotional understanding with cognitive and linguistic skills, so that they can apply them appropriately to a variety of contexts within and outside of school.

Each lesson lasts between 30 and 40 minutes and is presented by the children's primary teacher on a weekly basis throughout the school year. During these lessons, children learn social and emotional learning strategies through group discussions, story formation projects, self-control exercises and puppet play. Lessons are also supplemented with homework assignments sent home to the children's caregivers.

# **Confirming the intervention blueprint**

Classroom lesson plans are an obvious example of an intervention blueprint. For example, Table 3.1 links the details of the first 10 weeks of the Year 2 PATHS curriculum (case study 3) to its core activities, materials and learning objectives, which are then logically linked to the intervention's short-term goals. An intervention blueprint can be created for any kind of intervention model, however, as long as clear connections exist between the intervention's activities, objectives and short-term outcomes. There is also no right or wrong way to confirm an intervention's blueprint. While the participatory methods described in the previous steps are useful for specifying a blueprint's details, they are certainly not necessary. However, it is helpful if developers use their intervention blueprint as an opportunity to provide as much detail about their intervention's activities as possible, including what the activities are and how they will take place, as well as the materials that will be used.

A well-specified blueprint should also include information about the intervention's learning activities and these should be sufficiently varied to cater to a variety of different styles of learning. For example, some individuals learn best from visual materials, whereas others learn better from information that they hear. Studies also show that interventions stand a greater chance of being effective when participants have multiple opportunities to apply newly learned skills and receive feedback and coaching from the practitioner. Examples of activities that encourage practitioner feedback include group projects, role play, group discussions and homework assignments. The PATHS example in case study 3 makes use of a wide variety of engaging learning activities to reinforce children's knowledge of the intervention's four primary outcomes through group activities involving dramatic play, discussions and learning charts. Homework exercises sent to the children's parents also help ensure that learning will be reinforced at the children's home.

# TABLE 3.1

Weekly lesson plans for PATHS (Promoting Alternative Thinking Strategies) for first 10 weeks of year 2 content

Lesson	Content/activities	Parent activity or handout	Materials	Objectives	Primary short- term outcome
1	<b>Classroom Rules</b> . Introduce children to PATHS Help children internalise the importance of having rules and structures within a group Orient children towards becoming attentive listeners Allow children to participate in the process of creating rules and structures	Home activity sheet Home Letter	Mrs Brown's rules poster Pictures 1A–1B Paper, felt pens, markers	Children have greater awareness of PATHS rules Children learn rules for respectful and positive interaction Children appreciate the merits of team working	Social problem- solving Peer relations and self-esteem
2	PATHS Readiness Create an atmosphere that emphasises PATHS as a special and fun time Provide practice in paying attention to others Review same and different Review IS and IS NOT Review rules during PATHS time		Chart listing class rules 9 x Animal Pictures Actor/Actress Badge	Children gain an understanding of PATHS rules Increased understanding of respectful interactions Heightened awareness of teem working	Social problem- solving Peer relations and self-esteem

Lesson	Content/activities	Parent activity or handout	Materials	Objectives	Primary short- term outcome
3	PATHS Pupil for Today – Complimenting Introduce the concepts of complimenting and compliment Review the different feelings when receiving a compliment Practise what to do or say when receiving a compliment	PATHS Pupil for Today home letter for child chosen; for all pupils Home letter Home activity sheet	Two jars Strips of paper for each child Compliment list PATHS Pupil for Today Poster, felt tip, pens	Children have a deeper understanding of treating each other positively	Social problem- solving Peer relations and self-esteem Emotional understanding
4	Introduction to feelings Introduce the concept of feelings Illustrate that we can experience more than one feeling at a time Show feelings are inside but can show on the outside		Finished happy and sad faces for each child Happy faces and feeling rings	Children gain a deeper understanding of complex emotions	Emotional understanding
5	Happy, Sad, Private Introduce children to basic emotions and to a paradigm for thinking about them Introduce the ideas that we can guess how people form facial expressions, body postures and situational cues. Help children associate feeling labels and feeling expressions	Home activity sheet lesson 6	Feelings Charts – comfortable and uncomfortable Picture 5A Photographs 1–4 Feeling faces for happy, sad and private Examples of finished faces	Children have a deeper understanding of basic emotions	Emotional understanding
6	Fine, Excited, Tired Role-play feeling excited and tired Make a drawing of excited and tired Review all feelings		Feelings Charts – comfortable and uncomfortable Picture 7A–7G Photographs 5–10 Feeling faces & examples of finished faces	Have an increased understanding of feeling fine, excited or tired	Emotional understanding
7	Scared/Afraid, and Safe To role-play Scared/Afraid, and Safe To read and discuss stories with Scared/Afraid, and Safe situations		Feelings Charts Picture 9A–9H Photographs 11–14 Feeling faces examples Taylor's visit to the dentist	Understand the feelings of afraid and safe	Emotional understanding
8	Cross/Angry (Feelings vs Behaviours) To review Happy, Sad, Private, Fine, Excited, Tired, Cross/ Angry, Scared/Afraid, and Safe To provide practice in role- playing and in observing emotional cues		Feeling Charts – comfortable and uncomfortable Pictures 11A –11H Photographs for cross/angry Cross/ angry feeling faces Box or bag with 3x5 cards which have printed happy, sad,	Understand what one should do when one is cross Understand how to think again before taking action Have a deeper understanding of reading others' emotions	Emotional understanding

Lesson	Content/activities	Parent activity or handout	Materials	Objectives	Primary short- term outcome
9	Self-Control 1 Explore different ways for gaining self-control Introduce the concept of calming down Discuss the Red Light		Control Signal Poster 3 Steps for Calming Down Poster (red light) Teacher's Manual for the Control poster	Understand different ways for controlling difficult emotions Have strategies for calming oneself down	Emotional understanding Self-control
10	Self-control 2 Review the steps for calming down Provide practice in using these steps – stating the problem and feelings		Control Signal Poster 3 Steps for Calming Down Poster (Red Light)	Understand different ways for controlling difficult emotions Have strategies for calming oneself down	Emotional understanding Self-control

# **Step 3 checklist**

- The intervention has a blueprint or other document that clearly describes its core activities and links them to specific objectives that are related to the intervention's intended short-term outcomes.
- The relationship between the intervention's activities, objectives and outcomes is clear and plausible.

The intervention's activities are feasible.

- The opportunities to learn and master new skills are sufficient for the age, needs and vulnerabilities of the target population.
- The activities are interesting and engaging. When necessary, the blueprint specifies strategies to engage participants and overcome resistance. The activities are sufficiently diverse for a wide variety of learning needs.
- The blueprint has identified opportunities for intervention recipients to learn and practise new skills through direct feedback and coaching.

#### **Further resources**

#### Intervention blueprints

Jackson, C., & Dickinson, D. M. (2009). Developing parenting programs to prevent child health risk behaviors: a practice model. *Health education research*, 24(6), 1029–1042.

#### Case study 3: Promoting Alternative Thinking Strategies (PATHS)

- Greenberg, M. T., & Kusché, C. A. (2006). Building social and emotional competence: The PATHS curriculum. In S. R. Jimerson & M. Furlong (eds), *Handbook of school violence and school safety: From research to practice* (pp. 395–412). Lawrence Erlbaum Associates Publishers.
- Greenberg, M. T., Kusche, C. A., Cook, E. T., & Quamma, J. P. (1995). Promoting emotional competence in schoolaged children: The effects of the PATHS curriculum. *Development and Psychopathology*, 7, 117–136.
- Kam, C. M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science*, 4(1), 55–63.

# Step 4 Conduct a feasibility study



# A feasibility study (sometimes referred to as a process evaluation) tests whether the intervention can achieve its intended outputs.

These outputs include the intervention's core activities, as well as its ability to recruit and retain its intended participants. In Step 4 you will learn how to:

- specify the intervention's core activities and identify the factors that support or interfere with their successful delivery
- use qualitative research methods to understand which factors contribute to the success of the intervention from the perspective of those delivering it
- use qualitative methods to understand how those receiving the intervention benefit from it and whether these benefits are consistent with the intervention's original theory of change
- understand how best to recruit and retain participants
- use quantitative methods to make sure that the intervention is reaching those who need the intervention the most
- measure user satisfaction
- track and document intervention costs.

# Can the intervention work? Verifying feasibility

The term 'feasibility study' encompasses a wide range of activities that consider whether the key components of an intervention's logic model – including its resources, activities and population reach – are practical and achievable. In this respect, a feasibility study (also referred to as a type of process evaluation) is primarily concerned with whether an intervention *can* work, rather than whether it *does* work. Considering whether an intervention does work (in other words, has it achieved its intended child outcomes?) is the primary aim of Steps 5–7. Step 4 allows developers to investigate whether an intervention can work by systematically testing the intervention's progress towards its intended outputs as it is being implemented. Investigating an intervention's feasibility is essential for determining whether it is worth further investment. After all, there is little point to evaluating an intervention's impact if it is too difficult to deliver, expensive or its intended recipients are not willing to come to it. Feasibility studies help developers consider these issues by collecting information on how the intervention is being delivered, who it is reaching and how much it costs. More comprehensive feasibility studies also consider the value of the intervention from the perspective of those who are receiving it.

Feasibility studies frequently make use of a variety of quantitative and qualitative research methods to investigate an intervention's feasibility from a variety of perspectives. Quantitative methods are useful for determining whether the intervention is reaching its delivery and recruitment targets, whereas qualitative methods are beneficial for understanding the views of the intervention's recipients and whether these views are consistent with the intervention or service's ToC.

This chapter provides clear advice on the use of qualitative and quantitative methods for conducting a comprehensive feasibility study. While these methods are particularly valuable for determining an intervention's feasibility during the initial phases of its development, checking and monitoring feasibility is necessary for well-established interventions. More advice on the methods for monitoring the feasibility of interventions with proven child outcomes is provided in Step 8.

# Feasibility of delivery

# An implementation task list

Many activities – aside from the intervention itself – contribute to an intervention's successful delivery. These activities include securing the intervention's funding, recruiting a suitably qualified workforce, training and supervising practitioners, agreeing safeguarding arrangements, setting up monitoring and data management systems, locating a venue, etc. While many of these activities may have already been identified in the resources and outputs section of the intervention's logic model, others may be difficult to anticipate.

Detailing all of the activities necessary to implement an intervention in its logic model can also make it unnecessarily unwieldly, thus making the model impractical. Nevertheless, is helpful to have all of the intervention's implementation tasks documented somewhere, as any one of them could significantly delay or derail an intervention's successful implementation. Step 4 is the ideal time to do this, through the use of documentation, such as the implementation task log provided in table 4.1, to carefully record all of the activities involved in the delivery of the intervention as it is being implemented. Task logs can initially be populated with the resources and activities initially identified in the intervention's logic model and blueprint, and then added to as other important activities become evident. Task logs are also useful for documenting the cost of various activities, especially when used in conjunction with online time tracking tools, such as Toggl, Everhour or HourStack.

Task logs are particularly useful if they provide space for recording all of the activities contributing in an intervention's delivery alongside any corresponding barriers. Understanding barriers is particularly helpful for identifying 'universal threats' that could negatively affect the delivery of an intervention every time it is implemented versus one-off bottlenecks. For example, the repainting of the intervention's venue might occasionally create a bottleneck, but is unlikely to create a problem in future roll-outs. However, a complete lack of a permanent intervention venue likely represents a potential threat that could impact an intervention every time it is implemented.

As the intervention matures, task lists can be refined into groups of activities that represent significant implementation milestones when they are completed. These milestones can then

be monitored every time the intervention is implemented in a new location. For example, in case study 2 (Step 2) groups of implementation activities were classified into leadership, organisational and competency 'drivers' that needed to be in place before families could be recruited and the intervention was delivered.

## Understanding practitioner views

While implementation task lists are useful for documenting the feasibility of various intervention activities, they provide little information about the task's difficulty. Comprehensive feasibility studies therefore also often make use of qualitative depth interviews to understand implementation feasibility from the perspective of those delivering the intervention. These interviews need not involve all stakeholders involved in an intervention's delivery but should include a representative sample of practitioners and managers.

Ideally, practitioner views should be considered through one-to-one depth interviews, as was done in case study 4 and described in more detail in appendix C. The implementation task list can provide a useful starting point for the interview topic guide, although questions about practitioners' understanding the intervention's curriculum and core activities are also highly useful. A blank template is provided as table 4.1, and a partial example as table 4.2. Specific tasks can be adapted from programme to programme.

This knowledge is not only helpful for understanding intervention feasibility, but also for developing practitioner training materials, fidelity checklists and other quality assurance systems necessary for delivering the intervention at scale. Further details about how this can be done are described in Step 8.

## TABLE 4.1

Implementation task list: template

Activity description	Party responsible	Start date	End date	Notes (successes & challenges)	Bottleneck or universal risks?	Estimate time (in days) to complete	Cost
Resources							
1. Agree agency partnerships							
2. Agree budget and plan financial resources							
3. Recruit workforce							
4. Secure a venue							
5. Secure creche workers							
6. Secure/produce materials							
Practitioner training outputs							
7. Practitioners attend training							
8. Practitioners certified							
9. Supervision arrangements confirmed							
10. Fidelity and quality assurance arrangements confirmed							
Delivery outputs							
11. Referral and recruitment							
12. First session (visit; class, etc.)							
13. Second session							
14. Third session							

# **TABLE 4.2**

Implementation task list: partially completed example

Activity description	Party responsible	Start date	End date	Notes (successes & challenges)	Bottleneck or universal risks?	Estimate time (in days) to complete	Cost
Resources				1	1		
1. Agree agency partnerships	Partnerships manager (at programme developer and host agency)	July 2017	January 2018	First partnership between programme developer and host agency broke down so additional time was needed to identify an alternative host agency.	Risk to delay to intended programme start date.	X days	£xxx
2. Agree budget and plan financial resources	Director of Finance and Operations; project manager	June 2017	January 2018	More time was required to recruit participants than anticipated, so this phase was over budget.	Risk of exceeding planned budget.		
3. Recruit workforce	Host agency	January 2018	July 2018	The host agency was able to recruit the required staff, including recruiting existing employees who will receive additional training.	Inability to recruit sufficient staff would have delayed programme delivery.		
4. Secure a venue	Host agency	July 2018	August 2018	Individual responsible for securing an intervention location was on leave, so no permanent venue was identified. Intervention location thus needed to change with every session, creating transportation problems for some participants and derailing intervention momentum.	A lack of venue location could interfere with the successful delivery the intervention on all occasions; lack of a permanent location poses a universal threat.		

# Case study 4: Testing the feasibility of Group Family Nurse Partnership (gFNP)

### The intervention

Group Family Nurse Partnership (gFNP) is a modification of the Family Nurse Partnership (FNP) programme, which was first developed in the US to support the needs of single, firsttime teenage mothers and their children. FNP originally targeted this population because of their high vulnerability and the view that they would be more likely to welcome additional support, given that it is their first pregnancy.

Eligible mothers are enrolled into the programme at the time of their first antenatal booking and then visited on a biweekly or monthly basis (depending on the child's age) by trained and qualified family nurses until the child's second birthday. During these visits, family nurses build a supportive relationship with the mothers which then provides the context in which mothers learn positive parenting behaviours and strategies for supporting their own and their child's needs.

Multiple RCTs conducted in the US and the Netherlands have confirmed measurable improvements in FNP mothers' mental well-being, parenting skills, family relationships, education and future employment opportunities. Benefits for FNP children include improved cognitive and emotional development in early childhood and a decreased risk of poor academic and health outcomes in adolescence, including a significantly reduced risk of a preventable death.

A series of feasibility studies conducted between 2008 and 2010 observed that FNP's implementation milestones were achievable within the UK health system and that it was well-liked by the practitioners trained to deliver it. These studies also observed that FNP mothers viewed the intervention to provide value in a way that was consistent with FNP's original ToC.

Given FNP's initial UK success, the national FNP unit sought to develop a less intensive, group-based version of the programme for highly vulnerable families who did not meet the intervention's original strict eligibility criteria. Eligible participants thus did not need to be teenage, single or first-time, but they did need to be vulnerable. The revised eligibility requirements included mothers under the age of 20 expecting a second or subsequent child, and those aged 20 to 24 and expecting a first child with low educational achievement. This last criterion was added for 20 to 24-year-old mothers as an additional eligibility requirement after the first implementation study.

Other alterations to the FNP model included a reduction in length, lasting 18 months rather than 2.5 years, and explicit efforts to involve fathers, so they could also learn strategies for supporting their children's development.

## gFNP feasibility study

Although the gFNP providers assumed that the group format would be acceptable to vulnerable families, there were concerns that some may find the information less relevant for their circumstances or have difficulty attending the intervention for the entire 18 months of its duration. The gFNP feasibility study therefore sought to answer the following four questions:

- Were there any barriers to the intervention reaching its intended population?
- Were there any client factors that influenced attendance?
- · Could the intervention be delivered over 18 months?
- · Was the intervention acceptable to those delivering it and the families attending it?

The evaluation was in two phases, with the second phase using a modified staffing model based on issues emerging from the first phase of the evaluation. Quantitative and qualitative evaluation methods were used to answer all four questions. Quantitative methods included

information about the characteristics of 61 gFNP participants gathered from referral data and attendance records. Qualitative information was gathered through depth-interviews and open-ended feedback from practitioners completed after sessions (see appendix C for more information) conducted with 16 workers involved in the delivery of the programme and 35 FNP participants.

## Findings

### Question 1: Are there barriers to reaching the intended population?

Findings involving the attendance records revealed challenges in recruiting a sufficient number of mothers with the same delivery date for the groups to commence during the early stages of their pregnancies. Findings from the practitioner depth interviews suggested that these delays were mainly due to midwives having difficulties in identifying enough mothers who met the eligibility criteria, shared a similar delivery date and lived close enough to the group meeting location. Inconsistencies in the communication between the community midwives making the referrals and the family nurses also contributed to difficulties in recruiting mothers. This meant that some participants started later, meaning that that relationships with other group members were not always easy to establish and core content was often missed. Group size also appeared to influence the sustainability of the groups, as small groups had difficulty maintaining momentum.

Some family nurses also reported that the eligibility criteria may have interfered with their ability to recruit mothers during the second phase of the pilot work. In particular, issues arose around addressing questions about educational qualifications. Such questions were not required to recruit mothers to FNP, because their age and pregnancy status were collecting routinely as part of the initial booking.

### Question 2: Are any client factors related to attendance?

Information about the mothers' age, relationship status, number of other children, neighbourhood, income, education, and so on, gathered at the time of enrolment allowed the evaluators to compare the mother's demographic information with their attendance records. Although not statistically significant, these comparisons suggested that mothers with higher qualifications were more likely to attend more sessions. By contrast, attendance was lowest among mothers who were never employed or were living alone, meaning that the intervention was not fully reaching the most vulnerable parents.

## Question 3: Can programme delivery be sustained over 18 months?

Attendance records confirmed that attendance was highest during the mothers' pregnancies, and then dropped off during the months after the baby was born. On average, the mothers attended fewer than half of the scheduled group sessions. Reasons provided by the mothers for lack of attendance included other activities competing for their time, being tired and transportation difficulties. Some mothers also reported returning to work within several months following their baby's birth.

The practitioners reported that the successive nature of the content meant that mothers who missed sessions sometimes had difficulty catching up. Some practitioners perceived this as a negative by-product of the group format, since the traditional FNP home visits allowed sessions to be scheduled at the mothers' convenience, reducing the likelihood that core content would be missed.

#### Question 4: Is the programme acceptable to different stakeholders?

This question considered the extent to which the intervention was acceptable from the perspective of the mothers who attended gFNP and the practitioners who delivered it. Generally, the participants found that the group format was acceptable and in some cases close friendships were made that extended beyond the programme context. However, others found the group format less welcoming, especially if they were hesitant to participate or

viewed other group members as dominating. Some mothers also found that discussions involving domestic violence were challenging to deal with in a group context.

The mothers had mixed opinions about whether fathers should take part in groups, but men who were interviewed spoke positively about the experience. The content was nevertheless found to be acceptable and post-session user satisfaction ratings were almost always high.

Nurses and midwives additionally said they would have benefited from greater familiarity with the material, suggesting that more preparation time might have been needed before implementing the programme. In the first phase of implementation, all practitioners were family nurses, some of whom also had current fitness to practise as a midwife. Roles and responsibilities between the midwives, family nurses and support workers in the second phase of the pilot, based on a different staffing model, could also have been clearer. The non-FNP practitioners would have benefited from more background about the programme and about their role in relation to the family nurses.

### Conclusions

Although the first feasibility study of gFNP revealed a number of barriers to recruiting a sufficient number of mothers to convene the groups in the early stages of the mothers' pregnancy, these barriers were not necessarily insurmountable. However, consistent attendance for 18 months may have required more commitment from the mothers than they were willing to provide. The group format also may have made it difficult for some of the more vulnerable mothers to fully participate, thereby substantially reducing the dosage of the intervention they received. These limitations, combined with the initial disappointing findings from the UK FNP randomised controlled trial, resulted in the gFNP programme being discontinued, despite it being viewed favourably by some of its participants.

# Feasibility of recruitment, retention and reach

The success of any intervention is fundamentally determined by the degree to which participants are willing to attend and view its contents and activities to be interesting and worthwhile. In order for this to take place, participants need to (1) know about the intervention, (2) be motivated to attend, and (3) be able to attend on a regular basis. A comprehensive feasibility study should therefore have systems in place for collecting information on all three of these factors, as well as processes for understanding the intervention's 'reach' – in other words, the extent to which the intervention can recruit and retain the primary target population identified in its ToC. These systems include quantitative processes for collecting the demographic details of participants and tracking their attendance, and qualitative processes for understanding participants' views. These views should include participant satisfaction and the extent to which they perceive the intervention to provide benefits that are consistent with the outcomes identified in the intervention's ToC.

## **Recruitment feasibility**

Successful recruitment involves identifying eligible participants, making them aware of the intervention and motivating them to attend. All three of these processes can be challenging. In the example involving gFNP (case study 4), the practitioners had difficulty identifying enough eligible participants in time to create the groups so that they were aligned with the mothers' pregnancy due dates. Once these individuals were identified, practitioners then faced additional challenges in motivating some participants to attend – either because the participants viewed attendance as stigmatising or because other activities competed for their time.

These difficulties were first identified through the attendance records, which revealed a lack of attendance throughout the duration of the intervention. These findings were further confirmed through the depth interviews, which allowed the practitioners to explain the reasons for the recruitment challenges. The depth interviews also provided insight into why some mothers were motivated to attend the groups. However, the depth interviews offered relatively little information as to why some mothers did not attend in the first place or eventually dropped out of the groups.

Further information about participants' lack of attendance might have been gained through follow-up interviews with those who were not successfully recruited to gFNP or dropped out of the intervention at its early stages. The evaluators nevertheless gained some insight into the characteristics of those who dropped out by comparing the demographic details of those who remained in the intervention to those who left. This analysis revealed that single and isolated mothers were less likely to attend in comparison to mothers who were less vulnerable. Although the difference was not statistically significant, the evaluators interpreted these findings to mean that the intervention was not fully successful in reaching its primary target population by failing to engage the most vulnerable families.

## **Retention and user satisfaction**

Depth interviews with those attending the intervention are highly useful for understanding what participants liked about the intervention, as well as the aspects that were less successful. For example, there was consensus among the mothers attending gFNP that they did not like the video on infant cot deaths.

The depth interviews were also able to explore whether participants reported benefits for themselves and their children that were consistent with the intervention's primary goals and ToC. For example, a key aim of gFNP was to increase father involvement. However, some of the mothers viewed fathers attending as off-putting, so did not perceive this feature to be beneficial. The extent to which this feature resulted in programme attrition remains unknown, although this issue could have been explored in further detail if depth interviews had been conducted with those who dropped out of the intervention.

While staying in contact with those who have dropped out of an intervention can be challenging, it is achievable if participants are recruited to the evaluation independently of the intervention. For example, dropouts might be offered incentives (for example, gift vouchers) to participate in depth interviews, or telephone interviews that are specifically designed to find out why they left. Alternatively, participants might be recruited to the study independently of the intervention at the time of enrolment, and then followed up regardless of their level of intervention involvement.

User satisfaction surveys involving five or seven-point Likert scales (see figure 4.1) are commonly used to gauge participant satisfaction, although the information they provide is often highly limited. This is because users tend to respond positively to surveys, *unless* they are seriously disappointed – in which case they usually leave the intervention before it is time to complete the survey. This means that the findings from most user satisfaction surveys are highly skewed towards the views of satisfied participants.

Another drawback of user satisfaction surveys is that they provide limited information about what participants specifically like and dislike about an intervention, as all questions are predetermined by those who designed the survey. In the case of gFNP, many parents attended the intervention because they liked the group format and valued the opportunity to meet and engage with other parents. However, these views were not universally held by all parents and user satisfaction surveys would not have been able to capture important differences in these perceptions. Nevertheless, knowledge of these differences was vital for understanding the feasibility of the group format for reaching the most vulnerable families. While the group format did increase the acceptance of the intervention for some participants, it reduced acceptance among the more vulnerable families, thereby reducing the intervention's overall feasibility with its primary target population.

## FIGURE 4.1

An example of a five-item Likert scale

How satisfied are you with the intervention materials?					
	Extremely satisfied				
	Very satisfied				
	Moderately satisfied				
	Slightly satisfied				
	Not satisfied				

## Understanding reach

Intervention reach pertains to knowledge about who the intervention is and is not reaching. In the case of gFNP, reach was understood by comparing the demographic details of those who attended the intervention to those who did not. This analysis revealed that the intervention was most successful in recruiting and retaining the least vulnerable mothers meeting the eligibility requirements – suggesting that the intervention had limited success in reaching its target population.

Comparing the demographics of those who did and did not attend the intervention was therefore sufficient for understanding the reach of gFNP. However, such methods would not be sufficient for understanding the reach of universal interventions, when the demographic details of non-participants are typically not known. In these cases, intervention reach is best understood by comparing the demographic details of those attending the intervention to those of the local population. Information about the local population can be found within the various datasets published by the Office for National Statistics,<sup>7</sup> as well as local datasets. Demographic information should include anonymised details about the participants' age, neighbourhood (including level of deprivation and housing stock), family composition and ethnicity. Ideally, information on family income and parents' level of deprivation should also be collected when possible.

# **Tracking service use**

An intervention's usefulness can also be considered by tracking what happens to participants after they leave the intervention. For example, routine monitoring data that tracks teenage mothers' use of services after enrolment in the standard version of the FNP programme in the United States consistently reveals that FNP participants are more likely to go off

**10 STEPS FOR EVALUATION SUCCESS** 

STEP

<sup>7</sup> See: https://www.ons.gov.uk/

food benefits in comparison to non-FNP mothers with similar characteristics. Although inconsistencies in the quality of the monitoring data and the lack of an appropriately matched comparison group limit the conclusions that could be drawn from these observations, these immediate service outcomes are nevertheless viewed as a positive indication that the intervention is likely to achieve its longer-term parent and child goals.

Tracking service use is only feasible, however, if good data management systems are in place to routinely collect information on children and families. At the very least, this information should include:

- information about the demographic characteristics of eligible participants, whether or not they attend the intervention
- information about children and parents use and non-use of services
- the ability to track children and parents' use of multiple services.

Data management systems with these capabilities must adhere to data management protocols involving the anonymisation and sharing of information.

# Feasibility of cost

No feasibility study would be complete without some understanding of the costs of the intervention's inputs. Cost information is typically understood in terms of start-up costs (including training and materials) and running costs (the cost of running the intervention every time it is implemented). Once this information is gathered, a unit cost per participant can be calculated.

Unit costs can then be used to compare the expense of the intervention to alternative services. Clearly, the intervention may not be feasible if the unit costs are extremely high, particularly if the intervention's anticipated outcomes are similar to those that might be achieved with less costly services. However, high unit costs might be justified if there is reason to believe that intervention could provide benefits that might offset high intervention costs. For example, the high cost of the intervention described in case study 2 was immediately offset by substantial reductions in children going into care occurring as a result of the generationPMTO intervention. However, understanding the extent to which intervention costs might be outweighed by potential benefits is best understood through a cost–benefit analysis, which should only be conducted after a positive result has been confirmed through a high-quality impact evaluation, as described in Step 6.

An implementation task log as described in the previous section provides a useful starting point for considering intervention costs. Table 4.3 additionally provides an example of how intervention costs were documented for the set-up and running of the Incredible Years programme in Ireland in 2007.

## TABLE 4.3

Total and per-child costs of running parenting group over 12-session Incredible Years programme

	Mean (SD) unit cost (£)	Mean (SD) units	Total cost (£)
Non-recurrent initial training and group setup costs			
Materials (programme kit)	735.00	1	735.00
Initial group leader training:			
Training course fee	350.00 per leader	2 leaders/group	700.00
Time at training course for two leaders	22.94 (5.27)/hour	45 hours	1032.10
Travel time to training course	22.94 (5.27)/hour	8 hours	183.52
Mileage to attend course for two leaders	0.34/mile	160 miles	54.24
Subtotal			2704.86
Recurrent group running costs			
Supervision of group leaders before start of programme:			
Time for two group leaders with trainer	22.94 (5.27)/hour	6 hours	137.61
Travel time for two group leaders to supervision	22.94 (5.27)/hour	4 hours	91.70
Mileage	0.34/mile	640 miles	217.60
Trainer <sup>+</sup> costs <sup>‡</sup>	62.50/hour	1 hour	62.50
Recruitment of parents:			
Time for two group leaders spent in visits to recruit parents	22.94 (5.27)/hour	24 hours	550.56
Group leader travel time to recruit parents	22.94 (5.27)/hour	12 hours	275.28
Cost of telephone calls to recruit parents	0.03 per min	210 mins	6.30
Group costs:			
Group materials pack			611.45
Time for two group leaders running sessions	22.94 (5.27)/hour	51.81 (2.94) hours	1188.35
Time for two group leaders outside sessions (preparation, administration, follow-up with parents)	22.94 (5.27)/hour	139.11 (13.73) hrs	3190.51
Time for two group leaders in three-hour weekly supervision with trainer	22.94 (5.27)/hour	72 hours	1651.36
Travel time for two group leaders to attend weekly supervision with trainer	22.94 (5.27)/hour	48 hours	1100.91
Mileage	0.34/mile	1920 miles	650.88
Other running costs:			
IY trainer $^{\scriptscriptstyle \dagger}$ costs for weekly supervision	62.50/hour	12 hours	750.00
Costs of clerical support to group	9.70/hour	8 hours	77.60
Telephone calls to parents	0.03/min	1129.8 (688.8) ms	33.98
Transport and crèche facilities			1057.57
Venue rental and refreshments			1109.63
Subtotal			12763.65
Cost of establishing and running parenting group over 12-week pro	gramme:		
Total			15468.51
Cost/child based on 8/group			1933.56
Cost/child based on 12/group			1289.04
Costs of running parenting group excluding non-recurrent costs:			
Total			12763.65
Cost/child based on 8/group			1595.46
Cost/child based on 12/group			1063.64

Notes:

In some cases, total costs do not equal product of mean unit costs and mean units because of rounding.

<sup>†</sup>Consultant clinical psychologist. <sup>‡</sup>Supervision delivered to three sets of group leaders at a time.

Source: Edwards, R. T., Céilleachair, A., Bywater, T., Hughes, D. A., & Hutchings, J. (2007). Parenting programme for parents of children at risk of developing conduct disorder: cost effectiveness analysis. *British Medical Journal*, 334(7595), 682.

# Setting feasibility targets

Once the intervention or service has been run several times and information about implementation and attendance has been collected, it is possible to benchmark its feasibility against specific implementation targets. Examples of implementation targets applicable to the majority of interventions include:

- targets for obtaining intervention inputs, including a venue and intervention materials
- workforce recruitment targets
- workforce retention targets
- practitioner training targets
- fidelity for practitioners delivering the intervention as intended (intervention fidelity is described in greater detail in Step 8)
- participant recruitment targets, including the ability to recruit sufficient numbers of eligible participants
- participant retention targets
- targets for participant satisfaction
- targets for subsequent service use.

Further feasibility targets specific to the intervention might also be identified through the implementation task list or findings from depth interviews conducted with the practitioners and participants.

Feasibility targets should be realistic, but also meaningful. For example, the stakeholders involved in the selection of GenerationPMTO decided that the intervention would not provide meaningful value to the child protection service unless it could maintain a 70% recruitment rate. This target proved to be unrealistic, given the negative views the families had of the child protection system. Nevertheless, recruitment was close enough at 68%, and retention exceeded expectations at 98%. In the example of gFNP, no recruitment target was set prior to the feasibility study, although the analysis of attendance data demonstrated that while initial recruitment was high (84%), attendance steadily dropped off throughout the duration of the programme. In the end, participants attended just over half of all the sessions, significantly reducing families' exposure to the gFNP content.

Failure to meet feasibility targets does not necessarily mean that the intervention should be abandoned, although they do often suggest that revisions are likely necessary to improve the intervention's delivery and recruitment methods. Ideas for increasing the feasibility of delivery and recruitment processes can often be found in the content of the depth interviews, if done to a high standard. For example, findings from the depth interviews used in the GenerationPMTO feasibility study highlighted the resentment families had towards the child protection system. This knowledge in turn allowed the GenerationPMTO providers to modify recruitment processes to overcome these barriers, thus increasing recruitment rates over time.

While failure to meet feasibility targets may initially be disappointing, any resulting improvements are likely to dramatically improve the quality of the intervention, as well as increase the likelihood of a positive result once the intervention is ready to undergo a more rigorous evaluation (Steps 5, 6 and 7). However, in cases where interventions continue to struggle to meet feasibility targets, providers will need to be self-critical in considering whether the intervention is in fact needed, or some obstacles simply cannot be overcome. In these cases, it may be necessary to abandon the model all together or revisit the core assumptions of the logic model ToC. Regardless, these decisions should inevitably lead to an intervention that is both needed and feasible, and therefore more likely to provide meaningful benefits to children and families.

# Step 4 checklist

A comprehensive feasibility study should include methods to meet the following requirements.

A feasibility study been conducted to verify the resources necessary to implement the intervention.
Barriers to obtaining resources for delivering the intervention have been identified.
There is clear information on intervention costs, and on the extent to which intervention costs provide value for money in comparison to similar activities.
A feasibility study has been conducted to verify workforce requirements. Challenges in identifying a suitably qualified workforce have been identified.
The training requirements for the intervention are identified and some work has been done to understand their feasibility.
The feasibility study has confirmed a set of clear delivery activities and targets.
Information about the barriers to successful delivery have been identified, as well as the measures that can be taken to overcome these barriers.
Information is available about practitioner views, including whether they view it as valued and needed, and are able to deliver it.
There is feasibility evidence that practitioners can deliver the intervention with fidelity.
Information on how to deliver the intervention with fidelity, through the use of fidelity checklists, supervision and video-taping is available.
A feasibility study has carefully considered the ease with which participants can be recruited and retained in the intervention.
The intervention's typical recruitment and retention rates are sufficient to justify the need for the intervention.
The intervention model includes methods for increasing recruitment and retaining participants.
The local requirements for successful referral and recruitment are clear.
Information about barriers to successful recruitment and retention have been identified. Advice is available on how any barriers might be overcome.
Answers to the following two questions: What is known about the intervention's reach, including the participants who were not successfully recruited to the intervention? How are they similar or different to the intervention's target population?
There is information about the intervention's retention rates.

R	etention rates	are sufficient to	justify the need	for the intervention.
---	----------------	-------------------	------------------	-----------------------

Advice is available on how to increase retention rates.

- There is user satisfaction evidence that the intended target audience likes the intervention and finds the content to be worthwhile.
- Intervention participants typically report intervention benefits that are consistent with the intervention's ToC.
- Monitoring systems have been developed to gather information about delivery, recruitment, participant reach and retention.

#### **Further resources**

#### General

Asmussen, K. (2012). The evidence-based parenting practitioner's handbook. Routledge.

- Meyers, D. C., Durlak, J. A., & Wandersman, A. (2012). The quality implementation framework: a synthesis of critical steps in the implementation process. *American Journal of Community Psychology*, 50, 462–480.
- Wandersman, A., Duffy, J., Flaspohler, P., et al. (2008). Bridging the gap between prevention research and practice: the interactive systems framework for dissemination and implementation. *American Journal of Community Psychology*, *41*, 171–181.

#### Intervention costs

Edwards, R. T., Céilleachair, A., Bywater, T., Hughes, D. A., & Hutchings, J. (2007). Parenting programme for parents of children at risk of developing conduct disorder: cost effectiveness analysis. *British Medical Journal*, 334(7595), 682.

#### Case study 4 - Family Nurse Partnership

- Ball, M., Barnes, J., & Meadows, P. (2012). Issues emerging from the first 10 pilot sites implementing the Nurse-Family Partnership home-visiting programme in England. Department of Health.
- Barnes, J., Ball, M., Meadows, P., Howden, B., Jackson, A., Henderson, J., & Niven, L. (2011). The Family-Nurse Partnership programme in England: Wave 1 implementation in toddlerhood and a comparison between Waves 1 and 2a of implementation in pregnancy and infancy. Department of Health.
- Barnes, J. (2009). Nurse-family partnership programme second year pilot sites implementation in England the infancy period. Department of Health.
- Barnes, J. (2010). From evidence-base to practice: implementation of the Nurse Family Partnership programme in England. *Journal of Children's Services*, 5(4), 4–17.
- Barnes, J., & Henderson, J. (2012). Summary of the formative evaluation of the first phase of the group-based family nurse partnership programme. Department of Health.
- Barnes, J., & Stuart, J. (2016). The feasibility of delivering group Family Nurse Partnership. *Journal of Children's Services*, *11*(2), 170–186.
- Olds, D., Kitzman, H., Cole, R., & Robinson, J. (1997). Theoretical foundations of a program of home visitation for pregnant women and parents of young children. *Journal of Community Psychology*, 25(1), 9–25.
- Olds, D. L., Kitzman, H., Knudtson, M. D., Anson, E., Smith, J. A., & Cole, R. (2014). Effect of home visiting by nurses on maternal and child mortality: results of a 2-decade follow-up of a randomized clinical trial. *JAMA Pediatrics*, *168*(9), 800–806.

Robling, M., Bekkers, M. J., Bell, K., Butler, C. C., Cannings-John, R., Channon, S., ... & Kenkre, J. (2016). Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): a pragmatic randomised controlled trial. *The Lancet*, *387*(10014), 146–155.

# Step 5 Pilot for outcomes



# Pilot studies are relatively inexpensive evaluations which investigate an intervention's potential for improving its intended child outcomes.

Pilot studies are particularly useful for determining which measures are most appropriate for testing child outcomes, as well as how to best recruit and retain a sufficiently large and representative study sample. In Step 5, you will learn:

- the importance of validated measures and how to select and use them to measure preand post-intervention change
- methods for determining an adequate sample size based on the intervention's anticipated effects
- methods for recruiting and retaining participants from the intervention's target population
- methods for determining whether the intervention has brought about a statistically significant and positive child outcome
- how to interpret the findings from pilot studies and use them for designing more rigorous impact evaluations.

# The value of pilot studies

By definition, a pilot study is a preliminary and often small-scale investigation conducted to assess the feasibility of the methods to be used in a larger and more rigorous evaluation study. For this reason, it is not uncommon for evaluators to use the terms 'pilot study' and 'feasibility study' interchangeably to refer to evaluation designs that are not sufficiently rigorous to attribute causality to the intervention model (Step 6). However, from the perspective of the EIF evidence standards, the terms 'pilot study' and 'feasibility study' refer to two important but separate evaluation functions. While feasibility studies are primarily concerned with whether an intervention can achieve its intended *outputs*, pilot studies consider an intervention's potential for achieving its intended *outcomes*. Pilot studies therefore represent an important milestone in an intervention's development, as they are the first time the outcomes identified in the intervention's ToC will have been tested.

Pilot studies consider an intervention's potential for improving child outcomes through the use of validated measures that are developed to objectively assess changes in children's behaviour, feelings, thoughts or attitudes. Prior to this point, assessments of change might have been obtained through recipients' subjective recollections of what happened during the intervention, as described in Step 4. Although recollections are useful for understanding recipients' views, they offer little information about the amount of actual change that may have taken place during the course of the intervention. Pilot studies provide this information through the use of validated assessment tools that produce a reliable numerical value that can be used to measure the change occurring before and after the intervention. This value can be used to then test whether this change is statistically significant – meaning that the amount of change could not have occurred in the absence of the intervention.

Although pilot study findings are often subject to high levels of bias, a statistically significant positive result is frequently viewed to indicate that the intervention is ready for more rigorous testing. If this judgment is to be made with any confidence, however, the pilot study design should contain elements that help to ensure the findings are *objective* and *representative*. We recommend that at a minimum, these elements are consistent with the threshold for an EIF level 2 rating.

- Study objectivity is increased through the use of assessment instruments that have been
  psychometrically validated (see below) to measure the child outcomes identified in the
  intervention's ToC. Ideally, these measures should be completed once before participants
  receive the intervention and then again after the intervention is over, to measure pre- and
  post-intervention change. Psychometrically validated instruments can also be used to
  make comparisons between those who did and did not receive the intervention when pre/
  post assessments are not possible, although the intervention and comparison groups
  must be carefully matched.
- A study's representativeness is increased through the use of a sample that has the same characteristics as the intervention's target population and is sufficiently large to detect pre- and post-intervention change. The EIF level 2 threshold criteria requires that pilot studies involve no less than 20 participants who are representative of at least 60% of those originally recruited to the pilot study.

This chapter provides guidance on how to conduct pilot studies that meet the EIF level 2 threshold criteria. This includes advice on how to select measurement tools that are appropriate for the outcomes identified in the intervention's ToC, determine a sufficiently large sample size, collect data and run statistical tests for measuring change. This chapter also considers how pilot study findings should be used to inform the design of more rigorous evaluations, as described in Steps 6 and 7.

# Case study 5: A pilot study to test the potential efficacy of a brief format version of Triple P with Indonesian parents

The Triple P – Positive Parenting Programme® is an intervention system providing parents with strategies for reducing coercive family interactions (see case study 1) and encouraging positive child behaviour. The Triple P model is delivered through five tiers of increasingly complex interventions, ranging from brief, 90-minute seminars acquainting parents with Triple P's five principles of positive parenting (ensuring a safe environment, creating a positive learning environment, using assertive discipline, having realistic expectations, and taking care of oneself) to more intensive, individualised interventions aimed at reversing specific and entrenched family problems. Evidence from multiple RCTs conducted in Australia and other countries has consistently observed short-term improvements in children's behaviour when intensive versions of Triple P are made available to families experiencing specific problems with their children's behaviour (see case study 6).

This pilot study investigated the potential efficacy of a single 90-minute Triple P seminar delivered to Indonesian parents living in Australia, which aimed to increase their awareness and acceptance of the five Triple P parenting principles. The need for this intervention was identified through previous research observing that families living in Indonesia and other southeast Asian countries were at an increased risk of over-reactive and coercive family interactions and child behavioural problems. The primary aim of this study was to understand the potential of a brief version of the Triple P intervention for improving children's behaviour in Indonesian and other southeast Asian cultures.

### FIGURE 5.1

Short- and longer-term outcomes identified in a brief format version of Triple P for Indonesian parents



## Participants and sample size

All participants were recruited from the community surrounding the University of Queensland through mail-outs, flyers at Indonesian events and Facebook announcements. Parents were eligible if they were Indonesian and had a child between the ages of 2 and 12 who was living at home. Thirty-two eligible parents were recruited to the seminar and 27 attended on the day. Twenty-five parents completed validated psychometric questionnaires once before attending the seminar, three weeks afterwards and at a three-month follow-up. Thirty-two participants were recruited and 78% were retained. Analysis of the study sample's characteristics observed that all parents were married and highly educated, with over 43% having a postgraduate qualification. All research was carried out with the ethical approval of University of Queensland.

#### Measures

Three measures consistent with the Triple P seminar series' theory of change were used to measure pre/post intervention change.

- The Parent Acceptability Questionnaire (PAQ) was developed to measure parents' ratings of acceptability of the five positive parenting principles and determine specific learning outcomes of the Triple P programme.
- The Parenting Scale (PS) was chosen to consider changes in dysfunctional and coercive parenting behaviours, including behaviours that are too lax or over-reactive. The PS is a measure of ineffective parenting styles that includes subscales measuring overly permissive or lax behaviours, over-reactive parenting and verbosity (arguing). The PS has undergone extensive validity testing, demonstrating good internal consistency on all of its subscales with multiple populations around the world.
- The intensity subscale of the **Child Adjustment and Parent Efficacy Scale (CAPES)** was selected as a brief, low-burden measure to track improvements in the behaviour and emotional adjustment of children without clinical problems. The validity of the CAPES was established with Australian and Indonesian populations at the time it was first developed in 2013.

All three measures were administered in the form of questionnaires that could be completed in person, or online. These questionnaires were completed at three time points: before the start of the intervention, at a three-week follow-up, and then again at a further three-month follow-up. Demographic information was also collected at the time of recruitment and participants completed online user satisfaction surveys three weeks following the end of the seminar.

## Findings

T-tests (see appendix D) comparing pre- and post-intervention change confirmed statistically significant improvements in some, but not all aspects of the evaluators' ToC. In particular, the study did not observe any increases in parents' acceptance of the five Triple P parenting principles, as the majority of parents already endorsed these principles. Nevertheless, the study did observe *moderate* and *statistically significant* reductions in overly permissive parenting practices and child behavioural problems. The authors interpreted these findings to mean that a single seminar was potentially sufficient for Indonesian parents to learn strategies that would in turn help improve their children's behaviour.

The study did not, however, confirm reductions in over-reactive parenting behaviours. The authors noted this lack of a positive finding may have been related to the possibility that Indonesian parents may have greater difficulty managing angry emotions when dealing with challenging child behaviours. The authors therefore speculated that one seminar session covering the management of angry emotions may not have been sufficient to change over-reactive parenting styles. This conclusion was consistent with the feedback received from the parents in the user satisfaction surveys, requesting additional advice and support on how to implement various parenting strategies. In the end, the authors concluded that the intervention was ready for further, more rigorous testing after refinements to the original content were made.

# Selecting appropriate assessment measures

Positive pilot study findings are contingent upon the use of assessment measures that are appropriate for the child's age and consistent with the outcomes identified in the intervention's ToC. For example, the assessment tools used in case study 5 were chosen specifically to measure each of the short-term outcomes identified in the intervention's ToC, including improved parenting practices and reduced child behavioural problems.

If a literature review was conducted to inform the intervention's ToC in Step 1, the cited studies should provide information about measures which are suitable for assessing the child outcomes identified in the intervention's ToC. If relevant measures were not identified in Step 1, Step 5 is the time when this information must be done. As described in EIF's *Six common pitfalls and how to avoid them*,<sup>8</sup> **statistically significant evaluation results are highly unlikely when there is a poor fit between the intervention's ToC and the evaluation measures chosen to investigate pre- and post-intervention change.** 

# Validity

Measures are unlikely to yield a meaningful score unless they have been verified to be *valid*. This means that measurement tools have been psychometrically tested to have *construct validity*, meaning that they can consistently measure the behaviours, attitudes, feelings or thoughts (collectively referred to as constructs) that they were developed to measure.

<sup>&</sup>lt;sup>8</sup> See: https://www.eif.org.uk/resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

Establishing construct validity is a complex process, involving a series of statistical analyses to verify the *internal*, structural validity of the measure (for example, is there a high degree of convergence between all items which aim to measure the same construct?) and its *external* consistency with other measures. Establishing structural validity begins with a large pool of assessment items that is then gradually reduced to a smaller and more manageable set of questions that are internally consistent with each other. Many measurement tools are made up of subscales that consistently capture one or more of the constructs of interest.

While it is not particularly costly to verify the internal structural validity of a measure, determining the measure's external consistency with other measures requires significantly more resources, as subjects often must be recruited to participate in coded observations. For this reason, fewer measures have external construct validity that has been verified against observations of actual participant behaviours. For example, the CAPES measure, developed in 2013, has not yet undergone external validation against participant behaviour. It has therefore been judged as having moderate construct validity by independent reviewers. While lower construct validity does not mean that the measure cannot be used, it does mean that greater care should be taken when interpreting the measure's findings.

# Reliability

*Reliability* pertains to a measurement tool's ability to elicit the same response when repeated multiple times. Measurement tools must have established test-retest reliability to provide a valid estimate of pre- and post-intervention change. For example, a reliable measure of empathy should provide a similar, if not identical measure of empathy every time it is administered, unless something significant – like an intervention – changed the participants' level of empathy. Issues with reliability occur when measurement items are ambiguous or are subject to fleeting changes in study participants' moods. Reliability should therefore be established at the time the measure is first developed through the pool reduction process described in the previous section.

## Sensitivity

Measures that are both valid and reliable are more likely to be sensitive to pre- and postintervention change. Measurement sensitivity (also referred to as responsiveness) is particularly vital when small samples are investigated, which is often the case in pilot studies. Factors contributing to measurement sensitivity include (1) test items that are clear and specific, (2) more rather than fewer items that examine the full range of construct being investigated, and (3) an increased range of responses. For example, a five-item scale will likely be more sensitive than a yes/no question. As described in the previous section, measures should have also undergone some testing to ensure that they are valid and reliable for the population under investigation. This means that measures are also appropriate for the child's age and have been validated as culturally relevant for the target population.

It is worth noting that psychometric measures are developed to be parsimonious, meaning that the individual scales do not generally include any more items than are absolutely necessary for the measure to remain valid. This means that while it may be tempting to remove test items to reduce the length of a measure to make it easier to administer, modifying or removing items within individual scales stands to significantly reduce the measure's validity, reliability and sensitivity. **Researchers are therefore strongly encouraged not to modify validated measures, unless they are prepared to carry out additional statistical tests to ensure that item removal does not interfere with the validity of the measure. Even if these tests verify that validity has been retained, researchers should be aware that the sensitivity of the measure will likely be reduced, as sensitivity is often associated with the length of the measure. In situations where less cumbersome measures are required, it is sometimes feasible to administer a fewer number of subscales from a** 

validated measure. For example, in case study 5, the intensity subscale of the CAPES was substituted for the entire measure to assess children's behaviour.

## **Practical considerations**

A wide variety of psychometrically valid assessment tools are available to test a diverse range of child outcomes. However, these measures vary in terms of their length and ease of administration. As a rule of thumb, measures are more valid when they are administered in a *standardised* way, meaning that they are administered the same way every time.

Standardised measures are frequently administered in person, either by a researcher or practitioner who has undergone training to administer the measure. This training no doubt increases the precision and sensitivity of the measure, but also creates an additional burden in the administration of the test. These difficulties in turn can delay the completion of the study and potentially weaken the validity of the findings.

Less burdensome measures, such as those described in case study 5, include short-item questionnaires that can be completed by the study participants online. Such instruments are clearly much easier to complete from a practical standpoint, although they are frequently less sensitive than longer questionnaires or measures that are administered in person. The extent to which online or pen and paper questionnaires accurately measure meaningful changes in participants' behaviours is also less clear, unless the measure has been externally validated against actual behaviours as described in the previous section.

In some instances, comparing pre- and post-intervention scores will not be feasible. For example, the rapid pace of infant development makes pre/post measurement with the same test impossible. In these instances, pilot studies will require *norm-referenced* measures (see figure 5.2) that have undergone rigorous validation to permit comparisons between children's development and an established population average.

Pre- and post-intervention change can then be assessed by comparing children's relative standing against these norm-referenced averages. If used in a pilot study, comparisons will only be meaningful if they are made against a sample of similar children who did not receive the intervention. More information about recruiting appropriately matched comparison samples is provided in Step 6.

# Recruiting a sufficiently large and representative sample

## Representativeness

A study's representativeness, also referred to as its *external validity* (not to be confused with the external consistency of measures, as described in the previous section), refers to the extent to which its findings are generalisable to circumstances and populations outside of the study's context. It is difficult to understand the extent to which a study's findings can be replicated, unless the study was conducted with a sample that shares the same characteristics as the intervention's target population. For example, an aim of the Triple P study in case study 5 was to verify the potential of the intervention for improving parenting practices in non-western cultures. While the study made sure that only Indonesian parents participated, the authors concluded that the findings were potentially unrepresentative of families living in Indonesia, as the study participants were highly educated.

Study representativeness can also be undermined when there are high levels of study attrition. This is because high levels of attrition can substantially reduce the sample size and introduce bias, as attrition is frequently associated with participant characteristics. As described in case study 2, attrition was higher among more vulnerable families, thereby

reducing the generalisibility of the findings for those who were most disadvantaged. More information about how study attrition can undermine confidence in a study's finding is provided in Step 6 and in EIF's *Six common pitfalls and how to avoid them*.

## FIGURE 5.2

Norm-referenced measures permit comparisons against an average



Source: Renaissance (2018) 'What's the difference? Criterion-referenced tests vs. norm-referenced tests', blog, 11 July 2018. Available: https://www.renaissance.com/2018/07/11/blog-criterion-referenced-tests-norm-referenced-tests/

## Sample size

A study's representativeness increases with sample size. Although a positive outcome observed in a single participant could be representative of what might be observable in a larger population, it is impossible to know this with any certainty. This is because it is not possible to differentiate whether the result was associated with the intervention, characteristics of the individual or random occurrence. However, if the same finding is observed among many participants, one can be more confident that the findings are representative.

Larger samples are also necessary to ensure that a study is sufficiently 'powered' to detect a statistically significant finding. Statistical power refers to a study's ability to avoid a false negative, or 'type II' error, which involves concluding that the intervention did not make a measurable difference when it actually did. The risk of type II errors is higher in studies involving small samples, because pre/post differences are difficult to detect unless they are relatively large. The size of pre/post difference is referred to as the *magnitude of impact* and expressed in terms of an *effect size* which is conventionally understood to be small, medium or large. More information about calculating power on the basis of effect sizes can be found in appendix D.

When an intervention is first developed, it is difficult to anticipate the magnitude of impact, so making the risk of a type II error quite high in pilot studies involving small samples. It is possible to reduce this risk by estimating what potential effect size might be and then recruiting a sufficiently large sample accordingly. So, if one was confident of a very large effect size, a sample of 20 could be sufficient, although researchers should be aware that 20 is likely to be inadequate for interventions that anticipate small to moderate improvements in child outcomes.

We recommend that evaluators consult the following resources when first estimating the intervention's potential effect size:

- effect sizes in published systematic reviews of impact studies or similar interventions
- · effect sizes in pilot studies
- expert opinion regarding what is a meaningful or realistic effect size in the field.

In the absence of comparable effect size information, developers might anticipate their intervention's potential effect size in terms of the following three parameters.

- 1. The needs of the intervention's target population: As described in Step 1, population needs are understood on a continuum of low to high. An intervention's effect size is therefore likely to be larger if the population needs are high, or if the intervention aims to treat a pre-identified condition.
- 2. The intensity of the intervention: Highly intensive interventions offered to high-need families (see case study 2) tend to have higher impacts and thus higher effect sizes. Lower-dose interventions tend to result in lower effect sizes.
- 3. The extent to which the intervention adds value over what is currently available: For example, the Triple P seminar described in case study 5 was low dose yet expected to provide a moderate impact given that parenting advice of this nature was typically not offered through other services or intervention.

#### Study recruitment

Interventions are not likely to be ready for Step 5 until they can consistently recruit and retain a sufficient number of participants, as described in Step 4. It is up to those commissioning and delivering the intervention to determine what a sufficient recruitment rate should be, but it goes without saying that recruitment and retention should be sufficiently high for an intervention to be viewed as feasible and providing value for money.

From the standpoint of evaluation, a pilot study should be able to recruit a sample that is large enough to detect a statistically significant pre- and post-intervention impact and retain enough participants to ensure that the findings are representative of the target population. The EIF level 2 threshold for study retention is 60%. With a minimum sample size requirement of 20 participants, this assumes that a study that recruits only 20 participants will need to retain all 20. A minimum sample size of 34 is required if attrition is as high as 40%.

Recruiting and retaining study participants creates challenges that are independent from intervention recruitment. For example, research studies with parents and children will always require ethical approval from the local council, as well as the research organisation affiliated with the evaluator (such as a university) or professional body (for example, all studies that recruit families through health services require NHS approval). Ethical approval is typically contingent upon the perceived need for the study, the methods used to obtain informed consent, the extent to which data collection is not overly burdensome, the risk of unintended or negative consequences, safeguarding considerations and compliance with data protection standards. More details on ethical approval requirements are provided in the further resources section at the end of this chapter.

Once the study has gained ethical approval, participants must provide their consent and be available to participate in the study independently of the intervention before and after it takes place. Depending on the requirements of the measures, participants may need to meet with the researcher in person, unless measures can be easily completed online. Regardless, it is important that data collection takes place independently of the intervention. Independent data collection not only reduces the likelihood of conflicts of interest, but also improves data quality.

Study retention can be enhanced by rewards or incentives for participants to remain in the study. For example, it is not uncommon for participants to be offered gift vouchers involving nominal amounts as a thank you for participating in the study. Vouchers are especially helpful if measurement completion is time-consuming or continues months after the intervention has been completed. More information about recruiting and retaining study participants is provided in Step 6.

# **Measuring impact**

A variety of statistical analyses are appropriate for measuring impact. While it is beyond the scope of this guidance to provide advice on how to perform all of these analyses, we advise that for the first pilot study researchers make it as simple as possible, keeping in mind that the analyses must be appropriate for the nature of the data and the size of the sample. Further details about data analysis for small samples is provided in appendix D.

# Statistical significance

The aim of any statistical analysis is to observe whether the relationship between two phenomena is statistically significant, meaning that any observed association is not likely to have occurred by chance. When it comes to testing the effectiveness of interventions, the term 'chance' frequently refers to what may have occurred in the absence of the intervention.

Statistical tests are thus designed to avoid erroneous assumptions on the basis of random observations. In statistical language, this involves avoiding a mistaken decision to accept or reject the *null hypothesis*. The null hypothesis is the default assumption that no relationship exists between two or more phenomena (referred to in this guidance as the intervention and any observed child outcomes). Erroneous decisions to accept or reject the null hypothesis are classified as type I or type II errors.

- **Type I errors** involve mistakenly rejecting the null hypothesis by assuming that a relationship between two phenomena exists when it, in fact, does not.
- **Type II errors** involve accepting the null hypothesis, by assuming no relationship between two or more phenomena exists when it, in fact, does.

Statistical tests set parameters to reduce the likelihood of type I and II errors. Convention dictates that an acceptable level of a type I error risk is one out of 20 (.05). However,

researchers should be aware that the acceptability of this cut-off should be fundamentally informed by the distribution of the data and the potential risks created by type I and type II errors. For the purposes of a pilot study, however, a significance level at .05 (in other words, setting the risk of a type I error at 5%) is generally acceptable.

# Interpreting significant findings

The EIF evidence standards require that a pilot study must have observed a statistically significant improvement in at least one of the intervention's primary child outcomes to qualify for a level 2 rating. As this chapter makes clear, this is a substantial achievement and is viewed by many as a *preliminary* indicator that the intervention will prove to be effective after more rigorous testing has occurred. However, **positive findings from a pilot study should never be interpreted to mean that the intervention is effective.** This is because pilot study findings are subject to high levels of bias that can dramatically inflate the size of a positive result, even if the study involves a small sample. Small samples not only increase the risk of type II errors, but also the possibility that any statistically significant finding is actually false. Small samples also reduce the representativeness of the findings.

The lack of an equivalent comparison group in most pilot studies additionally makes it impossible to understand what might have happened in the absence of the intervention. Researchers should therefore remember that it is not uncommon for participant circumstances to improve even when they don't participate in interventions. For these reasons, promising findings obtained in initial pilot studies are often not replicated under more rigorous evaluation conditions as described in Steps 6 and 7.

# Interpreting non-significant findings

Failure to observe a statistically significant finding means that the intervention is ineligible for an EIF level 2 rating. However, a lack of a statistically significant result should not necessarily be interpreted to mean that the intervention is fundamentally ineffective. Rather, disappointing pilot study findings often mean that revisions to the intervention's content or evaluation design are necessary.

Revisions to the intervention's content frequently involve revisiting the intervention design by addressing three questions that were first considered when the ToC was confirmed.

- Who is the intervention for? Interventions that fail to observe a positive outcome through pilot research frequently will not have fully specified the needs of their target population. For example, the researchers in case study 1 assumed that all parents would benefit from parenting advice. It was only through ongoing testing that the researchers discovered that parenting interventions would have their greatest impact when provided to more vulnerable families experiencing specific parenting problems.
- Why will the intervention add value? Interventions sometimes fail to demonstrate a positive impact because they are not necessary. Returning to the example in case study 1, pilot testing confirmed that some information provided by parenting interventions does not provide much value over what many non-vulnerable parents already knew.
- How much of the intervention is needed? It is not uncommon for developers and providers to underestimate how much of an intervention is required for participants to learn and master the content in a way that is sufficient to improve child outcomes. For example, the Triple P seminar discussed in case study 5 may have been sufficient for parents to learn how to implement the key principles associated with lax parenting, but it was unlikely sufficient for parents to learn how to control over-reactive parenting behaviours.

It is also not uncommon for interventions to increase recipients' awareness of various problems, which may result in *increases* in reports of problems or negative perceptions.

Evaluators thus sometimes interpret this negative result to mean that the intervention will nevertheless result in improved child outcomes over time, once recipients have time to overcome their negative views. While this may be true, we strongly caution against this interpretation of negative findings. Instead, we recommend that developers address them through adjustments to their intervention's content to help recipients manage any negative views that may have developed during the course of the intervention.

Disappointing findings also occur because of weaknesses in the pilot study design. As described in previous sections, these weaknesses frequently involve issues pertaining to measurement or sample size. For example, the study may have failed to observe a positive result because the measures were not sensitive enough to detect pre/post intervention change. A second pilot study involving different measures may therefore be necessary.

It is also not uncommon for a study to fail to detect a significant result because the sample size was too small. It is thus worth keeping in mind that small to moderate impacts often become statistically significant once the sample size is increased. However, developers and commissioners still need to consider whether the size of the impact is worth the cost of the intervention. Interventions resulting in small to modest effects are likely to be worth the time and money if they are relatively inexpensive. However, an expensive intervention resulting in a small-to-modest effect size should be strongly reconsidered.

A final reason an intervention may fail to produce a significant positive result is that it may not have been implemented with fidelity – in other words, it was not delivered as it was intended. This problem might have been avoided had sufficient work been done during Step 4 when the intervention's feasibility was tested. However, a failure to implement the intervention with fidelity is not uncommon when sufficient feasibility testing has not taken place. More advice on how to increase intervention fidelity is provided in Step 8.

# Exceeding the minimum: The use of a comparison group and other pilot study alternatives

In this chapter, we have described what we view to be the *minimum* requirements of a pilot study to consider an intervention's potential impact for improving child outcomes. We must emphasise that these are *minimum* requirements, and strongly encourage evaluators to go beyond these requirements to more fully explore their intervention's potential. Useful methods that exceed these requirements include the use of a larger sample and measuring pre- and post-change in a comparison group who did not receive the intervention. More information about how to conduct a comparison group study is described in Step 6.

If resources permit, evaluators may also want to pilot more, rather than fewer measures to make sure they have identified assessment tools that are sensitive enough to detect change in their outcome of interest. In this respect, pilot studies provide an important opportunity for evaluators to identify the best measure for assessing their outcome of interest, as the findings from more rigorous studies are often viewed as less valid if they employ too many measures, as we describe in Step 6.

# **Step 5 checklist**

The minimum requirements for a pilot study that is consistent with the threshold requirements for an EIF level 2 rating:

The study has obtained ethical approval from the appropriate governing bodies.
The study tests a clearly stated hypothesis that is consistent with the intervention's ToC.
The study involves a minimum of 20 participants that are representative of the intervention's target population.
The study report includes the information about the characteristics of the study participants, including their demographic details and extent to which they meet the eligibility requirements.
The study has retained at least 60% of the originally recruited participants.
The study report includes details of the participants lost to follow-up, so a judgment can be made about the study's representativeness.
The study made use of psychometrically validated measures that are appropriate for the child's age and participant characteristics.
The study makes use of statistical analyses that are appropriate for the nature of the data. The data analysis and findings are clearly reported.
The study has observed a statistically significant positive result for at least one if its intended child outcomes.

## **Further resources**

#### **General information**

Asmussen, K. (2011). The evidence-based parenting practitioner's handbook. Routledge.

- Foster, E. M., Dodge, K. A., & Jones, D. (2003). Issues in the economic evaluation of prevention programs. *Applied Developmental Science*, 7(2), 76–86.
- Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, *13*(1), 16–23.
- Martin, J., McBride, T., Brims, L., Doubell, L. Pote, I. and Clarke, A. (2018) *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Early Intervention Foundation. Available: https://www.eif.org.uk/ resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

Reeves, B. C., Wells, G. A., & Waddington, H. (2017). Quasi-experimental study designs series – paper 5: a checklist for classifying studies evaluating the effects on health interventions – a taxonomy without labels. *Journal of Clinical Epidemiology*, *89*, 30–42.

#### Further resources about developing and using psychometrically validated assessment tools

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281.

- Darlington, R. B., & Bishop, C. H. (1966). Increasing test validity by considering interitem correlations. *Journal of Applied Psychology*, 50(4), 322.
- Fok, C. C. T., & Henry, D. (2015). Increasing the sensitivity of measures to change. *Prevention Science*, 16(7), 978–986.
- Fok, C. C. T., Henry, D., & Allen, J. (2015). Maybe small is too small a term: Introduction to advancing small sample prevention science. *Prevention Science*, *16*(7), 943–949.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741.
- Roach, K. E. (2006). Measurement of health outcomes: reliability, validity and responsiveness. *Journal of Prosthetics and Orthotics*, *18*(6), 8–12.

Souza, A. C. D., Alexandre, N. M. C., & Guirardello, E. D. B. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e Serviços de Saúde*, 26(3), 649–659.

Symonds, P. M. (1928). Factors influencing test reliability. Journal of Educational Psychology, 19(2), 73.

#### Information about valid and reliable measures

Asmussen, K., Feinstein, L., Martin, J., & Chowdry, H. (2016) *Foundations for Life: What works to support parent-child interaction in the early years*? Early Intervention Foundation. Available: https://www.eif.org.uk/report/foundations-for-life-what-works-to-supportparent-child-interaction-in-the-early-years

- Table 6: Measures commonly used to assess children's attachment security and attachment related outcomes (p. 65)

 Table 7: Measures frequently used to assess changes in children's behaviour and parenting strategies (pp. 93–94)

- Table 10: Measures frequently used to assess parent and child outcomes associated with cognitive and language development (pp. 131–133)

Child Outcomes Research Consortium: https://www.corc.uk.net/

California Evidence-Based Clearinghouse for Child Welfare's list of measures which they have reviewed: http://www.cebc4cw.org/assessment-tools/measurement-tools-highlighted-on-the-cebc/

Martin, J., McBride, T., Brims, L., Doubell, L. Pote, I. and Clarke, A. (2018) *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Early Intervention Foundation. Available: https://www.eif.org.uk/ resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them – Pitfall 4: Using inappropriate measures (pp. 18–20)

The Education Endowment Foundation's assessment of the validity and reliability of social and emotional skills measures: https://educationendowmentfoundation.org.uk/our-work/resources-centre/measuring-essential-skills/spectrum-database

#### Further information about participant recruitment and sample size

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.

Dumville, J. C., Torgerson, D. J., & Hewitt, C. E. (2006). Reporting attrition in randomised controlled trials. *British Medical Journal*, 332(7547), 969–971.

#### Further information about obtaining ethical consent

Association for the Directors of Children's Services research governance: http://adcs.org.uk/assets/ documentation/reason\_Research\_Governance\_Guide\_FINAL.pdf

ESRC. Research with children and young people: https://esrc.ukri.org/funding/guidance-for-applicants/researchethics/frequently-raised-topics/research-with-children-and-young-people/

NHS. Applying to a research ethics committee. https://www.hra.nhs.uk/approvals-amendments/what-approvalsdo-i-need/research-ethics-committee-review/applying-research-ethics-committee/

NSPCC. Research with children: ethics, safety and avoiding harm. https://learning.nspcc.org.uk/researchresources/briefings/research-with-children-ethics-safety-avoiding-harm/

The Research Ethics Guidebook: http://www.ethicsguidebook.ac.uk/Research-with-children-105

#### Further information about case study 5

Arnold, D. S., O'Leary, S. G., Wolff, L. S., & Acker, M. M. (1993). The Parenting Scale: a measure of dysfunctional parenting in discipline situations. *Psychological Assessment*, 5(2), 137.

Morawska, A., Sanders, M. R., Haslam, D., Filus, A., & Fletcher, R. (2014). Child adjustment and parent efficacy scale: Development and initial validation of a parent report measure. *Australian Psychologist*, 49(4), 241–252.

Sumargi, A., Sofronoff, K., & Morawska, A. (2014). Evaluation of a brief format of the Triple P-Positive Parenting Program: a pilot study with Indonesian parents residing in Australia. *Behaviour Change*, *31*(2), 144–158.

Wittkowski, A., Garrett, C., Calam, R., & Weisberg, D. (2017). Self-report measures of parental self-efficacy: A systematic review of the current literature. *Journal of Child and Family Studies*, 26(11), 2960–2978.

# Step 6 Test for efficacy



# An efficacy study is a rigorous evaluation designed to determine if an intervention works under ideal circumstances.

Efficacy studies do this through research designs that systematically reduce all sources of potential study bias, so that causality can confidently be attributed to the intervention model.

In Step 6 you will learn:

- · how to determine whether an intervention is ready for an efficacy study
- the ways in which potential sources of bias can 'threaten' the validity of a study's findings
- how a comparison group and methods such as random assignment can be used to reduce potential sources of study bias
- · strategies for reducing sources of potential bias throughout the duration of the efficacy study
- strategies for increasing the likelihood that the study will take place under ideal circumstances
- how to interpret findings from efficacy studies
- what to do when a rigorously conducted efficacy study fails to observe any positive effect on a child outcome of interest.

# Are you ready for Step 6?

Step 6 represents a monumental step in an intervention's development, as it is the point at which the intervention undergoes an evaluation that can determine whether it is *efficacious* – in other words, whether the intervention 'works'. This means that the evaluation is conducted in a way that minimises all sources of potential bias so that the extent to which the intervention *causally* improves child and family outcomes can be determined. A Step 6 evaluation is therefore considered by many to be the first litmus test of the intervention's ToC, because it not only verifies *whether* the intervention has made a positive difference, but *how much* of a difference it has actually made.

As we describe in this step, efficacy trials contain two features that make them substantially more rigorous than the evaluations described in the previous steps.

- First, efficacy trials include elements that increase their internal validity in other words, increase the confidence with which a judgment of causality can be attributed to the intervention model. These features involve processes such as random assignment that allow the study to systematically rule out (or control for) all sources of known and unknown bias that might otherwise limit one's confidence in the study's results.
- Second, efficacy trials are conducted in a way that ensures that the intervention is delivered the highest possible standard to maximise the size of its potential impact. Efficacy trials are therefore sometimes referred to as 'proof of concept' studies to determine that the intervention model is causal and its full potential is understood.

The design features of efficacy trials make them expensive and time-consuming. Interventions must therefore be *ready* to take Step 6 to ensure that they will provide valid and meaningful results.

## FIGURE 6.1

No skipping steps: interventions that have not undergone sufficient feasibility and pilot testing are more likely to fail when they complete their first efficacy study



#### Source: EIF

We recommend that evaluators and developers are able to confidently answer 'yes' to the following eight questions before they consider themselves ready to take Step 6.

- 1. Does the intervention aim to improve outcomes that are important for children's development?
- 2. Does the intervention model have a clearly defined target population that specifies the age of the children and their level of need?
- **3.** Is the intervention likely to provide measurable value for this population over what is currently available?
- 4. Can the intervention be feasibly delivered to a high standard?
- **5.** Are the intervention's costs known and are they proportionate to the intervention's intended outcomes?
- **6.** Is it possible to recruit participants from the target population at numbers sufficient to justify the cost and time of the intervention?
- **7.** Is it possible to retain participants from the target population at numbers sufficient to justify the cost and time of the intervention?
- **8.** Does the intervention have preliminary evidence from a pilot study meeting the minimum Step 5 requirements of improving at least one of its intended child outcomes?
If the answer to any one of these questions is 'no', there is a high probability that a rigorous efficacy study will fail to observe a statistically significant positive impact. This is either because there is an increased potential that the intervention's impact will be too small to detect, or that the study will contain flaws that will limit confidence in any judgment of causality. Processes that limit the validity of intervention findings include difficulties in recruiting or retaining a sufficiently large sample or implementing the intervention with fidelity.

The methods covered in Steps 1 to 5 of this guidance are designed to help evaluators and developers provide a confident 'yes' answer to our eight readiness questions. We therefore strongly recommend that interventions do not undertake an efficacy trial unless they have successfully completed Steps 1 to 5 of this guidance. It is not only wasteful to conduct an efficacy study when the intervention is not ready, but also unethical from the perspective of misusing study participants' time on a potentially futile endeavour.

# Increasing validity through an efficacy trial

Efficacy trials facilitate judgments of causality through experimental mechanisms that aim to increase the study's internal validity – in other words, reduce any biases (also known as *threats to internal validity*) that could potentially skew the study's findings in one direction or another. A list of common threats to internal validity and associated evaluation design features that aim to reduce them is provided in appendix E.

The use of a comparison (control) group can dramatically increase a study's internal validity if the group is equivalent to the intervention group in all respects, except that the control group does not receive the intervention. This way, it can be assumed that any positive changes observed within the intervention group are attributable to the intervention, rather than characteristics associated with the intervention group participants. If the changes within the intervention group are significantly greater than those in the control group, it can be assumed that the intervention has made a causal difference.

## Maintaining group equivalence through random assignment

A confident judgment of causality is only possible if the intervention and comparison groups are equivalent at the start of the study. Group equivalence is increased through methods (such as a coin toss or computer-generated numbers) that randomly assign participants to the intervention and control group. Random assignment, combined with a sufficiently large sample, thereby ensures that all known and unknown biases are randomly, and presumably equally, distributed across the intervention and control groups (see figure 6.2). Alternatives to random assignment include quasi-experimental designs that identify equivalent treatment and comparison groups within large sample sets. More information about random assignment and other quasi-experimental methods is provided in the further resources section at the end of this chapter.

### FIGURE 6.2

Non-random assignment versus randomisation: creating equivalence and preventing bias



Our hypothetical study is to evaluate the impact of a programme designed to reduce children's risk of experiencing poor mental health. Imagine that children shown by the gold circles have a lower risk while children shown by purple circles have a higher risk. In the first case, the sample is analysed so that we understand the pre-existing levels of risk, and the randomisation process ensures that the intervention and control groups are similar, by this key characteristic – that is, they are equivalent. At the end of the study, researchers find a lower risk rate among the intervention group, and can reliably conclude that this is due to the programme being effective.



In the second case, the sample is taken from two groups (perhaps from two different locations, clinics or schools). Instead of randomly assigning individuals from across the sample to the intervention and control groups, however, the researchers use a non-randomised method – perhaps they allow one group to volunteer to receive the programme, or assign one whole group to receive it. At this point, there may be significant pre-existing differences in the level of risk between the two groups – we cannot be confident that the two groups are similar, by this key characteristic. At the end of the study, researchers find a difference in risk level between the two groups – but in this case, we cannot be confident that this difference is due to the effectiveness of the programme rather than pre-existing differences, or bias.

### Source: EIF

Case study 6 describes how an efficacy trial involving three versions of Triple P made use of random assignment to create four groups that were equivalent in terms of all of the participant characteristics that were measured. These characteristics included demographic factors, as well as the extent to which the participants met the intervention's eligibility criteria. Random assignment thereby ensured that a primary threat to internal validity – selection bias – was eliminated from the outset of the study.

## Case study 6: An efficacy trial of Standard Triple P

### The intervention and faithful delivery

Triple P, first introduced in case study 5, is a suite of parenting interventions which aim to prevent child emotional and behavioural problems by enhancing parents' knowledge, skills and confidence in managing children's behaviour. The efficacy study described here aimed to understand the relative impact of three versions of the Triple P model – Enhanced Triple P, Standard Triple P and Self-directed Triple P – in comparison to no intervention.

All three versions of the Triple P model tested in this study are designed specifically for parents experiencing difficulties with their child's behaviour. In the Standard version, parents receive face-to-face support from a master's level psychologist through 10 individual one-hour sessions. These sessions cover 17 different strategies for managing their child's behaviour through role play, homework exercises and discussions involving videotaped examples of effective parenting.

Enhanced Triple P is the same as Standard Triple P, except that it contains two additional sessions covering parental coping strategies and couple support. The self-directed version of the intervention was given to the parents in the form of a workbook that could be completed over the same 10-week period as the intervention.

The practitioners delivering the programme received high levels of training and implementation support to ensure that both face-to-face versions of the intervention were delivered to a high standard. This support included the use of fidelity checklists completed by the practitioners at the end of each session to verify that all material was covered. Sessions were also videotaped and reviewed during supervision sessions with their managers. Fidelity to both models was high, with 100% of practitioners covering all the relevant content and homework assignments with all parents.

### Participants

The study recruited Australian families with a preschool child at risk of developing a conduct disorder. Families were recruited through flyers, newspaper articles and other community activities that encouraged parents to self-refer if they were experiencing problems with their child's behaviour. As summarised in the participant flow diagram in figure 6.3, 940 participants expressed interest in participating in the intervention and 724 returned screening questionnaires. Telephone calls were then used to verify that the parents met the intervention's eligibility criteria. The eligibility requirements assumed that parents would be experiencing clinical-level problems with the behaviour of a child between the ages of 3 and 4, as well as at least one additional risk factor. These risk factors included parental depression, marital dissatisfaction, single-parenthood or low family income. Parents also had to report that they were not receiving any other forms of support for their child's behaviour or other family issues. The study identified 381 families as eligible and 305 were successfully recruited to the study (see figure 6.3 below).

The successfully recruited 305 families were then randomly assigned to one of four treatment conditions: Enhanced Triple P (76), Standard Triple P (77), the self-directed version of Triple P (75), and a waitlist control group (77). The waitlist design meant that those in the control group would receive the intervention after the intervention pre/post measures were completed for the three treatment groups. Statistical tests verified that the characteristics of the participants in the intervention and comparison groups were equivalent before the interventions were delivered.

### Measurement

Intervention participants underwent assessments at four time points: before the start of the programme, immediately after intervention completion, one year after intervention completion and three years after intervention completion. The wait list control group did not participate

during this last assessment, as they had received intervention by that point and therefore no longer provided a valid comparison.

Parents completed a battery of validated measures of parenting practices and disruptive child behaviour at the first three assessments. These measures included the Eyberg Child Behaviour Inventory that contained two scales involving the frequency and intensity of disruptive child behaviour and the Parenting Scale described in case study 5 that included scales of overly lax and reactive parenting behaviours. The parents and their children also underwent an observational assessment at all three time points, where interactions were coded by trained independent observers during a series of difficult tasks using the validated Revised Family Observation Schedule.

At the three-year follow-up, children's behaviour was assessed through teacher ratings gathered through the Sutter–Eyberg Student Behaviour Inventory and diagnostic clinical interviews with the children's parents.

### FIGURE 6.3

Participant flow chart of study attrition at post-intervention, one-year follow-up and three-year follow-up



Source: Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The Triple P-Positive Parenting Program: A comparison of enhanced, standard, and self-directed behavioural family intervention for parents of children with early inset conduct problems. *Journal of Consulting and Clinical Psychology*, 68(4), 624–640.

### Attrition

Of the 228 families assigned to the three active intervention conditions, 183 (80%) were retained at the post-intervention time point, with 75% remaining in Enhanced Triple P, 83% remaining in Standard Triple P, 81% remaining in Self-directed Triple P. As a result, there were no significant differences in the completion rates for all intervention groups, although rates of attrition varied substantially from the waitlist group, which retained 92% of its participants.

The authors compared families who did not complete the post-intervention measurement with those who did, observing that those who left the study were significantly more likely to report severe problems with their child's behaviour than those who remained.

The authors further observed significant differences in the characteristics of the parents who left the intervention conditions in comparison to those in the waitlist control group.

In particular, mothers who left the enhanced and self-directed groups at post-test were significantly more likely to report feelings of negative affect at baseline than those leaving the other two groups. Conversely, mothers with higher levels of negative affect and disagreement with their partners were more likely to remain in the waitlist control group.

Statistical analyses were used to compare change within and between the four treatment groups. Participants were retained in the analysis regardless of their level of participation, in keeping with an 'intent to treat' design.

### Results

- **Post-intervention:** Immediately after intervention completion, significant improvements in children's behaviour and parenting practices were observed in the participants of all three Triple P interventions in comparison to the waitlist control group. However, the Enhanced and Standard versions positively impacted a greater number of parent and child outcomes. Most notably, both the Enhanced and Standard versions observed significant improvements in the coded observations of actual child behaviours, whereas the self-directed version did not.
- One-year follow-up: Comparisons between the control group and the three Triple P
  programmes were no longer possible at this point, as the control group had subsequently
  participated in the intervention. However, pre/post analyses for the three treatment groups
  indicated that improvements in the children's behaviours were sustained for the Enhanced
  and Standard Triple P groups, although the difference in impact between the two groups
  was not significant.
- **Three-year follow-up:** Positive effects were sustained at the three-year follow-up for all three of the treatment groups. Effect sizes were reported for change over time, with the impact on children's negative behaviour increasing in magnitude. Interestingly, the pronounced differences in impact between the face-to-face versions of Triple P (Enhanced and Standard) and the self-directed version were no longer present, with all three versions providing comparable benefits at the three-year follow-up.

## **Reducing measurement bias**

Once equivalent groups have been established, study conditions should be identical for both groups to avoid introducing any measurement bias. In case study 6, participants in all four groups were measured with the same instruments at the same point in time immediately before and after the intervention. Participants in the three treatment groups then underwent further identical measurement at the one- and three-year follow-ups. Measurement administration was carried out by researchers who were blind to group assignment to further reduce any measurement bias.

A noteworthy feature of case study 6 was the use of coded observations of parent and child behaviours in addition to validated self-report instruments to measure pre- and post-intervention change. Although self-report measures are easy to administer, they can be viewed as having less construct validity and an increased risk of testing effect biases. For these reasons, coded observations are viewed as preferable to self-report measures, as long as the researchers are blind to group assignment and inter-rater reliability is sufficiently high. Inter-rater reliability refers to the rate of agreement between two researchers independently coding the same observations. Inter-rater reliability is understood through Cohen's kappa statistic, and convention dictates that it be 80% or higher.

The use of coded observations alongside self-report measures also permits comparisons between the findings, which additionally increases the confidence in the results if the findings are consistent.

## **Study attrition**

Participant dropout – also referred to as study attrition – is common in studies involving family-based interventions. Study attrition can substantially reduce a study's validity in three ways: (1) by reducing the sample size, (2) reducing the generalisability of the findings, and (3) increasing bias through the introduction of differential attrition.

### Loss of power

High rates of attrition resulting in a substantial reduction in the sample size runs the risk of reducing the study's ability or 'power' to observe a statistically significant effect (see Step 5 and appendix D for more information). Evaluators should therefore anticipate how much attrition is likely and then over-recruit participants to ensure that the study's power is retained.

### **Overall attrition**

High levels of *overall attrition* can also reduce the generalisability of a study's findings when those who remain in the study are significantly different to those who leave. At EIF, we recommend that if overall study attrition is greater than 10%, evaluators conduct statistical comparisons to consider whether the difference between those who dropped out and those who remained is statistically significant.

In case study 6, overall study attrition was 17%, which is not uncommon for parenting interventions. However, comparisons between those who remained and those who left revealed that those who remained were significantly *less* vulnerable. Although the authors recognised that this discrepancy had the potential to reduce the generalisability of the findings, they also noted that the families who remained were nevertheless highly vulnerable. The authors therefore concluded that the findings were still generalisable for families experiencing clinical-level difficulties with their child's behaviour.

## Differential attrition

Study attrition is also problematic when it is *differential* – meaning that the characteristics of those retained in the treatment group are different to those who remain in the control group, thereby reducing group equivalence. Differential attrition was also an issue in case study 6. Not only did participants in the intervention group leave the study at higher rates than those in the waitlist control group (for example (17% for Standard Triple P in comparison to 8% for the control group), the most vulnerable participants were also significantly more likely to leave the treatment groups than they were the waitlist control.

Methods that reduce the potential biases created by differential attrition include 'intent-totreat' designs that retain all study participants regardless of the amount of the intervention they received, as was done in case study 6. The effects of differential attrition can also be mitigated through statistical analyses that control for group differences. However, it is important to recognise that there are limits to what these statistical techniques can achieve when differential attrition is high and the groups are no longer equivalent.

## Creating and maintaining ideal study circumstances

As described at the beginning of this chapter, a primary aim of efficacy trials is to consider the intervention's maximum impact when implemented under ideal circumstances. Otherwise, it will be difficult for evaluators and others to determine if disappointing findings are a result of poor implementation or whether the intervention has, in fact, provided no effect. Methods for maintaining ideal circumstances include:

- 1. Clear eligibility criteria. Recall from previous steps that a poorly specified target population can significantly dilute an intervention's impact. Efficacy studies therefore include strict eligibility criteria and methods for retaining it. In case study 6, participants went through two rounds of screening to ensure that they met the intervention's eligibility criteria.
- 2. Taking the participants' needs into account. Participants are more likely to drop out of studies if they are difficult to participate in or require too much time. Evaluators should therefore establish clear protocols for recruiting participants and meeting their needs throughout the course of the study. These protocols should also be consistent with local ethical requirements, as discussed in Step 5.
- **3. Methods for reducing contamination:** 'Contamination' refers to the extent to which those assigned to the control group are participating in similar interventions that could reduce or 'contaminate' any potential intervention effects. In case study 6, contamination was reduced through the eligibility criteria which also required that participants were not receiving therapy or participating in other similar sorts of interventions.
- **4. Quality assurance systems**. As described in Step 4, it is important that the intervention is delivered with fidelity, meaning that it is delivered as intended and to a high standard. In case study 6, fidelity was ensured through checklists and practitioner supervision that involved the review of videotaped sessions.
- **5. Practitioner qualifications.** Interventions often assume that practitioners will be sufficiently skilled and qualified to deliver the intervention to a high standard. Efficacy studies therefore often do not begin until sufficiently trained and qualified practitioners are available to deliver the intervention.
- 6. Developer involvement. During an efficacy trial, developer involvement can substantially improve the quality of delivery and facilitate the troubleshooting of any problems. However, as we describe in Step 7, developer involvement may also artificially inflate findings and reduce the scalability of the intervention model. Developer involvement is therefore ideal during efficacy trials, in order to maximise the intervention's impact. Developer involvement should be reduced in future studies, however, once the efficacy trial is completed.

# **Statistical analyses**

It is beyond the scope of this guidance to provide advice on the various statistical analyses that can be used to test an intervention's impact during an efficacy trial. However, it is worth noting that a fundamental goal of these analyses is to determine whether the amount of change occurring between the treatment and comparison groups is statistically significant. This was the case in case study 6, which used analysis of variance as an extension of the t-test described in appendix D to determine whether the pre/post change was statistically significant within all four groups, and that these within-group changes were significantly different between each other.

It is also not uncommon for the analyses to include various statistical techniques to additionally control for any potential non-equivalence between the treatment and control groups. However, it is worth noting that the use of these statistical controls is not without controversy, and viewed by some as unnecessary if random assignment was successful. Reporting should therefore include a justification for why and how the analyses were used.

# **Understanding impact**

The ultimate goal of any efficacy trial is to understand the extent to which an intervention has made a positive impact under ideal circumstances. If the study has been conducted to a sufficiently high standard – meaning that it has adequately addressed all major threats to the study's internal validity as we have described here – then it can be *cautiously* assumed that the intervention has causally improved a child outcome of interest at one point in time, under one specific set of circumstances.

There is no question that this is a significant step in an intervention's development. However, developers and evaluators should be aware that until further rigorous testing takes place, findings from efficacy trials are not generalisable. In other words, **a positive finding observed in an efficacy study is no guarantee that the finding will be repeated again.** In addition, evaluators and developers should be circumspect when interpreting the overall meaning of the positive findings, taking into account the consistency of the findings, the size of their impacts and the extent to which these impacts are meaningful from a public health perspective and children's development.

## Consistency

The EIF level 3 criteria require that interventions must have observed at least one statistically significant improvement in one child outcome reflecting one or more of the seven outcome categories described at the beginning of this guidance. However, it should be remembered that statistically significant differences can occur at random (see appendix D), so consistency across multiple, similar outcomes is preferable over a single positive effect. The Society for Prevention Research recommends that efficacy trials are more meaningful when there is consistency in findings.

In case study 6, all three versions of Triple P observed statistically significant short-term improvements in parents' reports of children's behaviour immediately after programme completion. In addition, the two most intensive versions of the programme (Enhanced and Standard Triple P) observed actual improvements in children's behaviour through the coded observations. This consistency thus increased the evaluators' confidence in the short-term positive effects for the two most intensive versions of the programme.

While consistency in findings is important, we must caution that the need for consistency not be used as an excuse to 'fish' for positive findings through the measurement of multiple outcomes with the hope that at least one will be positive. Such practices ultimately reduce the confidence in a study's findings and are therefore often viewed with suspicion, *unless* the statistical analyses also correct for the number of tests. Trial protocols submitted prior to the study help to ensure that only the intervention's primary intended outcomes are sufficiently tested. Guidance on test correction and research protocols can be found in the further resources section at the end of this chapter.

## Size of impact

As explained in Step 5, an intervention's impact is frequently judged as small, medium or large. While a large impact is typically regarded as better, small impacts should not be viewed negatively if they are statistically significant and the intervention is not expensive.

The example described in case study 6 was designed to consider impact in three separate ways:

 The impact of three Triple P programmes against no treatment at all was considered through separate comparisons between each intervention model and a waitlist control. The comparisons confirmed that each version of the Triple P programme was significantly more effective than no treatment.

- The relative *superiority* of the three Triple P programmes was considered through comparisons between the three models, observing that the most intensive versions of the programmes (Enhanced and Standard) provided the strongest short-term impact through improving a greater number of parent and child outcomes.
- Longitudinal comparisons between the three Triple P programmes were used to consider whether the interventions' impacts were sustainable over three years. The loss of the waitlist control group meant that the study could no longer compare the impacts of the three Triple P interventions with no treatment. However, comparisons between the three treatments were nevertheless possible, revealing that while the pre/post change within the three groups remained statistically significant and positive, the between-group change was no longer significant.

The study design in case study 6 was therefore useful for the developers to consider the relative value of three versions of the programme. They observed that Enhanced Triple P provided little value over the standard version, while the standard version provided significant value over the self-directed version in the short term. Surprisingly, all three programmes appeared to provide similar longer-term value at the three-year follow-up. However, these findings should be interpreted with caution, as it was not possible to compare the three intervention's efficacy to no treatment at all.

## Value of outcome

When interpreting positive findings, it should always be kept in mind that not all outcomes provide equal value to children's development. Findings involving observed changes in children's behaviour therefore have more value than changes in children's self-reported attitudes. Researchers and evaluators should be careful not to overinterpret the value of various outcomes, especially when they are fairly distant from the primary outcome of interest or their public health relevance is relatively small. For example, an increase in children's awareness of the dangers of drug use is not the same as actual reductions in drug use.

## No effect

Occasionally, a rigorously conducted efficacy trial will observe that no effect can be causally associated with the intervention model. Such findings are inevitably disappointing, although they should not be interpreted to mean that the intervention will *never* work. Rather, they should be viewed as meaning that further work to the intervention model is required. As described in Step 5, common reasons for disappointing findings include the specificity of the target population and the intensity of the intervention's dosage. Measurement and recruitment issues can also contribute to disappointing results, although these issues should be avoided if sufficient work has taken place in Steps 4 and 5.

In very rare instances, a rigorous study may observe a causal relationship between the intervention and a negative outcome. These cases should be interpreted to reflect serious issues with the intervention model, thus highlighting the need for substantial revisions and further testing. Observing a negative effect in a rigorous study can often be avoided if sufficient pilot testing during Step 5 has taken place.

# Step 6 checklist

The checklist here reflects the minimum threshold requirements for an EIF level 3 rating. We strongly recommend, however, that intervention evaluators and developers consult other guidelines, such as CONSORT and GRADE, for conducting and reporting efficacy trials (see further resources below).

- There is good evidence that the intervention is ready for an efficacy trial, from previous evidence gathered through feasibility studies and pre/post pilots.
- The efficacy trial has made use of suitable methods, such as random assignment or rigorously conducted statistical matching, to create equivalent treatment and comparison groups.
- The study has clear eligibility requirements and methods for recruiting and retaining participants.
- The study has been approved by all relevant ethics committees.
- The characteristics of the recruited sample are consistent with the eligibility requirements and the intervention's primary target population.
- Measurement is valid and appropriate for the outcome under consideration and consistently administered to all treatment and comparison conditions at the same point in time.
- All serious threats to internal validity are reported and mitigated.
- Rates of attrition are clearly reported throughout the length of the study.
- In the event that overall attrition is higher than 10%, comparisons are made to verify the extent to which those leaving the intervention differ from those who remain.
- Differential attrition should be kept at a minimum. Most journals recommend that it not exceed 5%, although in some instances up to 10% may be acceptable if it occurs completely at random. The characteristics between the intervention and the control group should be reported if differential attrition exceeds 5%.
- Statistical analyses should be appropriate for the nature of the data and the design of the study. Statistical analyses should include corrections for the number of tests when appropriate.
- Analyses should make use of an 'intent-to-treat' design that retains participants in the study regardless of their participation in the programme. Methods for dealing with missing data should be robust and appropriate.
- Positive findings should be interpreted with caution, taking into account the consistency of the findings, the size of impact and the nature of the impact. The findings should also be interpreted in light of limitations of the study's design and execution, including the potential impact of overall and differential attrition.

### **Further resources**

#### General

CONSORT guidelines: http://www.consort-statement.org/

The GRADE working group: http://www.gradeworkinggroup.org/

- EIF evidence criteria: https://guidebook.eif.org.uk/eif-evidence-standards
- Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15*(5), 451–474.
- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, 16(7), 893–926.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3), 276–282.
- Martin, J., McBride, T., Brims, L., Doubell, L. Pote, I. and Clarke, A. (2018) *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Early Intervention Foundation. Available: https://www.eif.org.uk/ resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them
- Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology*, 5(1), e45.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

#### **Differential attrition**

Avellar, S. A., & Silman, T. (2014). What Isn't There Matters: Attrition and Randomized Controlled Trials. Mathematica Policy Research.

#### **Triple P evaluation**

- Bor, W., Sanders, M. R., & Markie-Dadds, C. (2002). The effects of the Triple P-positive Parenting Programme with co-occurring disruptive behaviour and attentional/hyperactive difficulties. *Journal of Abnormal Child Psychology*, 30, 571–587.
- Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The Triple P-Positive Parenting Program: A comparison of enhanced, standard, and self-directed behavioural family intervention for parents of children with early inset conduct problems. *Journal of Consulting and Clinical Psychology*, 68 (4), 624–640.
- Sanders, M. R., Bor, W., & Morawska, A. (2007). Maintenance of treatment gains: A comparison of enhanced, standard, and self-directed Triple P-Positive Parenting Program. *Journal of Abnormal Child Psychology*, 35(6), 983–998.

### Measurement bias

Hoskin, R. (2012) The dangers of self-report. http://www.sciencebrainwaves.com/the-dangers-of-self-report/

Kaminska, O., & Foulsham, T. (2013). Understanding sources of social desirability bias in different modes: Evidence from eye-tracking. ISER Working Paper Series, no. 2013-04.

### Reporting standards

de Boer, M. R., Waterlander, W. E., Kuijper, L. D., Steenhuis, I. H., & Twisk, J. W. (2015). Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1), 4.

# **Step 7 Test for effectiveness**



## An effectiveness study is a rigorous evaluation designed to determine if the positive child outcomes observed in the efficacy study can be replicated in real-world circumstances.

From the perspective of EIF, it is also useful if an effectiveness study (or previous efficacy study) can consider whether the intervention can be confidently associated with child benefits that are sustainable for a year or longer.

In Step 7, we describe:

- · how effectiveness studies can be conducted in real-world circumstances
- · methods for measuring change for a year or longer
- how effectiveness studies can be used to understand for whom and under what circumstances the intervention has its greatest impact
- how to interpret disappointing findings observed in effectiveness studies.

## **Efficacy versus effectiveness**

According to the Cambridge dictionary, the meaning of the words efficacy and effectiveness are virtually identical. Efficacy is described as the ability, especially of a medicine or method, of achieving something to produce the intended result, whereas it refers to effectiveness as the ability to be successful and produce the intended results.

However, when it comes to evaluation work, the terms efficacy and effectiveness refer to rigorous studies that perform two separate functions:

• Efficacy studies aim to determine the impact of an intervention under ideal circumstances. The primary aims of an efficacy study are to therefore 1) consider whether there is a causal association between the intervention model and a positive child outcome and 2) understand the magnitude of the intervention's impact under ideal circumstances.

• Effectiveness studies consider whether the impact will remain when the intervention is implemented in real-world settings. As a result, effectiveness trials can only occur after an efficacy trial has taken place. When possible, effectiveness trials should also consider for whom and under what circumstances the intervention is most effective.

The primary aim of Step 7 therefore is to conduct a second rigorous comparison group study meeting all of the requirements of Step 6, but in a natural setting with a population whose characteristics are more heterogeneous than those recruited in the first efficacy trial. From this perspective, effectiveness trials can be considered as tests of an intervention's *external validity* (as described in Step 5), whereas efficacy trials are essentially tests of an intervention's *internal validity*.

From the perspective of the EIF evidence standards, it is also useful for effectiveness trials to rigorously consider whether any observed positive child impacts can be sustained over time – especially if this has not already been considered in the first efficacy trial. This means that the study will have gained follow-up information over a year or longer on all child outcomes of interest in *both* the treatment and the control groups. Such information is particularly useful for determining whether any positive, short-term impacts remain or fade-out. Rigorously conducted follow-up studies can also verify whether the intervention is associated with any positive 'sleeper effects' involving outcomes originally observed to be significant, or outcomes not previously measured.

In this chapter, we describe what can be achieved through a well-designed effectiveness trial through an example involving the Incredible Years (IY) Preschool programme. We also discuss how a lack of positive findings from an effectiveness trial might be interpreted in light of the positive results achieved in previous effectiveness and efficacy studies.

## Case study 7: A long-term effectiveness trial of Incredible Years Preschool

### The intervention

Incredible Years Preschool Basic Programme is for parents with concerns about the behaviour of a child between the ages of three and six. Parents attend 18 to 20 weekly group sessions where they learn strategies for interacting positively with their child and discouraging negative and aggressive child behaviour. Each session is facilitated by two group facilitators with a master's qualification in a helping profession (psychology, teaching, etc.) who lead parents through group discussions that make use of video vignettes, problem-solving exercises and structured practice activities.

The Incredible Years (IY) intervention has repeatedly demonstrated short-term improvements in children's behaviour in multiple rigorously conducted RCTs taking place in the United States and other countries. This case study considers the 10-year outcomes of two effectiveness studies carried out with vulnerable children living in the UK.

### Study 1: Short-term outcomes

This study involved a previous and shorter (12-week) version of the Incredible Years Basic intervention for children between the ages of 4 and 7, delivered through the NHS Child and Adolescent Mental Health Services in south London and West Sussex. Participants were 141 low-income preschool children identified through the services as having clinical-level behavioural problems.

Children's antisocial behaviour was assessed with the parent account of child symptoms interview, as well as validated parent report measures of child behavioural problems. Parents and children also participated in an 18-minute structured play task that was videotaped and coded by trained researchers blind to treatment assignment.

Thirty-one families dropped out of the trial (17 treatment, 14 control), representing 22% overall attrition, but minimal differential attrition. Comparisons between those who remained in the trial and those who left revealed no significant differences with respect to family demographics or the outcomes of interest. Significant and often substantial short-term improvements involving large effect sizes were observed with respect to all parent and child outcomes in the treatment groups in comparison to the control for the 120 families who remained in the study.

### Study 1: 10-year follow-up

Ten years later, 94 of the original participants (recruited from the original treatment samples and subsequent waitlist group) and 26 families receiving usual care were successfully recruited to a follow-up study, representing 78% of the original sample. No significant differences were observed between those who were and were not successfully recruited. There were also no statistical differences between those remaining in the treatment and control groups.

All families participating in the 10-year follow-up completed validated, self-report measures of parent and child behaviours and participated in semi-structured psychiatric interviews. Children's IQ and reading were also assessed, and teachers were asked to complete the same standardised scales as the parents. The quality of the parent–child relationship was additionally assessed through a five-minute audio sample of a parent–adolescent discussions, as well as a 20-minute videotape of parent–adolescent interactions. Both observations were coded by trained researchers who were blind to the treatment condition.

The 10-year follow-up observed that IY children were significantly less likely to exhibit aggressive behaviours and more likely to have a higher IQ and read at higher levels in comparison to the control group children. The IY parents also demonstrated higher levels of warmth and supervision, although direct observation of parenting behaviours showed no differences between the intervention and control groups.

## Study 2: Six-year follow-up

The 10-year findings from the first study were compared to six-year findings observed in children whose parents participated in a second IY effectiveness trial which overlapped with the previous study. However, the aim of this second study was to consider the effects of IY when offered at the targeted selected level to high-risk families who *were not* experiencing clinical-level difficulties with their child's behaviour. In other words, this second trial considered the effectiveness of IY when offered as a form of prevention to vulnerable families, rather than as a form of treatment. Otherwise, all aspects of the two trials were similar, with parents and children participating in videotaped observations before and after programme completion.

While this study also observed significant short-term improvements in IY children's behaviour, these benefits were notably smaller in comparison to those in the trial testing IY with a targeted indicated sample. The short-term improvements were also not sustained at the six-year follow-up.

The authors speculated that the lack of a sustained impact for the second group could not readily be explained by child or family characteristics. However, they also noted that the parents in the targeted selected group attended fewer sessions, suggesting that perhaps they were less motivated to learn the IY content in comparison to those in the targeted indicated group.

# Three critical questions answered by effectiveness trials

Effectiveness trials are often conducted to determine whether an intervention should be made available at scale. This means that evaluators should have sufficient confidence in the intervention's impacts to assume that they can be achieved across a diverse range of circumstances in the absence of rigorous testing. Effectiveness trials are therefore most useful if they can determine (1) whether the finding is replicable under real-world circumstances, (2) whether the finding is sustainable, and (3) for whom and under what circumstances the intervention has its greatest impact.

## Is the finding replicable under real-world circumstances?

As mentioned previously, the primary purpose of effectiveness trials is to determine whether positive results are replicable under real-world circumstances. In case study 7, the Incredible Years Preschool programme underwent two rigorous effectiveness studies meeting the requirements described in Step 6 to determine whether the intervention could replicate its US benefits in the UK.

Both studies confirmed that the intervention was feasible (retention rates were reasonable) and could provide benefits that were similar in magnitude to those observed in previous US studies. In fact, the first effectiveness study resulted in impacts that were the same as those observed in the intervention's first efficacy trial.

## Is the finding sustainable?

Findings from the first effectiveness study described in case study 7 also confirmed that the benefits associated with the IY programme were sustainable. The study achieved this by retaining the details of the families who participated in the original study and then successfully recruited them to the second follow-up study. This study observed that antisocial children whose parents attended the Incredible Years programme 10 years previously were not only less aggressive but were also reading at higher levels than those whose parents did not attend the intervention.

These findings are noteworthy for several reasons.

- First, they were observed in an effectiveness trial, rather than an efficacy trial. More typically, positive long-term outcomes are confirmed in follow-up studies of efficacy trials conducted under ideal circumstances.
- Second, these findings not only confirmed that IY's primary intended outcomes were sustainable, but additional benefits associated with children's reading were also possible.
- Third, positive benefits consistent with IY's ToC were observed 10 years after intervention completion. Although many child and family interventions intend to provide long-term benefits for children's development, relatively few have had the opportunity to test and confirm them.

# For whom and under what circumstances does the intervention have its greatest impact?

Although it is not a requirement for effectiveness studies to test for differential impacts, it is highly preferable if they do so. The Incredible Years intervention accomplished this by comparing two effectiveness studies: one that examined the effectiveness of the programme with vulnerable children identified as having behavioural problems, and a second conducted with equally vulnerable children who were not identified as having behavioural problems.

By comparing the two studies' findings, the evaluators observed that the impacts of the intervention were greatest and more sustainable when the intervention was provided at the targeted indicated level. This finding is highly useful for understanding how the intervention should be targeted when offered at scale, indicating that the intervention is most effective when offered to families experiencing specific difficulties with their children's behaviour.

Comparing the intervention's outcome when offered to two different samples, experiencing differing levels of needs is an ideal way of understanding for whom the intervention has its greatest impact. However, it would also have been possible for the authors to consider this by conducting *subgroup analysis* within one of the samples. For example, it may have been possible for the authors to subdivide the second sample into groups of more and less vulnerable families on the basis of their demographic or baseline characteristics and then compare their outcomes. If the outcomes of one of the groups were significantly greater, the authors would have had a deeper understanding of for whom the intervention worked best.

While some subgroup analyses may be specified in a study protocol, it is important to emphasise that subgroup analyses are fundamentally post hoc – meaning that they are conducted *after* viewing the data, and so are not true tests of the evaluator's original hypothesis. Nevertheless, findings from such studies can usefully inform the intervention's eligibility criteria in subsequent studies – thereby increasing the intervention's impact over time.

# Effectiveness trials observing no effect

It is becoming increasingly common for effectiveness studies to observe no effect, so interpreting the findings from these studies can be challenging. This is because it can be difficult to determine whether positive findings can *never* be replicated in real-world circumstances, or whether there were circumstances unique to the effectiveness trial which diminished the positive impacts initially observed in the efficacy trial.

It is useful to consider the following issues when effectiveness trials fail to replicate the positive results observed in efficacy studies.

- First, were their particular problems with the implementation of the intervention, resulting in it not being delivered to a high standard or with fidelity? Or, did those delivering the intervention significantly alter the model so that content was missing or substantially reduced? In Step 8, we describe in greater detail the measures that should be taken to ensure that an intervention's fidelity is retained every time it is delivered.
- Second, did the effectiveness trial consider the same outcomes as the efficacy trial? While it is highly useful for effectiveness trials to consider outcomes in addition to those first tested in efficacy trials (as was the case in case study 7), these new outcomes should not be considered at the expense of the intervention's original primary child outcomes.
- Third, was the sample significantly less disadvantaged than in the efficacy study? Typically, an intervention's impact will be greater for more vulnerable populations, particularly when the intervention is specific to their needs. It is therefore possible that the increased diversity of the effectiveness study's sample substantially reduced the intervention's impact. Evaluators can check for whether this occurred by comparing the characteristics of those participating in the effectiveness study to those who originally participated in the effectiveness trial.
- Fourth, did the intervention represent value over what was currently available in the new setting? This question is particularly important when transporting interventions into new countries, where business as usual is already at a high standard. In this respect, the answer to this last question should be viewed as a primary reason for conducting an effectiveness trial in the first place. This means a result observing no effect is

nevertheless valuable. However, it should be kept in mind that the intervention still has the potential to provide value in other circumstances where business as usual may not be as effective. This issue is explored in greater detail in Step 9.

## Step 7 checklist

- Has a rigorous study meeting all of the Step 6 requirements been conducted in realworld circumstances?
- Has the study confirmed significant positive child outcomes?
- Has the evaluation considered for whom the impacts are strongest and under what circumstances?
- Have longer-term impacts been verified through a 12-month or longer follow-up involving the original efficacy study or subsequent effectiveness trials?

### **Further resources**

#### General

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, 6(3), 151–175.

- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926.
- Martin, J., McBride, T., Brims, L., Doubell, L. Pote, I. and Clarke, A. (2018) *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Early Intervention Foundation. Available: https://www.eif.org.uk/ resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

#### **Case study 7 references**

- Scott, S., Spender, Q., Doolan, M., Jacobs, B., & Aspland, H. (2001). Multicentre controlled trial of parenting groups for childhood antisocial behaviour in clinical practice. *British Medical Journal*, 323, 1–7.
- Scott, S., Briskman, J., & O'Connor, T. (2014). Early prevention of antisocial personality: Long-term follow-up of two Randomized Controlled Trials Comparing Indicated and Selective Approaches. *American Journal of Psychiatry*, 171, 649–657.

# Step 8 Refine and monitor



Once an intervention has confirmed that it can provide sustainable benefits for children that are meaningful from a public health perspective, further testing is necessary to develop quality assurance systems to ensure that these benefits remain replicable.

In Step 8, you will learn:

- how evaluation methods can be incorporated into the running of an intervention to improve its quality on an ongoing basis
- · how to monitor child outcomes on an ongoing basis
- how monitoring systems can be used to determine when an intervention is appropriate for an individual child's needs or when referral to other services may be necessary
- how rapid cycle evaluations and microtrials can be used to test and refine an intervention's active ingredients
- evaluation methodologies for testing an intervention's workforce requirements.

# A new set of steps

Once you arrive at Step 8, you should have a fairly good understanding of what your intervention can do. This includes knowledge that your intervention is feasible and that children and families will come to it and stay. This also includes confidence – gained through the robust evaluation evidence obtained in Steps 6 and 7 – that the intervention can deliver meaningful child outcomes and these outcomes can be achieved in real-world settings.

For some, this knowledge might mean that the intervention's evaluation journey has come to an end. However, if the aim is to make the intervention more widely available by taking it to scale, this conclusion could not be farther from the truth. If this is the case, Step 8 is only the first in a series of new and more iterative evaluation steps necessary to ensure that the intervention will remain effective when offered to new and diverse circumstances at the same time.

These new steps include the development of quality assurance and improvement systems to ensure that the outcomes observed in Steps 6 and 7 can be replicated *when delivered by someone other than the intervention's developer*. Examples of quality assurance systems include standardised installation and implementation procedures (including training, supervision and fidelity monitoring) to help organisations, independent from the intervention's developer, deliver the intervention to a high standard. Good systems also include guidance on how the intervention can be adapted so that it remains effective within a variety of different contexts.

Steps 9 and 10 cover methods for adapting and developing interventions to ensure that they remain effective when offered at scale. Here, in Step 8, we describe how quality improvement cycles can be used to test and refine various aspects of the intervention model so they can be standardised. These processes are informed by the 'plan-do-check-act' (PDCA) cycles commonly used by many organisations (figure 8.1). We describe here how the short-term outcomes confirmed in Steps 6 and 7 can be incorporated into these cycles to preserve an intervention's quality when it is delivered.

### FIGURE 8.1



Principles of the plan-do-check-act cycles that inform quality improvement

Source: Based on the Deming cycle; see Moen, R. (2009). 'Foundation and History of the PDSA Cycle'. Associates in Process Improvement. USA.

We also consider how an intervention's outcomes can be used to systematically manipulate and test the intervention's key components (also referred to as the 'active ingredients') to differentiate those which are essential from those which are not. Knowledge of these ingredients can then be used to standardise intervention procedures so that all active ingredients are maintained as the intervention is implemented (figure 8.2).

### FIGURE 8.2

Using quality improvement cycles to standardise practices and further test impact



Source: Based on the Deming cycle; see Moen, R. (2009). 'Foundation and History of the PDSA Cycle'. Associates in Process Improvement. USA.

The active ingredients of child and family interventions typically fall into one of three categories:

- · the specific curriculum or content
- · the eligibility criteria and systems for maintaining it
- the workforce requirements and processes for training, certifying and supervising practitioners.

In this chapter, we provide examples of how an intervention's active ingredients can be evaluated within each of these three categories. This includes advice on the importance of robust monitoring systems for ensuring that child outcomes are achieved as the intervention is being implemented. We also consider strategies for systematically testing and refining key aspects of the intervention's content and eligibility criteria. We then conclude with advice on evaluation that can be used to inform decisions about the intervention's workforce requirements.

## **Outcome-based monitoring systems**

In Step 4, we discussed the importance of monitoring an intervention's progress towards its intended outputs. These ouputs included meeting pre-identified delivery and fidelity targets, rates of recruitment and retention and participant satisfaction. By the time an intervention reaches Step 8, monitoring systems should be in place to routinely collect information on all of these outputs as the intervention is implemented.

Now that child outcome information has been confirmed in Steps 5, 6 and 7, integrating information about child outcomes into previously established output monitoring systems should be a relatively straightforward process. However, collecting child outcome data often represents a cumbersome extra step to the delivery process. As described in Step 5,

outcome data should ideally be gathered twice – once at the start of the intervention and then again after it is completed. If measures are lengthy, this may add considerable extra time. For example, it is not uncommon for child and family interventions to require at least one additional session at the beginning of the programme to collect all necessary child outcome information.

It is also worth noting that some data collection processes pose ethical challenges because they can lead to child protection disclosures. Care therefore must be taken to ensure that data collection methods are properly embedded within local safeguarding procedures and comply with data protection standards. In addition, procedures should be established to ensure that data collection processes are carried out reliably, otherwise the validity of the data will be compromised and collecting it will have been a waste of time.

In many instances, the outcome measures used as part of the intervention's previous evaluations may not be practical in real-world settings. This may be because they require videotaping and expensive, coded observations by independent researchers. Some measures may also not be sensitive enough to measure the pre/post changes occuring in the small samples receiving the intervention. It is therefore recommended that over time, developers create their own companion outcome metrics that are bespoke to their intervention model.

As described in Step 5, creating measures that are valid, reliable and sensitive to change is a time-consuming process. Testing and validating new outcome measures is feasible, however, if data is collected on a continuous basis as the intervention is implemented. While it is not essential that bespoke measures undergo the same level of validation as the measures used in Steps 5, 6 and 7, work is still necessary to ensure that their findings are consistent with more robust and established measures.

Bespoke monitoring measures offer a number of advantages that can offset the extra time and expense required to develop them. These advantages include the ability to monitor participants' progress as the intervention is being implemented. This information can then be used to determine whether participants are benefiting from the model so that appropriate, mid-course corrections can be made. Examples of common mid-course corrections include referring participants on to more intensive services if it is clear that they are not benefiting from the intervention.

Bespoke monitoring tools are also useful for supporting communication between participants and practitioners to inform decisions about the intervention's advice and intensity. For example, the Parent Daily Report (PDR) described in case study 8 is a valdiated tool for assessing pre- and post-intervention changes in problematic child behaviour. However, it is also highly useful for tailoring parenting advice to families' specific needs. When used in this fashion, the measure in fact becomes part of the intervention, meaning that it is no longer sufficiently independent to be used as an assessment tool during more rigorous impact evaluations. Evaluating child outcomes is nevertheless possible if other similar, validated instruments are then used as part of the study.

Outcome monitoring systems are significantly enhanced by electronic software that facilitates the collection, storage and analysis of data. Once these systems are set up, comparisons can then be made between the outcomes achieved during normal delivery and those observed in efficacy and effectiveness trials. As more outcome data is collected, findings can be merged and norm-referenced to benchmark the performance of individual settings, practitioners and participants.

Outcome-based monitoring is highly useful for a variety of quality improvement purposes. However, it is worth remembering that monitoring data is seldom as robust as the data collected during efficacy and effectiveness trials. This is because data is often systematically missing or incomplete, and procedures for collecting it are typically less standardised. Monitoring data is also frequently collected in the absence of a comparison group, meaning it is impossible to understand the extent to which findings are related to participant biases. Therefore, monitoring data should never be considered a replacement for the findings observed in robust efficacy and effectiveness studies, which should continue as developers prepare their intervention for going to scale.

## **Case study 8: The Parent Daily Report**

The Parent Daily Report (PDR) was developed by Gerald Patterson and colleagues in 1969 to measure the frequency of problematic child behaviours such as hitting, fighting, teasing, temper tantrums and talking back. These behaviours were first identified through classroom observations of preschool children interacting (described in Step 1), which were then consolidated into a checklist of 34 difficult child behaviours that parents could rate on a daily basis.

The PDR has since been validated as a reliable tool for testing changes in problematic behaviours of preschool children. However, the checklist nature of the tool also helps parents identify specific behavioural problems that can be discussed with practitioners and inform tailored parenting advice. Individual progress can then be monitored on a regular basis to determine the success of various parenting strategies and make mid-course corrections if certain strategies are not working. For example, the GenerationPMTO programme (case study 2) and Incredible Years Preschool (case study 7) both use of the PDR as an intervention tool to tailor advice to the needs of individual families.

# Rapid cycle evaluations for testing an intervention's active ingredients

Rapid cycle evaluations have recently been proposed as a 'fast' and theoretically less expensive way of testing an intervention's active ingredients. Their premise is similar to that of a superiority trial, as the aim is to compare different forms of the same treatment. The Triple P efficacy trial described in case study 6 served this purpose, as the relative efficacy of three forms of the intervention were compared to each other, as well as business as usual.

Rapid cycle evaluations differ from superiority trials, however, as they make use of data that is already collected on an ongoing basis to compare the value of intervention enhancements in a relatively fast manner. In this respect, the aim of rapid cycle studies is not to prove that something works, but to understand how to make it work better.

Let's say, for example, that one wants to test the relative value of five different parenting strategies taught during a parenting intervention. A rapid cycle evaluation design could be used to randomly assign parents already recruited to the intervention to one of two versions: one that teaches parents all five strategies and another that teaches only four. The outcomes at the end of the interventions could then be compared and further, similar testing could be repeated with the remaining strategies. Once all of the cycles were completed, developers would have a fairly good idea about which strategies were the most effective, as well as the participants who most benefited from them.

The interpretation of the findings from rapid cycle evaluations are not without limitations, particularly if they are conducted with small samples or with measures that have not been fully validated. Nevertheless, they are a useful form of in-house piloting to compare the relative effectiveness of minor changes in intervention content. More rigorous rapid cycle evaluation are also possible in situations where evaluators have access to large samples and outcome datasets, as in schools or community services. In these cases, it is imperative that

data collection methods are sufficiently robust so that biases are not introduced through high levels of missing data.

## **SMART evaluation**

A more advanced form of rapid cycle evaluation, referred to as SMART (Sequential Multiple Assignment Randomised Trial), makes use of randomisation sequences to inform processes which adapt interventions to participants' individual needs. As illustrated in figure 8.3 (below), participant treatment response is monitored on an ongoing basis and then used to inform decisions about further randomisation into subsequent treatments. Outcome data is then collected at the end of the trial and is used to further refine the treatment and its referral protocols.

It is fair to say that the sophisticated nature of SMART designs means that they are often expensive, and not particularly fast. However, they are well suited to studies of various technologies that include a personalisation component, including computer software and mobilephone apps. Less rigorous hybrids of SMART designs are also useful for comparing the outcomes of personalised referral and treatment decisions. For example, the Family Nurse Partnership (FNP) programmes is currently making use of rapid cycle techniques to adapt and personalise the FNP model to the needs of different participant groups.

## FIGURE 8.3

Flow diagram of a Sequential Multiple Assignment Randomised Trial (SMART)



STEP

# **Determining workforce needs**

As interventions are taken to scale, they will inevitably be delivered by practitioners with differing skills and qualifications. Studies show that practitioner skills can strongly impact intervention outcomes and can even result in negative effects when interventions are delivered by practitioners lacking sufficient skills. Given that practitioner qualifications also impact the cost of the intervention, developers should have a good understanding of the qualifications and skills required to successfully implement their intervention before it is delivered at scale. This information is also useful for developing and testing practitioner training manuals and certification procedures.

There are multiple ways to assess practitioner skills, and it is beyond the scope of this guide to desribe all of them. Instead, we provide two examples that expand upon the outcomebased methods introduced previously. However, we encourage developers to seek further advice on how to evaluate practitioner skill within their discipline to ensure that all of their training and supervision processes are effective and appropriate.

The first example (case study 9) involves the Family Nurse Partnership (FNP) programme's second effectiveness trial, which compared the impact of the intervention when delivered by trained and qualified family nurses, paraprofessionals and treatment as usual. The study observed that although the mothers visited by the paraprofessionals valued the support and advice they received, mothers who were visited by family nurses were more likely to reduce their smoking, delay their next pregnancy and seek employment. On the basis of these findings, the authors concluded that the family nurses' professional training enabled them to better learn the intervention content and thereby communicate it in a way that FNP participants better learned it.

## Case study 9: Comparing the effectiveness of the Family Nurse Partnership programme when delivered by nurses and paraprofessionals

The Family Nurse Partnership (FNP) programme was originally developed as a 2.5 year home visiting intervention for single, first-time teenage mothers. As first described in case study 4, trained and qualified family nurses visit mothers on a weekly basis from their first pregnancy booking onwards and then for six weeks after the child's birth. Visits then continue on a fortnightly basis until the child's second birthday, resulting in a total of 64 visits. During these visits, mothers are supported by FNP practitioners to take care of the physical and mental needs and support their baby's development.

At the time of this study, FNP had already undergone two rigorously conducted RCTs, both observing reductions in mothers' health risk behaviours and improved child outcomes. However, weekly and fortnightly visits by qualified nurses are expensive, and some studies have found that home visiting has the potential to be effective if delivered by sufficiently trained and supervised paraprofessionals. A third effectiveness trial was therefore conducted to determine the relative value of FNP in comparison to a paraprofessionally-led home visiting service.

Although mothers were less likely to drop out of the paraprofessional group (23% for paraprofessionals vs 29% for the nurse-led group), the short-term mother and child outcomes were consistently better for those visited by nurses than those visited by paraprofessionals. In particular, FNP mothers were less likely to smoke or have a repeat pregnancy and more likely to be employed in comparison to the mothers visited by the paraprofessionals or control group mothers. FNP infants were also significantly more likely to engage positively with their mother at six months and demonstrated improved language at 21 months.

The research team attributed these differences to the fact that there was higher turnover in the paraprofessional group and that nurses spent more time with the mothers on health-related issues. The authors concluded that while the training for nurses and paraprofessionals was identical, the nurses' background facilitated a more sophisticated understanding of the training content, which in turn helped them to cover important content during the visits in a way that mothers could learn it.

Our second example involves a subsample of practitioners delivering Incredible Years Preschool in the RCT described in case study 7. In this instance, a coded observational tool was developed to assess the skill of the 15 facilitators delivering the intervention to the 90 families participating in the treatment arm of the trial. The study observed that child outcomes were significantly correlated with practitioner skill, and that when skill was particularly low, child outcomes actually became worse (figure 8.4). The study additionally observed that practitioners with previous mental health training delivered the intervention with greater skill, as measured through the coded observations.

### FIGURE 8.4



Impact of therapist skill levels on the outcomes of children whose parents attended an Incredible Years programme

Source: Scott, S., Carby, A., & Rendu, A. (2008). Impact of therapists' skill on effectiveness of parenting groups for child antisocial behavior. Institute of Psychiatry, Kings College London.

In a separate study, a similar observational coding scheme was used to measure practitioner skill in an RCT comparing two separate training packages for the Incredible Years intervention: a standard, three-day practitioner training programme, or an enhanced version that was augmented by ongoing coaching and supervision as the practitioners delivered the intervention. The study observed that while intervention fidelity was similar for both groups, practitioners receiving the enhanced coaching demonstrated greater skill in providing practical support to parents and mediating the vignette-led group discussions. This information was then fed into the ongoing development of the intervention's training.

## **Step 8 checklist**

Below is a list of common quality assurance activities recommended for interventions aiming to be offered at scale.

The intervention has developed robust data collection systems for monitoring delivery targets, intervention fidelity, recruitment, retention, child outcomes and parent or practitioner outcomes when relevant.
Data collection tools and software is available for host organisations.
The intervention has well-developed eligibility criteria that has undergone some testing.
The intervention model has methods for collecting information on participant outcomes on an ongoing basis to determine whether the intervention is appropriate and participants are benefiting from it.
Quality improvement testing has taken place to identify the intervention's active ingredients and the conditions that support them.
The workforce requirements of the intervention have been specified and tested.
Methods are available for assessing practitioner skill.
Practitioner training and supervision has undergone testing and development.

### **Further resources**

### General

Catch-22. Realising ambition insight 8: Proving vs Improving: Routine outcome monitoring to support service improvement efforts. https://www.catch-22.org.uk/services/realising-ambition/publications/

### CORC: https://www.corc.uk.net/

- Haynes, L., Goldacre, B., & Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Cabinet Office.
- Leadbeater, B. J., Dishion, T., Sandler, I., Bradshaw, C. P., Dodge, K., Gottfredson, D., ... & Smith, E. P. (2018). Ethical challenges in promoting the implementation of preventive interventions: Report of the SPR task force. *Prevention Science*, *19*(7), 853–865.

Moen, R. (2009). 'Foundation and History of the PDSA Cycle'. Associates in Process Improvement. USA.

### The Family Nurse Partnership evaluation

Dartington Service Design Lab and Family Nurse Partnership National Unit (2018). Interim report of the FNP ADAPT project. https://fnp.nhs.uk/fnp-next-steps/adapt/

Olds, D. L., Robinson, J., O'Brien, R., Luckey, D. W., Pettitt, L. M., Henderson, C. R., ... & Talmi, A. (2002). Home visiting by paraprofessionals and by nurses: A randomized, controlled trial. *Pediatrics*, 110 (3): 486–496.

### The Incredible Years evaluation

Scott, S., Carby, A., & Rendu, A. (2008). Impact of therapists' skill on effectiveness of parenting groups for child antisocial behavior. Institute of Psychiatry, Kings College London.

Webster-Stratton, C. H., Reid, M. J., & Marsenich, L. (2014). Improving therapist fidelity during implementation of evidence-based practices: Incredible Years program. *Psychiatric Services*, 65(6), 789–795.

### Rapid cycle evaluation and microtrials

American Institute of Research. Rapid cycle evaluations: https://www.air.org/service/client-services-rapid-cycleevaluations

Mathematica Policy Research: https://www.mathematica-mpr.com/our-capabilities/rapid-cycle-evaluation

- Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A 'SMART' design for building individualized treatment sequences. *Annual review of clinical psychology*, *8*, 21–48.
- Leijten, P. H. O. (2014). Toward Improved Parenting Interventions for Disruptive Child Behavior: Engaging Disadvantaged Families and Searching for Effective Elements (Doctoral dissertation, Utrecht University).
- Leijten, P., Dishion, T. J., Thomaes, S., Raaijmakers, M. A., Orobio de Castro, B., & Matthys, W. (2015). Bringing parenting interventions back to the future: how randomized microtrials may benefit parenting intervention efficacy. *Clinical Psychology: Science and Practice*, 22(1), 47–57.

STEP

# Step 9 Adapt and transport



## As interventions are taken to scale, the diversity of the contexts in which they will be offered will inevitably increase. If this involves 'transporting' interventions into new cultures, substantial changes are particularly likely.

In step 9, we introduce evaluation methods that can determine:

- the extent to which intervention contents are relevant within new cultures and countries
- whether the intervention's intended child outcomes are upheld through ongoing piloting
- the extent to which interventions developed in one country are needed and will 'fit' within the context of another.

# Managed adaptation

As interventions are taken to scale, the diversity of the contexts in which they will be offered will inevitably increase. This is particularly true if interventions are offered in countries where the cultural views, language and service delivery infrastructures are entirely different from those where the intervention was first developed. In these instances, substantial adaptations will need to be made to the intervention's content to ensure that the model remains appropriate and relevant.

In Step 8, we introduced the importance of understanding the intervention's active ingredients so that quality assurance processes can be developed to preserve them. Key ingredients include features of the intervention's training activities, delivery requirements, content, materials, dosage and eligibility criteria. If any one of these elements are changed, the intervention's effectiveness could be substantially reduced. For these reasons,

intervention adaptations must be managed with care. This means that evaluation methods, such as those described in Steps 5 and 8, should be used to pilot adaptations. If adaptations significantly alter the design or delivery of the intervention, it is likely that further efficacy and effectiveness trials will be necessary to confirm the adapted intervention is at least as effective as previous versions. The box below provides a list of common adaptations that require rigorous piloting, efficacy and effectiveness testing.

# Adaptations that are likely to impact an intervention's effectiveness

- · reducing the number or length of sessions
- · lowering the level of participant engagement
- eliminating key messages or skills learned
- removing topics
- changing the theoretical approach
- using staff or volunteers who not adequately trained to deliver the programme
- using fewer staff members than recommended.

At some point, modifications to content will alter the original assumptions underpinning the intervention's original ToC. This is especially true if the intervention aims to target a new period in children's development, or a different level of family need. **Time and again**, **studies have shown that intervention models that have been proven to be effective for one stage of children's development are not necessarily effective for another.** It is therefore strongly recommended that when interventions are adapted to a new age group, another literature review is conducted to confirm that the ToC is fully relevant for the children's age and needs. Once the new ToC has been confirmed, it will inevitably have knock-on effects on the intervention's logic model, blueprint, and so forth, meaning that Steps 1 to 7 will need to be revisited to ensure that the new intervention is also effective. Indeed, the evaluations described in case studies 4 and 5 involved adaptations of interventions to service the needs of different target populations.

Here, in Step 9, we describe the methods involved in adapting interventions so that they can be transported from one culture to another. In these instances, dramatic alterations in the intervention's content and format are less likely than when interventions are altered on account of children's age, although a considerable amount is still necessary to ensure that the materials are culturally appropriate and appealing to the new target audience.

# Adaptation for transportation

The term 'transportation' is commonly used to describe the process of adapting interventions developed in one country for use in another. One might assume that this simply means translating the intervention's materials into a new language and making any visual materials more culturally relevant. However, as described in case study 5, adapting interventions for transportation also involves attention to differences in cultural norms, especially those involving children's development, child-rearing practices, socially desirable behaviours, educational needs and employment opportunities.

Transportation also involves a careful consideration of the new contexts in which the intervention will be delivered. For example, many US interventions developed for children between the ages of 11 and 14 are delivered through middle schools, which do not exist

in the UK or many other western countries. Structural differences between middle schools and other secondary school contexts include the age range of the children attending the school (US middle schools are typically for children between the ages of 11 and 14), teacher training and qualifications, teacher-to-child classroom ratios and curriculum content. All of these differences are likely to influence aspects of the intervention's key ingredients and thus need careful consideration when US middle school interventions are transported to other countries.

Other issues relevant to intervention transportation include:

- the extent to which the intervention's outcome measures have been translated and validated within the new host country
- the skills and qualification of the new country's workforce
- the facilities and technologies available to deliver the intervention
- the laws and attitudes governing responses to truancy, smacking, antisocial behaviour and drug use.

Most importantly, policymakers and other stakeholders involved in transportation decisions must carefully consider the extent to which the intervention has the potential to add value over the services that are currently available in the new country. This means taking all of the above factors into account to make careful comparisons between the effectiveness of the intervention in its original country comparison to the effectiveness of current business as usual. This work includes careful planning and testing that makes use of the consensus building and evaluation methods described in previous steps.

In case study 10, we describe the testing that took place to inform the transportation of the US-based intervention, Strengthening Families 10-14 (SF 10-14) to the UK. SF 10-14 was first identified by the UK Alcohol Education and Research Council as a result of concerns involving rising rates of teenage drug and alcohol misuse. The Council identified SF 10-14 as having potential for reducing these rates on the basis of long-term findings from a rigorously conducted US efficacy trial observing reductions in the risk of substance misuse and improvements in children's school achievement. UK providers were, however, aware that substantial differences existed between the US and the UK contexts. Careful planning was therefore necessary to ensure that the intervention's aims were relevant for UK children and its delivery mechanisms were feasible within UK service structures.

These initial studies were highly useful for understanding how to revise the SF 10–14 model to increase its relevance within the UK. These studies identified initial issues with the US materials and difficulties in recruiting families to the study. These issues have since been addressed and the intervention is now undergoing further testing in seven UK sites through the Big Lottery Realising Ambition Fund.

# Case study 10: Cultural adaptation of Strengthening Families 10–14

### The programme

Strengthening Families 10–14 (SF 10–14) is a parenting and family strengthening programme for families with children aged between 10 and 14. The programme aims to enhance family protective processes involving communication and child supervision as a way of reducing the risks associated with risky adolescent behaviour. Parents and their children (aged 10 to 14) attend seven weekly sessions which use videotapes to teach parents how to communicate effectively with their teenage child and set appropriate limits on various risky behaviours.

At simultaneous sessions, children learn strategies for resisting peer pressure, increasing prosocial behaviour and managing stress and negative emotions. Several rigorously conducted studies in the US have observed small to moderate reductions in adolescent children's drug and alcohol initiation, risky sexual behaviour and school achievement.

On the basis of these findings, the UK Alcohol Education and Research Council identified SF 10-14 as a potential intervention for reducing rising rates of drug and alcohol misuse in UK adolescents. The council therefore funded a two-phase study to explore the feasibility of the intervention for transportation to the UK.

### **Phase I evaluation**

Phase 1 made use of participatory methods and focus groups to determine the extent to which the SF 10–14 model was transportable within the UK context and extent to which intervention materials needed to be modified.

- Participatory methods were used to build consensus among practitioners and host agencies involved in youthwork on initial adaptations to the SF 10-14 materials. These methods involved gathering views from group members independently, circulating ideas to the group until all ideas were exhausted and duplicates removed, then reaching a group consensus about agreed adaptations through discussion. The group decision was subsequently checked and agreed by all group members. Revisions to the US materials were then made to create adapted UK SF 10-14 materials.
- Focus groups assessed both the original US SF 10-14 materials and the revised UK SF 10-14 materials to provide further feedback on the US materials and to comment on whether the changes made in the UK version were necessary and/or useful. The focus groups comprised parents/guardians and youth from four sites in the UK.

### **Phase I findings**

Both methods resulted in the recommendations for multiple changes in the original US materials.

- Participants felt that some of the language in the US version was old-fashioned and overly sentimental. Further, it was recommended that idiomatic American expressions be translated into UK relevant terms.
- In the videos, voiceover narration was recommended rather than on-screen narrators, who were seen as patronising.
- Participants felt that the religious connotations in all materials should be removed: for example, it was recommended that the word 'creed' (perceived as a religious word) should be changed to 'motto' (a nonreligious word).

Participants also agreed with the US materials in some key areas.

- Participants did not identify discrepancies in definitions of undesirable behaviour, for example regarding drug misuse and antisocial behaviour.
- Participants did not comment that the recruitment systems or service providers were inappropriate.

### **Phase II evaluation**

Phase II involved a pilot study that made use of quantitative and qualitative methods similar to those described in Steps 4 and 5 of this guidance.

• Quantitative methods included a pre/post investigation of 53 parents and 69 children who attended the intervention in three UK sites. Parents and children completed measures that had been validated for use with the programme once prior to attending the intervention, then again upon intervention completion and then at a three-month follow-up.

• Qualitative methods included focus groups with purposive samples of 16 parents and 14 young people attending the intervention.

### **Phase II findings**

The study revealed significant challenges in recruiting participants to the study and programme. However, once recruited, retention rates were high.

The study further observed no significant pre/post changes in parent and child behaviours. The authors attributed the lack of positive change to difficulties in recruiting a sufficient number of participants, resulting in the study being underpowered to observe moderate intervention effects.

Findings from the focus groups nevertheless confirmed that families liked the intervention and perceived improvements in family functioning that could potentially lead to reductions in a variety of adolescent risks and improved school achievement. On the basis of these findings, it was decided that SF 10–14 had promise within the UK context, although further work involving Steps 4 to 7 is necessary to confirm that the model is effective in this country.

## **Step 9 checklist**

Interventions are best suited for transportation when they meet the following criteria:

- The original model has evidence of effectiveness from at least one efficacy trial.
- Intervention 'fit' has been reviewed and agreed through consensus building techniques with key stakeholders.
- There is good reason to think the intervention will add value over what is currently available in the new country.
  - The impact of various adaptations on the intervention's key ingredients have been carefully considered.
- Feasibility testing to consider the acceptability of the intervention in new contexts (including recruitment and retention rates) is planned.
- Pre/post pilot testing involving the intervention's previously confirmed outcomes and meeting the requirements of Step 5 is planned or under way.

### **Further resources**

### General

Asmussen, K. (2011). The evidence-based parenting practitioner's handbook. Routledge.

- Barrera Jr, M., & Castro, F. G. (2006). A heuristic framework for the cultural adaptation of interventions. *Clinical Psychology: Science and Practice*, *13*(4), 311–316.
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327.

- Fraser, M., Richman, J. M., Galinsky, M. J., & Day, S. H. (2009). *Intervention Research: Developing Social Programs*, Oxford University Press.
- Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, 6(3), 151–175.
- Gardner, F., Montgomery, P., & Knerr, W. (2016). Transporting evidence-based parenting programs for child problem behavior (age 3–10) between countries: Systematic review and meta-analysis. *Journal of Clinical Child & Adolescent Psychology*, 45(6), 749–762.
- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926.
- O'Connor, C., Small, S. A., & Cooney, S. M. (2007). Program fidelity and adaptation: Meeting local needs without compromising program effectiveness. *What works, Wisconsin research to practice series*, *4*, 1–6.
- Resnicow, K., Baranowski T., Ahluwalia J. S., & Braithwaite R. L.(1999). Cultural sensitivity in public health: Defined and demystified. *Ethnicity and Disease*, 9, 10–21.
- Guide for adapting Evidence-Based Programmes: https://txicfw.socialwork.utexas.edu/wp-content/ uploads/2016/09/Guide-to-Adapting-an-Evidence-Based-Intervention.pdf

#### Strengthening Families 10–14

- Allen, D., Coombes, L., & Foxcroft, D. R. (2006). Cultural accommodation of the Strengthening Families Programme 10–14: UK Phase I study. *Health Education Research*, 22(4), 547–560.
- Barrera, M., Castro, F. G., & Steiker, L. K. H. (2011). A critical analysis of approaches to the development of preventive interventions for subcultural groups. *American Journal of Community Psychology*, 48(3–4), pp.439–454.
- Coombes, L., Allen, D., & Foxcroft, D. (2012). An exploratory pilot study of the Strengthening Families Programme 10–14 (UK). *Drugs: Education, Prevention and Policy, 19*(5), 387–396.
- Gorman, D. M. (2017). The decline effect in evaluations of the impact of the Strengthening Families Program for Youth 10–14 (SFP 10–14) on adolescent substance use. *Children and Youth Services Review*, *81*, 29–39.



## Going to scale is the last step in our 10 Steps framework. Reaching this step does not mean, however, that the intervention's evaluation journey is over. Instead, it means that evaluation cycles are embedded within the intervention's delivery systems to verify that it remains effective when offered at scale.

In Step 10, we describe all of the quality assurance processes necessary for offering interventions at scale, including those which help local systems determine if they are ready to offer an intervention in a way that will ensure that it remains effective.

In Step 10, you will learn:

- · methods for assessing local system readiness
- · the role of the intervention provider for informing system readiness
- the ways in which technical support can be used to inform system readiness and install interventions within local systems
- methods for offering and using technical support, including licensing, purveyors and independent intermediaries.

# Systems for scale-up

Step 10 represents the pinnacle of the intervention's development. By this point, the intervention will not only have robust evaluation evidence confirming that it provides meaningful benefits to children and families, it will also have developed systems to ensure that these benefits can be achieved on a regular basis in diverse circumstances. This means that the intervention is ready to be offered at scale. It does not, however, mean that its evaluation journey has come to an end.

Table 10.1 provides a list of the Society for Prevention Research's standards for intervention scale-up. In previous steps, we briefly introduced some of the activities that make up these standards, including methods for monitoring recruitment and reach, delivering the intervention with fidelity, monitoring effectiveness, and specifying and training an appropriately qualified workforce. Here in Step 10, we describe two additional standards not previously covered: readiness assessment and technical assistance. We also briefly discuss the importance of replicating findings under circumstances that are independent of the intervention's original developer.

## **TABLE 10.1**

Society for Prevention Research (SPR) standards for taking interventions to scale

Evidence	The SPR standards assume that interventions are not suitable for scale-up unless they have evidence demonstrating a positive and statistically significant impact on at least one child outcome that has practical significance in terms of public health impact. This evidence should support a clear statement involving who the intervention is for, what it has achieved and under what circumstances. This evidence should be substantiated through at least one efficacy and one effectiveness trial conducted to the rigour described in Steps 6 and 7 and the criteria outlined in the SPR standard. Ideally, this evidence should also have been achieved in circumstances independent of the intervention's developer.
Readiness assessment	The SPR standards recommend that it is desirable that the intervention providers be able to provide some guidance about the local systems and infrastructures necessary to successfully implement the intervention. This guidance can be provided through a psychometrically validated community assessment tool or through technical advice supported by research conducted as part of the intervention's efficacy or effectiveness trials. Processes for determining intervention readiness are described here in Step 10.
Intervention costs	The intervention should have developed guidance on potential costs that includes cost tracking and analysis tools. Guidance on collecting cost information is provided in Step 4, although it is assumed that this work will be ongoing as developers continue to test and refine their intervention materials, training costs and technical advice (see below). It is also desirable for interventions to be able to provide cost-effectiveness and cost-benefit information when available.
Materials	Interventions ready for scale-up should have developed materials that describe all of the activities involved in the set-up and delivery of the intervention, training manuals and any materials (books, audio-visual materials, etc.) required to deliver the intervention. The materials should also include a clear statement of the conditions necessary to implement the intervention, including characteristics of the setting, including the expected qualifications and experience of the practitioners delivering the intervention. Steps 3 and 4 cover the core principles underpinning these materials, although it is assumed that they will have undergone continual refinement through the ongoing research and development described in Steps 5 to 9.
Training	Interventions that are ready for scale-up should have developed and tested robust processes for training practitioners and certifying their learning. This training should cover the intervention's core theories, processes for implementing all active ingredients. The training should also provide practitioners with opportunities to practise new skills and receive feedback. Training should also cover permissible adaptations and processes that cannot be changed. Training should specify prerequisite practitioner skills and qualifications, as described in Step 8. The training should also be comparable to what was provided during the intervention's efficacy and effectiveness trials.

Technical assistance	The intervention must be able to provide post-training support to those setting up and delivering interventions. This support can include advice on how to 'install' the intervention through support and advice on workforce recruitment, referral systems, participant recruitment and engagement, staff supervision, and monitoring systems. Technical assistance can also take the form of additional teaching, coaching and practitioner certification. Methods for providing technical support through intervention purveyors, licensing and consulting are described in Step 10.
Fidelity monitoring	Tools must be available to those setting up the intervention to monitor fidelity of all aspects of the intervention's set-up and delivery. These tools should cover delivery and training targets, caseload management, recruitment, referral and retainment and fidelity to the intervention's content during delivery. Processes for monitoring intervention fidelity are covered in Step 4.
Monitoring reach	The intervention should provide clear eligibility criteria and systems for monitoring recruitment and retention against these criteria. Monitoring should include an assessment of local barriers to participation and identification of strategies to overcome these barriers. Recruitment efforts should be monitored on an ongoing basis and planning should be renewed as often as necessary to ensure high participation rates. Processes for monitoring recruitment, retention and reach are described in Step 4.
Monitoring effectiveness	Systems should be established to monitor the intervention's effectiveness as it is implemented through processes that collect data associated with the intervention's primary outcomes. The processes involved in setting-up monitoring systems are described in Step 8.
Evidence on adaptations	Systems for understanding the effectiveness of various intervention adaptations to the intervention's active ingredients should be ongoing as part of the intervention's development. The processes involved in adapting interventions and testing their effectiveness are described in Steps 8 and 9.
Normative data	Over time, the intervention should be able to provide normative data on desired levels of implementation milestones, fidelity, reach and effectiveness. The use of normative data within monitoring systems is briefly described in Step 8.

Source: Adapted from Society for Prevention Research

## **Readiness assessment**

Readiness refers to a local system's capacity to successfully adopt a new intervention. As described in previous steps, evidence-based interventions can be complex to set up and deliver, and often require substantial changes to local infrastructures to make them effective. Studies consistently show that local systems are unlikely to make these changes unless they are *ready* to make them. System readiness refers to three processes.

- System-wide motivation to adopt the new intervention. This means that there is a shared understanding across the system that the intervention is needed. This understanding then serves as the motivation for the organisation to initiate and see through the changes necessary to install and implement the new intervention so that it is effective and sustainable.
- A general capacity to adopt any intervention. This means sufficient financial resources, a suitably trained workforce, interagency relationships and processes for storing, sharing and analysing data.
- Intervention-specific capacity refers to the particular skills required to deliver the intervention to a high standard. This includes understanding the intervention's model and materials, necessary caseloads, methods of engaging and working with target population, familiarity with other similar models, and so on.

On the face of it, implementation readiness is not something that intervention providers have much control over. Nevertheless, by the time interventions have reached Step 10, there should be a reasonably good understanding of the local conditions that contribute to its successful implementation. This information can then be used to help local decision-makers determine whether the intervention has the potential to add value within their community,
which should in turn influence their motivation or 'readiness' to adopt and implement it. Readiness information comes in a variety of forms, including bespoke advice provided as technical assistance (see below) or checklists like the one provided for the PATHS programme in appendix F. The EIF Maturity Matrix also provides useful advice for local areas to self-assess their level of readiness.<sup>9</sup>

# Licensing, purveyors and other forms of technical assistance

Once the decision has been made to implement an intervention, a fair degree of work is still necessary to successfully install it within local systems. This often requires expertise that is well beyond what those delivering the intervention will have, even if highly detailed manuals and training processes have been produced. Implementation is therefore more likely to be successful if technical assistance is made available on an ongoing basis through licensing arrangements, a purveyor organisation, an intermediary organisation – or any combination thereof. Technical support can also be provided directly through consultation support from the developer, but this is less feasible when interventions are offered at scale by multiple organisations at the same time.

#### Licensing

Some interventions require licensing as a quality assurance mechanism to ensure that the intervention is set up and delivered to a high standard. Licensing requires that host agencies pay a fee and then engage in processes to demonstrate that they are meeting the licensing agreements. Processes typical of licensing arrangements include independent reviews of practitioner delivery (through videotapes or site inspections), quarterly reporting and the sharing of monitoring data.

Licensing arrangements are most effective when they are accompanied by consulting or technical assistance in return for the fee. This assistance can include helping local systems understand their population's demographic characteristics and need, strategies for setting up interagency referral routes, practitioner training and supervision, fidelity monitoring, outcome assessment, and data management.

Licensing may be offered through purveyors or independently of them. Although licensing can be viewed as expensive and cumbersome, it has several advantages for both the developers/providers and host agencies. From the perspective of the providers, it helps maintain a certain quality standard for the intervention's 'brand.' From the host agency's point of view, licensing, when accompanied by technical support, is an additional resource for local systems to deliver the intervention.

#### Purveyors

An intervention purveyor is an umbrella organisation set up specifically to provide ongoing technical support to all national or regional sites delivering an individual intervention. Purveyor organisations are often set up and managed by the original developer, but can also be delivered by a government unit or charity.

At the very least, purveyor organisations are responsible for the coordination and provision of training activities. However, they work best when they can also provide technical advice on the installation and ongoing delivery of the intervention. Ideally, this support should reflect the shared learning from other host agencies facing similar issues. Purveyor organisations provide less value when they only provide practitioner training.

<sup>&</sup>lt;sup>9</sup> See for example: https://www.eif.org.uk/resource/eif-maturity-matrix-0-19-years

#### Intermediary organisations

An intermediary organisation is similar to a purveyor, but it is not affiliated with any specific intervention model or brand. Rather, intermediary organisations are set up to provide advice on the installation and delivery of a variety of different evidence-based interventions and activities. This assistance may include processes developed to increase system readiness and understand local population need. Case study 11 uses the Communities that Care model as an example of the kinds of processes used by intermediary organisations to help communities understand their population needs, select interventions, install and deliver them.

#### **Case study 11: Evidence-based technical assistance provided through Communities that Care**

Communities that Care (CTC) is a model used by intermediary organisations to help local communities form partnerships to identify and implement preventive, evidence-based interventions to reduce youth crime and other risky adolescent behaviours. Through a five-phase model of technical assistance, CTC helps communities profile risk and protective factors within their population to then make decisions about the set-up and delivery of a portfolio of interventions with evidence of improving youth outcomes.

- **Get started**: Community leaders interested in beginning the CTC process define the scope of the prevention effort, assess their community readiness, address barriers to community readiness and engage key community stakeholders.
- **Get organised:** CTC leaders form a community board to oversee and implement the CTC process.
- Develop a profile: CTC leaders are trained to conduct a community needs assessment of risk and assessment factors through the use of the CTC validated community assessment tool. Risk and protective factors are assessed within four ecological domains: (1) the family, (2) school, (3) community, and (4) individuals, friends and peers. Data from this exercise are then analysed to prioritise risk and protective factors unique to the community.
- **Create a plan**: CTC leaders develop a plan to select and implement evidence-based programmes that are likely to reduce the risk and protective factors identified in the CTC profile.
- **Implement and evaluate**: The plan is implemented with the aim of supporting fidelity so that programmes can achieve their maximum effectiveness.

Once the plan is implemented, CTC becomes an ongoing community process. The process of monitoring implementation progress and community-level changes in risk, protection and youth outcomes is repeated every two years by reconducting the CTC needs assessment. Based on a review of the findings, CTC boards revise their action plans as needed.

CTC has undergone multiple, rigorously conducted RCTs in communities across the US demonstrating consistent reductions in young people's reports of illegal drug use and antisocial behaviour. Findings from its most recent trial demonstrate sustained improvements eight years after the partnerships were set up and three years following the end of technical assistance.

## **Independent replication**

It is a well-known phenomenon that evaluation findings are more positive when they are achieved through studies that have been conducted by or with the intervention's original developer. For example, a recent review of RCTs involving criminal justice interventions observed that effect sizes were twice as high in studies conducted by the intervention's original research team in comparison to studies conducted by independent researchers.

This phenomenon, also referred to as 'designer effects', occurs for a number of understandable reasons, including the fact that the original developer is often in the best position to set up and deliver the intervention. However, findings achieved under these circumstances have less real-world generalisability, so are not particularly useful for determining whether an intervention is ready to be delivered at scale.

## Step 10 checklist

The checklist below represents a range of activities that indicate that an intervention is ready to be delivered at scale. Completion of this list, however, should not be interpreted to mean that the intervention's evaluation journey is over. Rather, completion means that the intervention has begun a more iterative phase of testing and development to ensure that the intervention continues to provide meaningful benefits for children within a diverse range of populations and settings.

This testing and development will take place through many of the quality assurance mechanisms described in Steps 8 to 10, as well as further rigorous studies which aim to reverify previous positive findings, as well as investigate the circumstances under which the intervention's effects are strongest.

#### Evidence

- The intervention has evidence from at least one rigorously conducted efficacy trial conducted under ideal circumstances confirming short-term benefits for children in an outcome with public health significance.
- The intervention has evidence from at least one rigorously conducted effectiveness trial conducted in real-world circumstances confirming short-term benefits for children in an outcome with public health significance.
- At least one of the intervention's rigorously conducted trials provides evidence of child benefits lasting a year or longer.
- At least one rigorously conducted trial should have confirmed the intervention's child benefits by a research team independent of the intervention's original developer.

#### Quality assurance and technical assistance

- Detailed information about the intervention's costs should be available. This information should include all training, materials and licensing costs, as well as information about the ongoing running of the intervention. It is desirable for this information to provide an indicative unit cost per child during the course of the intervention.
- Readiness information is available about the conditions necessary to successfully deliver the intervention. Ideally, the intervention providers will also be able to provide advice on how the intervention addresses community needs.
- Materials for delivering the intervention are available. These include training manuals and implementation guidance, as well as all materials necessary for delivering the intervention.

A comprehensive package of practitioner training has been developed and evaluated. This training should specify desirable practitioner qualifications and experience and include processes for verifying that the information provided during training has been learned and properly applied. This package should also include high-quality 'train the trainer' mechanisms so the intervention can be offered at scale through multiple purveyor bodies at the same time.
Technical assistance on system readiness, and the set-up and delivery of the intervention is available through licensing, a purveyor organisation or intermediary support.
Information and processes for monitoring implementation fidelity is available through training, checklists and technical support.
The intervention has clear eligibility criteria and advice on how to recruit, refer and retain eligible participants. Systems are also available for monitoring intervention reach.
Systems have been developed for monitoring the intervention's impact on child and family outcomes on an ongoing basis. Ideally, these systems are supported by computer software that enables normative comparisons with findings from other host organisations and previous evaluations.
The intervention is underpinned by a programme of ongoing quality improvement that includes the rigorous testing of the intervention's active ingredients and various adaptations involving diverse settings and populations.

#### **Further resources**

- Collaborative for Academic, Social, and Emotional Learning, National Center for Mental Health Promotion and Youth Violence Prevention. (2011). Leading an SEL school: Steps to implement social and emotional learning for all students. 5/20/11, Education Development Center. http://www.promoteprevent.org/sites/www. promoteprevent.org/files/resources/leading\_an\_sel\_school.pdf
- Dymnicki, A., Wandersman, A., Osher, D., Grigorescu, V., Huang, L., & Meyer, A. (2014). Willing, able ready: Basics and policy implications of readiness as a key component for implementation of evidence-based practices. ASPE Issue Brief. Office of the Assistant Secretary for Planning and Evaluation, Office of Human Services Policy, US Department of Health and Human Services.
- Fixen, D. L., Blasé, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531–540.
- Glisson, C., & Schoenwald, S. K. (2005). The ARC organizational and community intervention strategy for implementing evidence-based children's mental health treatments. *Mental Health Services Research*, 7(4), 243–259.
- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926.
- Meyers, D. C., Durlak, J. A., & Wandersman, A. (2012). The quality implementation framework: a synthesis of critical steps in the implementation process. *American Journal of Community Psychology*, 50(3–4), 462–480.
- Rand. Getting To Outcomes<sup>®</sup>: Improving Community-Based Prevention. https://www.rand.org/health-care/ projects/getting-to-outcomes.html
- Spoth, R., Rohrbach, L. A., Greenberg, M., Leaf, P., Brown, C. H., Fagan, A., ... & Hawkins, J. D. (2013). Addressing core challenges for the next generation of type 2 translation research and systems: The translation science to population impact (TSci Impact) framework. *Prevention Science*, *14*(4), 319–351.

- Wandersman, A., Duffy, J., Flaspohler, P., et al. (2008). Bridging the gap between prevention research and practice: the interactive systems framework for dissemination and implementation. *American Journal of Community Psychology*, *41*, 171–181.
- Wanless, S. B., & Domitrovich, C. E. (2015). Readiness to implement school-based social-emotional learning interventions: Using research on factors related to implementation to maximize quality. *Prevention Science*, *16*(8), 1037–1043.

#### **Communities that Care**

- Oesterle, S., Hawkins, J. D., Kuklinski, M. R., Fagan, A. A., Fleming, C., Rhew, I. C., ... Catalano, R. F. (2015) Effects of Communities that Care on males' and females' drug use and delinquency 9 years after baseline in a community-randomized trial. *American Journal of Community Psychology*, 56, 217–228.
- Quinby, R. K., Hanson, K., Brooke-Weiss, B., Arthur, M. W., Hawkins, J. D., & Fagan, A. A. (2008). Installing the Communities that Care prevention system: Implementation progress and fidelity in a randomized controlled trial. *Journal of Community Psychology*, *36*, 313–332.
- Rhew, I. C., Hawkins, J. D., Murray, D. M., Fagan, A. A., Oesterle, S., Abbott, R. D., & Catalano, R. F. (2016). Evaluation of community-level effects of Communities that Care on adolescent drug use and delinquency using a repeated cross-sectional design. *Prevention Science*, *17*, 177–187.

## **Appendix A Strategies for searching the scientific evidence base**

Here we provide a list of strategies and tips for quickly locating the most recent and robust scientific evidence to confirm each of the theory of change (ToC) questions summarised in Step 1.

We must emphasise here that not all sources of information are equally robust. For example, the methods used for selecting articles in peer-reviewed journals are far more rigorous than those used in sector magazines or government reports. We therefore signpost to critical appraisal tools for determining the rigour of scientific methods and journal selection criteria at the end of this appendix.

## Q1: What is the intervention's primary intended outcome? And Q2: Why is the primary intended outcome important?

The primary aim of these questions is to justify why the intervention's intended outcomes are important from the perspective of children's development. Therefore it is useful to have an up-to-date understanding of the most recent scientific evidence regarding normal child development within your intervention's area of interest. A good place to start is to identify articles published in top academic journals summarising the most recent evidence. A list of the top journals associated with each of the seven EIF outcome domains is provided at the end of this appendix.

The landing page for all of these journals will have a search box. Here, you can type in key terms or phrases relevant to your programme, including the domain (such as cognitive development or child maltreatment), the child's age range (such as infancy, preschool or adolescence) and whether it pertains to parents or schools.

These searches will result in a list of titles and abstracts that can be reviewed. Journals will typically allow you to refine your search further, such as by topic (you may limit your 'physical development' search to 'adolescents'), date (you may choose a specific date range) or document type (you may choose journal articles, webpages or summaries).

Aim to find overview articles to help you understand child developmental milestones and processes. This will help you identify which primary outcome(s) are important, specific and realistic for your theory of change. Choose recent articles, written in the last 5–10 years. While journals frequently charge to download the full article, the title might be located for free on Google Scholar, or by contacting the author directly via Research Gate.

Online handbooks, encyclopaedias or compendiums also provide comprehensive overviews of child development within specific areas of interest. For example, the *Encyclopedia of Early Childhood Development*<sup>10</sup> contains up-to-date summaries of the most recent evidence in a variety of topic areas relevant to the early years. The *Handbook of Child Psychology and Developmental Science*<sup>11</sup> also provides comprehensive summaries of children's development within a wide range of relevant topic areas. Although the series is costly, authors will

<sup>&</sup>lt;sup>10</sup> See: http://www.child-encyclopedia.com/

<sup>&</sup>lt;sup>11</sup> See: https://bit.ly/2TCzmkR

frequently provide copies of their contributions for free, if contacted through their institutions or ResearchGate.

#### Q3: Why is the intervention necessary?

The answer to this question needs to explain how, from the perspective of the evidence, the intervention will support important developmental processes by addressing various developmental risks. Information about important developmental process is often available in the overview articles identified for answering questions 1 and 2. However, findings from cohort studies also provide a useful starting point for understanding how various risks impact developmental processes over time.

Birth cohort studies are population-wide observational studies that recruit families at the child's birth and then measure developmental outcomes at regular intervals throughout childhood. The design of cohort studies allows researchers to consider relationships between various risks and later development. The Millennium Cohort Study (MCS)<sup>12</sup> is an example of an ongoing UK cohort study tracking the development of approximately 19,000 children born in 2000–01. The MCS publications and resources tab<sup>13</sup> helps users to identify relevant information. A list of recent birth cohort studies in western countries is provided at the end of this appendix. Articles summarising the findings from cohort studies can also be found in the concluding list of top journals.

#### Q4: Why will the intervention add value?

The answer to this question should consider how the intervention will add value to children's development in comparison with the support that is currently available. In order to best answer this question, developers will need to have a good understanding of the strength of evidence underpinning interventions that are already available.

While information about the efficacy of individual interventions is often provided in the journals listed at the end of each step, systematic reviews also summarise what is known about the strength of evidence underpinning a group of practices with a similar aim. Systematic reviews are also quite useful for identifying where there are evidence gaps. The Cochrane Library<sup>14</sup> is a particularly useful resource for understanding the strength of evidence underpinning a wide variety of healthcare practices. The Campbell Library<sup>15</sup> similarly conducts reviews considering the evidence underpinning social interventions.

#### Q5: Who is the intervention for?

The answer to this question should carefully consider who the intervention is and is not for. This answer should cover, at the very least, the age range and level of need of the intervention's target population. Cohort studies are a good place to gain a preliminary understanding of children's needs at various age ranges. However, information pertaining to this question can also be found in articles summarising the findings from individual interventions (as described below) within the sections describing the characteristics of the study participants.

#### Q6: What will the intervention do? And Q7: How much of the intervention is required?

While the details of this question will not be fully answered until Steps 2 and 3 are completed, having as much knowledge of how similar interventions operate is a useful starting point for confirming the details of your own. 'What Works' clearinghouses, such as the EIF Guidebook,<sup>16</sup> are a good place to find out more about the ways in which effective

<sup>&</sup>lt;sup>12</sup> See: https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

<sup>&</sup>lt;sup>13</sup> See: https://cls.ucl.ac.uk/publications-and-resources/

<sup>&</sup>lt;sup>14</sup> See: https://www.cochranelibrary.com/

<sup>&</sup>lt;sup>15</sup> See: https://campbellcollaboration.org/library.html

<sup>&</sup>lt;sup>16</sup> See: https://guidebook.eif.org.uk/

interventions operate. Blueprints for Healthy Youth Development<sup>17</sup> and the Penn State University EPISCenter<sup>18</sup> also provide the details of various intervention's logic model.

For those wishing to identify additional studies relevant to their programme's ToC, Google Scholar or other relevant academic databases can be searched using key terms. The steps below provide a starting point for searching in Google Scholar for relevant texts. This does not constitute a systematic review process, more details of which are in the summary of resources section below.

- Work out what question you are trying to answer for example: 'What kinds of programmes are effective in improving children's social and emotional development?'
- 2. Begin by typing key words into Google Scholar.<sup>19</sup>
- **3.** Consider how helpful the results are for your purposes. You may find some useful results, but you may also wish to broaden your results and/or make them more relevant.
- 4. Have a play with the key words you are typing in. Develop a list of synonyms and alternative terms. You might use terms you have come across in handbooks or articles encountered when answering the previous questions. Try these words in different combinations.
- 5. If you find exactly the kinds of papers that answer your question, look at the key words in the title or abstract. This will help you to add to your list of alternative words.
- 6. Use these word combinations in a new search. Remember, the process is iterative.
- 7. You can also use 'operators' such as 'AND' and 'OR' to define relationships between groups of words:
  - OR broadens the search to retrieve any of the words it separates
  - AND **narrows** the search to retrieve **all** the words it separates.

Highlighting the key words and phrases in the question in different colours – 'What kinds of **programmes** are **effective** in improving **children's social and emotional development**?' – can help us to identify alternative words and phrases to include in the search.

#### TABLE A.1

Alternative words for iterating a key question search

Programme and effectiveness	Children	Social and emotional development
Effect	Child	Social and emotional competence
Intervention	Youth	Social and emotional learning
Programme	Adolescent	Mental health and wellbeing
Effective	Young person	Resilience
Evaluation	Children	Non-cognitive skills
Evaluate		Character skills
Impact		Soft skills
Result		

<sup>&</sup>lt;sup>17</sup> See: https://www.blueprintsprograms.org/

<sup>&</sup>lt;sup>18</sup> See: http://www.episcenter.psu.edu/

<sup>&</sup>lt;sup>19</sup> See: https://scholar.google.co.uk/

Example 'search strings' to type into Google Scholar might be:

- effective social and emotional learning programmes for children
- effective AND interventions AND children
- programmes AND character skills AND (adolescents OR young people OR child OR children).

Relevant search strings can also be used to answer the previous five questions, should the overview sources provide insufficient information.

As mentioned at the beginning, some sources and research findings are more robust than others, and we recommend you consider this when you are searching for sources for each of these questions. Risk of bias is increased through methodological features such as having a very small sample size in impact evaluations or disproportionately drawing on articles expressing a particular view in literature searches. There are many reasons why studies may not meet ideal methodological standards, often related to limited resources or being preliminary, pilot studies, prior to more rigorous research. However, it is worth considering how such factors will impact on your confidence in results, so you can select the best research for your purposes.

The EIF evidence standards (figure 4, Introduction) provide one method for appraising the risk of bias in impact evaluations. You can also use the Critical Appraisal Skills Programme checklists,<sup>20</sup> which provide guidance for assessing the robustness of a range of studies, depending on the research design used (for example, a literature review vs an impact evaluation).

#### **Further resources**

#### **Overviews of child development**

- The Handbook of Child Psychology and Developmental Science: https://bit.ly/2TCzmkR
- Encyclopedia on Early Childhood Development: http://www.child-encyclopedia.com/
- Harvard Centre on the Developing Child: https://developingchild.harvard.edu/
- Zero to Three: https://www.zerotothree.org/early-development
- Outcomes framework: An Equal Start: https://www.cypnow.co.uk/digital\_assets/an\_equal\_start.pdf
- Outcomes framework: Measuring What Matters: http://www.instituteofhealthequity.org/resources-reports/ measuring-what-matters-a-guide-for-childrens-centres/measuring-what-matters.pdf
- Handbook: Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities: http://www.prevencionbasadaenlaevidencia.com/uploads/PDF/RP\_Preventing\_young\_people\_ disorders\_NRCIM.pdf

#### **Birth cohort studies**

- The Avon Longitudinal Study of Parents and Children (ALSPAC; also referred to as Children of the 90s) has been tracking two cohorts of 14,000 children living in the Bristol area since 1991 and 1992: http://www.bristol. ac.uk/alspac/participants/
- The Canadian Healthy Infant Longitudinal Development (CHILD) study has been tracking the development of 3,400 children since 2009: http://childstudy.ca/
- The Dunedin Study in New Zealand has been tracking the development of 1,037 babies since 1972: https:// dunedinstudy.otago.ac.nz/
- Growing up in Australia has been following the development of 10,000 children since their births in 2003: https://growingupinaustralia.gov.au/research-findings
- Growing up in Scotland had been tracking the development of children in three cohorts: https:// growingupinscotland.org.uk/

<sup>&</sup>lt;sup>20</sup> See: https://casp-uk.net/casp-tools-checklists/

- Birth cohort 1: 5,217 children, born in 2004/05
- Child cohort: 2,858 children, born in 2002/03
- Birth cohort 2: 6,127 children, born in 2010/11
- The Millennium Cohort Study (MCS) has been tracking over 19,000 British children since 2000: https://cls.ucl. ac.uk/cls-studies/millennium-cohort-study/

#### Systematic reviews

- Cochrane Database of Systematic Reviews: https://www.cochranelibrary.com/cdsr/about-cdsr
- The Campbell Collaboration: https://campbellcollaboration.org/library.html
- EPPI Centre, a centre focusing on systematic reviews and research use: https://eppi.ioe.ac.uk/cms/. Among other resources, it contains an index of systematic reviews, with summaries, under the Publications tab.

#### What works clearinghouses

- The EIF Guidebook: https://guidebook.eif.org.uk/
- Blueprints for Healthy Youth Development: https://www.blueprintsprograms.org/
- Penn State University EPISCenter: http://www.episcenter.psu.edu/
- California Evidence-Based Clearinghouse for Child Welfare: http://www.cebc4cw.org/assessment-tools/ measurement-tools-highlighted-on-the-cebc/
- National Institute of Justice: https://www.nij.gov/Pages/welcome.aspx
- National Registry of Evidence-based Programmes and Practices (SAMHSA): https://www.samhsa.gov/nrepp
- Centre for Analysis of Youth Transitions (IFS): http://cayt.mentor-adepis.org/

#### **Critical appraisal**

 Critical Appraisal Skills Programme, which provides a number of critical appraisal checklists for assessing the robustness of a range of studies, depending on the research design used: https://casp-uk.net/casp-toolschecklists/

#### Journals

#### General child development

- American Psychologist: https://www.apa.org/pubs/journals/amp/
- British Journal of Developmental Psychology: https://www.apa.org/pubs/journals/amp/
- Child Development: https://onlinelibrary.wiley.com/journal/14678624
- Child Development Perspectives: https://onlinelibrary.wiley.com/journal/17508606
- Developmental Psychology: https://www.apa.org/pubs/journals/dev/
- Developmental Science: https://onlinelibrary.wiley.com/journal/14677687
- Infant and Child Development: https://onlinelibrary.wiley.com/journal/15227219
- Infant Behaviour and Development: https://www.journals.elsevier.com/infant-behavior-and-development
- PlosOne: https://journals.plos.org/plosone/
- Prevention Science: https://link.springer.com/journal/11121
- Psychological Bulletin: https://www.apa.org/pubs/journals/bul/

#### Cognitive development

- · Cognition: https://www.journals.elsevier.com/cognition
- Journal of Experimental Child Psychology: https://www.journals.elsevier.com/journal-of-experimental-childpsychology
- Social & emotional development, and self-regulatory/behavioural development
- Development and Psychopathology: https://www.cambridge.org/core/journals/development-andpsychopathology
- Journal of Consulting and Clinical Psychology: https://www.apa.org/pubs/journals/ccp/
- Journal of Child Psychology and Psychiatry: https://onlinelibrary.wiley.com/journal/14697610

#### **Physical development**

- British Medical Journal: https://www.bmj.com/
- Journal of the American Medical Association: https://jamanetwork.com/journals/jama
- The Lancet: https://www.thelancet.com/
- Pediatrics: http://pediatrics.aappublications.org/content/143/1?current-issue=y

#### Child abuse and neglect

- Child Abuse and Neglect: https://www.journals.elsevier.com/child-abuse-and-neglect
- Child and Family Social Work: https://onlinelibrary.wiley.com/journal/13652206
- Child Maltreatment: https://journals.sagepub.com/home/cmx
- Child Welfare: https://www.cwla.org/child-welfare-journal/
- Children and Youth Services Review: https://www.journals.elsevier.com/children-and-youth-services-review
- Development and Psychopathology: https://www.cambridge.org/core/journals/development-andpsychopathology

#### Substance misuse

- Addiction: http://www.addictionjournal.org/
- American Journal of Community Psychology: https://onlinelibrary.wiley.com/journal/15732770
- Frontiers of Psychiatry: https://www.frontiersin.org/journals/psychiatry
- Journal of Adolescence: https://www.journals.elsevier.com/journal-of-adolescence/
- Journal of Adolescent Health: https://www.jahonline.org/
- Journal of Community Psychology: https://onlinelibrary.wiley.com/journal/15206629
- Prevention Science: https://link.springer.com/journal/11121

#### **Risky sexual behaviour**

- American Journal of Community Psychology: https://onlinelibrary.wiley.com/journal/15732770
- · Journal of Adolescence: https://www.journals.elsevier.com/journal-of-adolescence/
- Journal of Adolescent Health: https://www.jahonline.org/
- Journal of Community Psychology: https://onlinelibrary.wiley.com/journal/15206629
- Prevention Science: https://link.springer.com/journal/11121

## **Appendix B Targeting interventions on the basis of need**

Interventions are most likely to achieve their greatest impact when offered to children and families who most need them. Public health frameworks commonly classify population needs in terms of **universal, targeted selected** and **targeted indicated** provision. This appendix provides an overview of the factors that should be considered when targeting interventions on the basis of child and family needs.

### **Universal support**

Universal support refers to activities that are offered to all children and families on a population-wide basis. As figure B.1 makes clear, universal coverage is often less intensive, but nevertheless aims to provide some benefits to all children. This means that universal interventions must be substantially better or different than what is currently available in order to make a difference that is measurable and meaningful. For example, a new maths curriculum will only be effective if it is substantially better than the curriculum children currently receive.

#### **FIGURE B.1** Population-wide levels of need Specialised Targeted indicated Families services support with identified needs Additional Families at risk support Basic coverage Population Intensity of intervention coverage

#### Source: EIF

As mentioned in Step 1, universal support is frequently not sufficient for meeting the needs of all children and families. More often than not, universal support is redundant for some and insufficient for others. This does not mean, however, that universal support cannot provide any measurable value. It can, but the benefits are frequently smaller and larger samples are required for these benefits to be observed.

It is beyond the scope of this guidance to describe in detail how to measure the impact of universal interventions with large samples. However, when developing a ToC, it is helpful to consider ahead of time the potential magnitude of an intervention's impact relative to the needs of its target population. The scientific literature provides a useful starting point for considering these issues.

## **Targeted selected interventions**

The term targeted selected pertains to activities that target or 'select' families with characteristics that place them at greater risk of experiencing problems. Studies show that these risks often interfere with many of the key processes (as covered in question 3 in Step 1) known to contribute to children's development. Targeted selected interventions are therefore offered to children and families on the basis of these risks with the aim of preventing problems from occurring in the first place. Although not exhaustive, figure B.2 provides an overview of many of the factors that either support children's development or place it at risk.

#### **FIGURE B.2**

Risk factors occurring in a child's life at the level of the child, family, community and society



Source: EIF

These factors exist at the level of the child, family, community and society.

- **Child-level factors** are those specific to the individual child. Child-level factors associated with children's development include gender, genetic inheritance, antenatal outcomes, physical health and maturation. Children's previous developmental outcomes also impact children's functioning at any given point in time.
- **Family-level factors** include the characteristics of the caregivers, interactions occurring within the home and processes directly impacting the family's day-to-day living. Examples of family-level processes known to be particularly associated with children's development include the caregivers' mental and physical health, the caregivers' beliefs, attitudes and knowledge about child rearing, the level of interparental conflict and the parents' conflict resolution skills, the size of the family, family income and the caregivers' level of education.
- **Community-level factors** involve the nature and quality of community resources. Community-level factors that particularly influence children's development include the quality of their educational experiences and the characteristics of their peer group. Other community factors found to be influential include families' access to healthcare, including mental health services, community safety, the quality of the housing stock and the quality of transportation systems.
- Society-level factors are those which influence the population's attitudes, beliefs and knowledge, and determine the resources that are available to communities and families. These include government policies which influence employment opportunities and citizens' access to education and healthcare.

The relative impact of risk factors will vary, depending on the child outcome and the presence or absence of other risk factors. Understanding the relative contribution of various risks factors on specific child outcomes can be challenging, but nevertheless essential for maximising an intervention's impact. The scientific literature provides an excellent starting point for understanding the impact of various risks on important child outcomes. Many EIF reports, such as *Foundations for Life*,<sup>21</sup> *Interparental conflict and outcomes for children in the contexts of poverty and economic pressure*,<sup>22</sup> and *Key competencies in children's early cognitive development*<sup>23</sup> provide introductory information about the relative impact of various risks on child development. Appendix A provides further guidance about where and how information about risk and protective factors and child development can be obtained.

When targeting interventions on the basis of risk, it is important to recognise that risks rarely exist in isolation. Intervention designs therefore need to include elements that address the presence of multiple risks, including procedures for referring children or caregivers on to other services. For example, interventions targeting children's behaviour are often made available on the basis of low family income. However, within this group, there is a greater chance of reaching participants who struggle with additional issues that may interfere with their ability to benefit from the intervention. Targeted selected interventions must therefore include elements that help more vulnerable participants overcome various barriers, or have systems in place to refer them on to other services when necessary. Methods for how to do this are described in greater detail in Step 8.

An intervention's ToC should also recognise when an intervention may not be appropriate for various participants. This means specifying who the intervention is *not* for. For example, studies repeatedly show that therapies offered to couples to strengthen the couple relationship may actually make things worse when domestic violence is present. In other

<sup>&</sup>lt;sup>21</sup> See: https://www.eif.org.uk/report/foundations-for-life-what-works-to-support-parent-child-interaction-in-the-early-years

<sup>&</sup>lt;sup>22</sup> See: https://www.eif.org.uk/report/interparental-conflict-and-outcomes-for-children-in-the-contexts-of-poverty-andeconomic-pressure

<sup>&</sup>lt;sup>23</sup> See: https://www.eif.org.uk/report/key-competencies-in-early-cognitive-development-things-people-numbers-and-words

words, traditional couples' therapy is *contra-indicated* when domestic violence is present. Targeted selected interventions made available to families on the basis of risk (for example, family income) should therefore identify when the intervention may not be appropriate by additionally specifying ineligibility criteria. The ways in which validated measures can be used for specifying intervention eligibility and ineligibility is described in detail in later steps.

## **Targeted indicated interventions**

The term targeted indicated pertains to interventions or activities that are made available to children or caregivers when they are 'indicated' – in other words, when the child or caregiver is coping with an identifiable problem. In this respect, the activity is no longer trying to prevent the problem from occurring in the first place, but is instead offered as a form of treatment or a way of preventing it from becoming worse. Examples of indicated problems commonly targeted by early interventions include behavioural issues, learning disabilities and childhood obesity.

In order to be effective, targeted indicated interventions need to have clear eligibility criteria that is used at the time children and caregivers are referred or recruited into the intervention. Eligibility criteria are often best defined through the use of validated instruments that can accurately assess the severity of the child or caregiver's problems, as well as measure progress.

The scientific literature contains a wealth of information about the availability of validated measures and their use in the assessment of targeted indicated problems. We also provide more information about the use of these measures for the purposes of evaluation in Step 5, as well as in many of the previously listed EIF reports.

## Appendix C The EIF five-phase participatory model for confirming a sciencebased theory of change

The primary purpose of a participatory ToC exercise is to help all stakeholders involved in the set-up and delivery of an established intervention develop and confirm a ToC that is consistent with the scientific evidence base. The five-phase model described in this section does this by providing clear structures for stakeholders to learn from each other, as well as the scientific evidence base. These structures include processes for stakeholders to:

- participate in the planning of the ToC exercise
- develop a shared understanding of the history of the intervention
- · share their views about the core purpose of the intervention and how it operates
- develop a shared understanding of the intervention's core purpose that is supported by scientific evidence
- confirm a ToC that is consistent with the evidence base and provides a solid framework for taking the intervention forward through future testing and evaluation.

We describe first how to conduct such an exercise under ideal circumstances, as well as provide suggestions for how it might be modified so that it is practical and affordable for a variety of different circumstances. We recommend that a team of *at least* three individuals external to the organisation be available to conduct the exercise. These individuals should have experience in facilitating workshops, conducting process evaluations and conducting systematic literature reviews.

It is worth keeping in mind that ToC exercises are useful for ensuring that all stakeholders understand the intervention's core purpose and that this remains consistent with the scientific evidence base. We therefore recommend that ToC exercises be repeated at key points throughout an intervention's development, and provide suggestions throughout this guidance as to when these exercises would be ideal. Repeating ToC exercises also provides flexibility for allowing stakeholders who may not have been available for earlier ToC exercises to be included in subsequent exercises. However, we feel strongly that the scientific evidence base be included as a primary stakeholder in every ToC exercise. **We also feel that it is important that organisations develop and maintain systems for consulting the scientific evidence base on an ongoing basis**.

## Phase 1: Planning the exercise

The primary purpose of the planning phase is to agree the details of the exercise, including its scope, participants, timetable and key deliverables. All of these details will vary depending on the project's budget, the number of stakeholders who will contribute, their availability and the size of the team conducting the exercise. A highly comprehensive ToC exercise will

include representatives from all of the intervention's stakeholder groups, including recipients of the intervention. The findings of this exercise will then be summarised in at least one final report. A less comprehensive exercise might include only the stakeholders involved in the design and delivery of the intervention and the findings will be verbally fed back to participants at a final workshop. Under either scenario, the result will be a written statement of the intervention's theory of change that documents its relationship to the scientific evidence base.

Decisions about the scope of the ToC exercise should be determined by:

- the history of the intervention: is it relatively new or well established?
- the intervention's reach: does it run in only a few local sites, or is it running nationally?
- the stakeholder groups involved in its delivery: is it still being offered by the original developer, or is it being delivered to multiple sites through a purveyor organisation?
- the purpose of the exercise: is it just to confirm the intervention's theory of change, or will the findings be used to inform other organisational decisions?

Decisions about the exercise's timetable will be based on whether or not the interviews and literature reviews will be conducted simultaneously or sequentially and whether findings will be presented at one or more workshops. Timetabling decisions will also be informed by participants' availability for attending the workshops and the depth interviews.

Decisions about deliverables include the number of workshops and their arrangements, the number of interviews that will be completed, the scope of the literature review, and whether written reports will be produced. At the very least, a ToC exercise should produce a ToC that is agreed by all relevant stakeholder groups. However, we strongly advise there is some written summary describing the outcomes of the exercise. It is also not uncommon for the findings from the literature review to be contained in a separate report.

### Phase 2: Documenting the intervention's history

Understanding the intervention's history is essential for planning the depth interviews and determining the parameters of the literature review. Depending on the scope of the exercise, this information can be gathered during a collaborative workshop, where the stakeholders provide this information via storyboard techniques. Alternatively, some of this information might be conveyed during meetings or email exchanges taking place during the planning phase. Any documentation about the intervention, including training materials, previous reports, evaluation findings and monitoring data will also be useful. At the very least, those conducting the exercise will need to know:

- How long has the intervention been running?
- Who currently manages the intervention?
- How is the intervention funded?
- What are the intervention's current objectives and what outcomes does it attempt to achieve?
- What are the intervention's core activities? What is its format and duration?
- Who delivers the intervention and what are their qualifications?
- How are those who deliver the organisation trained and supported?
- What other systems exist for implementing the intervention and ensuring quality?
- What is the intervention's reach? How many children/caregivers/families receive the intervention?

- What are the characteristics of the intervention's recipients? What is their age and level of need?
- What are the procedures for recruiting families?
- How much attrition occurs?

The team conducting the literature review will then use this knowledge to determine the search terms used to interrogate scientific literature to answer the seven questions described in the previous section. Information about the intervention's history will also be useful for determining the interview schedule that will be used during the depth interviews.

### Phase 3: Gathering the evidence

A primary objective of the five-phase model is to help those involved in the design and delivery of the intervention compare their views to the scientific literature, and ideally, those who receive the intervention. Two evidence-gathering activities should take place during this phase – qualitative depth interviews with representatives from the intervention's primary stakeholder groups, and the scientific literatures review. Both of these activities should gain answers to the seven questions set out in Step 1: 1. What is the intervention's primary intended child outcome? 2. Why is the primary intended child outcome important? 3. Why is the intervention necessary? 4. Why will the intervention add value? 5. Who is the intervention for? 6. What will the intervention do? 7. How much of the intervention is required?

#### Qualitative depth interviews

A comprehensive ToC exercise will gain the views of representatives from all of the intervention's key stakeholder groups, including those involved in the frontline delivery of the intervention (commissioners, managers, practitioners) and the intervention's recipients. The scope of the exercise will determine how many individuals participate in the exercise, but we recommend that representation should be proportionate to the size of the stakeholder group. Recruitment and consent procedures should be compliant with data protection requirements and standards for ethical consent. If the exercise will involve children and caregivers, additional ethical approval will be necessary. Further details about how to gain ethical approval are described in Step 5.

Ideally, the interviews should last no more than one hour and take place in person, although some could also be conducted over the phone. The interview schedules should be built around the seven questions, allowing sufficient time for follow-up to ensure that the participants understand the questions and can provide meaningful answers. For example, when asked *Why will the intervention add value?* participants should be allowed to answer this question first, and then prompted to give their views about how the intervention adds value to children's development if they do not initially provide this information.

It is worth noting that the depth interviews also provide an excellent opportunity to explore additional issues identified by the intervention's management during the planning phase. Additionally, it is worth noting that not all questions will apply to all participants. For example, the questions, *Who is the intervention for*? and *How much should they receive*? are unlikely to be relevant to recipients of the intervention.

If resources are scarce, the depth interviews could be replaced with focus group sessions. However, we strongly caution against this, as the outcome of group sessions frequently reflects the dominant views on the day, which can be biased towards the views of senior management. The ToC team may also choose to adopt a hybrid approach, where the views of some stakeholder groups will be gathered through individual interviews and others will be gathered through group sessions. For example, individual interviews might be conducted with members of the developer and intervention purveyor teams, and group interviews might be conducted with commissioner and practitioners. Similarly, group interviews might be conducted with intervention recipients – although group techniques are often less successful when participants are highly vulnerable. Regardless of the methods, the findings from the interviews and group sessions should be recorded and transcribed and then analysed through thematic methods.

#### The literature review

The literature review should be fit for purpose, meaning that it does not need to be a fullfledged systematic literature review, although it needs to be balanced and reflect the most recent evidence. This means that the researchers responsible for conducting the review should consider a variety of sources, including reputable journals, when answering each of the seven Step 1 questions.

Ideally, the results from the literature review should be written up in a report which summarises the evidence within each of the question categories. Alternatively, the findings can be shared in a presentation that will be reviewed at the learning workshop. Under any scenario, it is advisable for the review's findings to undergo some level of peer review from an expert in the field who is external to the intervention and the ToC exercise team.

## Phase 4: Learning from each other

Once the depth interviews and literature reviews are complete, the findings can be shared during a learning workshop involving the stakeholder groups who participated in the interviews, as well as other representatives from the intervention's organisation. We recommend that between two and four hours be allocated to this workshop – to ensure that there is sufficient time for the findings to be presented and discussed among the participants.

- **Presenting the findings:** We recommend that the findings from the depth interviews be presented first, followed by the findings from the literature review. This will allow participants to understand first how their views may be similar or different from each other's, and then compare them to what is known from the scientific evidence base.
- **Building consensus:** Once the findings have been shared, group discussions can be facilitated to discuss the meaning of the findings and identify steps in taking them forward to confirm the intervention's ToC.

## Phase 5: Confirming the theory of change

Once the findings have been shared and discussed, it is possible to confirm the intervention's ToC. This can take place on the same day as the learning workshop, but we recommend that this occur on a separate day to allow stakeholders to further reflect on the exercise's findings.

Confirmation should take place at a participatory workshop that includes representatives from all relevant stakeholder groups. These groups should include the 'evidence', who will be represented by a researcher or researchers involved in the literature review. The result should be a ToC that is fundamentally linear (see template at end of this appendix) and provides the following information for each theory of change.

• What is the intervention's primary intended child outcome? The ToC diagram should identify no more than two child outcomes that are logically linked to one of the seven outcome domains described in Step 1. Relevant short- and long-term outcomes should be identified and they should be consistent with what is known about children's development more generally. The outcomes should be justified by specific references to scientific articles.

- Why is the outcome important? The supporting ToC documentation should more fully explain why these outcomes are important from the perspective of children's development, using evidence to describe how earlier child behaviours contribute to longer-term child outcomes. The answer to this question should also consider any other shorter-term intervention outcomes that may be necessary, but not sufficient for the intervention to achieve its intended goals. For example, if a short-term outcome is to increase children's self-esteem, this section should be able to explain the relationship between children's self-esteem and the intervention's intended longer-term goals. This section should also reflect on whether change occurring on account of the intervention will be sufficient for the longer-term outcomes to be achieved.
- Why is the intervention necessary? The ToC should identify specific child development processes that the intervention will address, and the need for these processes should be justified by scientific evidence.
- Why will the intervention add value? The answer to this question will not be provided within the diagram but explained in the supporting documentation. The answer should include any gaps that may have been identified in the literature review, as well as any gaps in local provision.
- Who is the intervention for? The ToC should identify who the intervention is and is not for, not only in terms of recipient's characteristics (including the age of the child), but also their level of need (is the intervention universal, targeted selected or targeted indicated?). This target population should be justified by links to the scientific evidence.
- What will the intervention do and how much of the intervention will recipients receive? The ToC should include a brief explanation of the intervention's activities, its format, who will deliver it and how much recipients will receive. Links to the research literature should be used to justify the length and format of the intervention for its intended target population.

When answering these ToC questions, stakeholders will need to be self-critical in asking themselves if the answers are primarily driven by the evidence or their own personal opinions. We therefore recommend that the researcher representing the evidence take on the role of a 'critical friend'. Stakeholders should also be ready to critically examine the extent to which their intervention's outcomes provide sufficient benefits to children's development from a public health perspective. The intervention's short-term outcomes should also be plausibly linked to the intervention's longer-term child development goals and this relationship should be supported by scientific evidence.

Once the ToC has been confirmed at the workshop, it should be written up and the shared for further comments and sign-off. Gaining additional external peer review at this final stage will no doubt add credence to the final statement.

**TEMPLATE: LINEAR THEORY OF CHANGE** 



Source: Early Intervention Foundation, 10 steps for evaluation success: https://www.eif.org.uk/resource/10-steps-for-evaluation-success

## Appendix D Statistical power, effect sizes and t-tests

Here we provide brief advice on statistical methods that can be used to determine whether findings are statistically significant in pre/post pilot study.

# Hypothesis testing: Accepting or rejecting the null hypothesis

The research designs covered in Steps 5, 6 and 7 are set up to test the hypothesis that the intervention contributed to a child outcome in a way that was greater than what would be expected by chance. What occurs by chance is also referred to as the 'null hypothesis', expressed as  $H_0$ .

Statistical tests aim to avoid two different kinds of mistaken assumptions (see figure D.1): assuming that a relationship between phenomena did not happen by chance when it did (referred to as a type I error), and assuming that the relationship occurred by chance when it didn't (a type II error).

#### **FIGURE D.1**

Rejecting or accepting the null hypothesis (H<sub>0</sub>) and type I and II errors

	Do not reject H <sub>0</sub>	Reject H <sub>0</sub>
H <sub>0</sub> is true	Correct decision	Incorrect decision: Type I error or α
H <sub>0</sub> is false	Type II error or β	Correct decision

Statistical tests are designed to reduce the likelihood of type I and II errors. Convention dictates that an acceptable level of a type I error risk be less than 0.05, meaning the likelihood of the relationship occurring by chance is less than one out of 20.

However, researchers should be aware that the acceptability of this cut-off should be fundamentally informed by the distribution of the data and the risks associated with making type I and type II errors. For example, if the consequences of a type I error are life-threatening, then a cut-off of 0.001 - a 0.01% of mistakenly rejecting the null hypothesis – would be more appropriate.

130

## Testing for significance with a t-test

Many statistical tests are available to examine whether a statistically significant result exists. Researchers should choose a statistical test that matches with the nature of their data and the 'assumptions' of their chosen statistical test. For example, chi-squared tests compare differences in frequencies and are thus only appropriate for categorical data (for example, yes/no answers, gender categories, ethnic categories). T-tests compare differences between averages, and so can only be used with continuous data involving scores.

We describe a paired sample t-test in detail in this appendix because it compares two means which are paired in a natural way and is an ideal analysis for small-scale studies wishing to compare the pre and post scores of the same participants. The paired sample t-test (also called a matched sample or dependent sample t-test) determines the extent to which the difference between two sets of paired observations is zero. Paired sample t-tests assume that the dependent variable is measured on a continuous scale rather than a categorical scale and that the dependent variable is approximately normally distributed, among other assumptions. In case study 5, the authors used paired sample t-tests to evaluate the effects of the seminar by comparing the difference of the parents' ratings of their children's problematic behaviour before they attended the Triple P seminar and then again afterwards.

A t-test is relatively straightforward to calculate using statistical software, an online t-test calculator or Microsoft Excel (using the Data Analysis add-in).

#### Example

Imagine that this is your data. It shows pupils' scores on an academic test before and after an intervention (a teaching module).

#### TABLE D.1

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Sample data for a t-test

Source: Shier, R. (2004). Paired t-tests: Mathematics Learning Support Centre. http://www.statstutor.ac.uk/ resources/uploaded/paired-t-test.pdf

The t-test would be calculated as follows:

- Calculate the difference between the two observations in each pair (the difference between the pre-score and the post-score), making sure you distinguish between positive and negative differences.
- 2. Calculate the mean difference in the full sample.
- **3.** Calculate the standard deviation of the differences and use this to calculate the standard error of the mean difference.
- 4. Calculate the t-statistic, based on this the standard error and the mean difference.
- 5. Use tables of the t-distribution to ascertain the p-value for the paired sample t-test.<sup>24</sup>

### **Calculating the effect size**

The size of the difference between two means is often referred to as the *effect size*. An effect size reflects the importance or magnitude of difference between the outcomes in two samples or sets of scores (such as pre-intervention scores and post-intervention scores). Effect sizes allow us to go beyond the question 'Did it work?' and move to the question 'How well did it work in practical terms?' A common effect size, known as Cohen's *d*, is also relatively easy to calculate using statistical software or an online calculator.<sup>25</sup> The calculation (based on Walker, 2007) is as follows:

- 1. calculate the mean of the pre-intervention scores and post-intervention scores
- 2. calculate the difference between the means of the pre-intervention scores and postintervention scores
- **3.** calculate the average of the standard deviations of the pre-intervention scores and postintervention scores
- 4. divide the difference in means by the average of the standard deviations.

Figure D.2 demonstrates the effect size, *d*.



Source: Reproduced from Lenhard & Lenhard, 2016

<sup>&</sup>lt;sup>24</sup> For an online t-test calculator (paired sample t-test), see: https://www.socscistatistics.com/tests/ttestdependent/default.aspx

<sup>&</sup>lt;sup>25</sup> For an online effect size calculator, see: https://www.psychometrica.de/effect\_size.html

The effect size compares the mean of one sample or set of scores (as shown in purple) to the mean of another sample or set of scores (as shown in gold). This means that if we see a d of 1, we know that the two samples' means differ by one standard deviation; a d of 0.5 tells us that the two samples' means differ by half a standard deviation; and so on.

Cohen suggested that d=0.2 be considered a 'small' effect size, d=0.5 represents a 'medium' effect size and d=0.8 a 'large' effect size. This means that if two means don't differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant (Walker 2007).

### Power and sample size calculation

Statistical tests are more likely to detect significant differences when the sample size is larger, because the 'power' of the study increases. Statistical power refers to a study's ability to avoid a false negative, or type II error, which involves concluding that the intervention did not make a measurable difference when it actually did. The risk of type II errors is higher in studies involving small samples, because pre/post differences are difficult to detect unless they are relatively large – as previously explained in the section on effect sizes.

Sample size calculations depend on several factors and we recommend that developers liaise with experts to complete them. Generally, at least the following factors would be included in the calculation.

- **1. Acceptable level of significance:** normally *P* < 0.05.
- 2. Power of the study: normally 80%, which means that researchers accept that one in five times (that is, 20%) they will miss a real difference.
- **3. Expected effect size:** the smaller the expected effect size, the larger the sample required to detect the effect.
- **4. The population variance of a given outcome:** the greater the variation in the population (estimated by standard deviation), the larger the sample size needed. A smaller sample size would be acceptable for more homogeneous populations.<sup>26</sup>

The EIF level 2 threshold criteria require a minimum sample size of 20 to ensure a nominal degree of representativeness and increase the likelihood of observing a statistically significant impact. However, a sample size of 20 only has an 80% likelihood of observing this impact if the anticipated effect size is above 0.55. If the anticipated impact is lower – for example, 0.30 – than the chances of observing a statistically significant effect drop to 25%. Thus, **EIF's level 2 threshold sample size criterion of 20 is likely to be inadequate for interventions that anticipate small to moderate improvements in child outcomes.** 

#### Appendix references

- Kadam, P., & Bhalerao, S. (2010). Sample size calculation. International journal of Ayurveda research, 1(1), 55.
- Kirby, A., Gebski, V., & Keech, A.C. (2002). Determining the sample size in a clinical trial. *Medical Journal of Australia*, 177(5), 256–257.
- Lenhard, W., & Lenhard, A. (2016). Calculation of Effect Sizes. https://www.psychometrica.de/effect\_size.html
- Shier, R. (2004). Paired t-tests: Mathematics Learning Support Centre. http://www.statstutor.ac.uk/resources/ uploaded/paired-t-test.pdf
- Walker, I. (2007) 'Null hypothesis testing and effect sizes', Statistics for Psychology. http://staff.bath.ac.uk/ pssiw/stats2/page2/page14/page14.html

<sup>&</sup>lt;sup>26</sup> For an online power/sample size calculator, see: http://powerandsamplesize.com/

## **Appendix E Common threats to internal validity**

#### TABLE E.1

Eight common threats to internal validity

Threat/bias	Description	Example	Evaluation features to mitigate threat	Case study 6 features used to mitigate threat
Ambiguous temporal precedence	It is difficult to differentiate between the cause and effect between two variables. This threat is typically eliminated through experimental designs.	This typically occurs in retrospective studies which compare the outcomes between two groups at a single point in time. For example, an association between high levels of cannabis use during adolescence and the onset of schizophrenia is frequently observed, causing some to speculate that cannabis may increase the likelihood of schizophrenia. However, this association cannot rule out the likelihood that teenage children experiencing initial symptoms of schizophrenia may be more likely to engage in cannabis use at high rates, thus making it difficult to differentiate cause and effect (see Gage et al, 2017).	Prospective research designs that take place over a sufficient period of time to understand cause and effect relationships. Pre/post study designs.	Change was measured before and after the intervention and at a one- and three-year follow-up.
History	External events, other than the intervention, have the potential to influence the study results. This threat is potentially high if treatment and control group data are collected at historically separate time points.	The exam results of pupils in subsequent years before and after the introduction of a new curriculum. Although the pupils may be highly similar in terms of their characteristics, events occurring in the year the curriculum was introduced could bias the results for the intervention group.	Measurement for the intervention and comparison group take place at similar points in time.	Measurement for the intervention and comparison group take place at similar points in time.

Threat/bias	Description	Example	Evaluation features to mitigate threat	Case study 6 features used to mitigate threat
Maturation	During the study, change occurs due to the maturation of the participants, thus making it impossible to determine if change occurred on account of the intervention or participant maturation.	The self-control skills of Year 2 pupils receiving an intervention are higher than Year 1 pupils not receiving the intervention.	Intervention and comparison group are similar in all respects, including age. Measurement occurs at the same time for both groups, once before and then again after the intervention within a relatively short time period, and age-normed measures, as described in Step 5, are used.	Children in all four treatment groups were within the same one-year age range and age-normed instruments were used.
Testing or measurement effects	Exposure to a pretest or intervening assessment sometimes influences performance on a post-test, or make the subjects more sensitive to treatment.	Symptoms of depression are measured before and after an intervention. Post-intervention scores are then compared to those not receiving the intervention who completed the scale at only one point in time.	Control group participants complete the identical measures as the intervention group at similar points of time, despite not receiving the intervention so any effect from the pre-test should be observed in both groups and any difference between them is assumed to be due to the intervention.	Participants in all four groups completed the same set of measures pre/post intervention so presumably testing effects would be balanced across all groups. Changes in actual behaviours were also considered to reduce the likelihood of testing effects.
Instrumentation (or instrumentality)	Testing instruments or conditions are not sufficiently validated with the study population or inconsistently administered, contributing to non- random differences in measurement results.	Separate teams of researchers are used to administer measures for the intervention and comparison group.	The same set of researchers is used to administer assessment tools for both groups. Ongoing testing (inter- rater reliability) takes place to ensure that the researchers' scores are consistent with each other.	All self-report measures were completed by participants on pen and paper. Observational measures were carried out by the same researchers who retained high levels of inter-rater reliability and were blind as to whether the subjects they were observing were in the treatment or control group.
Experimenter bias	Experimenter bias occurs when the researchers who administer the measures affect the outcome by behaving differently with the intervention and treatment groups.	Researchers with a vested interest in a positive outcome inadvertently encourage intervention participants to respond in a more positive way, through coaching or other cues. Oftentimes, researchers may not be aware that they are doing this.	The best way to mitigate this is by ensuring that the researcher does not know or is 'blind' to whether the participant was assigned to the intervention or comparison group.	Measurement was conducted independently of the delivery of the intervention and researchers carrying out the observational assessment were blind to group assignment.
Participant bias (placebo effect)	Improvement in the intervention group occurs because participant perceives it to provide value independently of the intervention's content.	A common example of a placebo effect is participants perceiving improved symptoms after taking a sugar pill instead of a potent drug.	Participants do not know their intervention or comparison group assignment.	Participants had knowledge of whether they were assigned to a treatment or control group. The extent to which they were aware that there were three separate treatment groups was not reported, however.

Threat/bias	Description	Example	Evaluation features to mitigate threat	Case study 6 features used to mitigate threat
Regression to the mean	The statistical tendency for very high or low pre- test scores to 'regress' towards the average during retesting.	Highly depressed patients demonstrate artificial improvements on independent measures because it is not possible for score to become any higher and levels of error within the measure.	The use of a comparison group and a sufficiently large sample to increase the range of the scores and ensure the number of very low and high scores are small relative to the overall sample.	Baseline analyses revealed that the group was highly vulnerable, although nevertheless heterogeneous.
Selection bias	Systematic differences exist in the characteristics between intervention and comparison group participants. These differences may or may not be known at the time the groups were selected, although they are often unknown.	More physically fit adults signing up for a weight-loss intervention.	Study participants do not have the opportunity to select treatment and agree to random assignment.	Random assignment was used to assign participants to the control and three treatment groups.
Differential attrition	As people leave the study the characteristics of those who remain in the intervention are different than the characteristics of those who remain in the treatment group, thus introducing selection bias and reducing the groups' comparability.	It is not uncommon for more vulnerable families to have difficulty completing an intervention, resulting in higher levels of study attrition in comparison to the control group. The higher the levels of differential attrition, the greater the likelihood the groups will no longer be equivalent. High levels of attrition also suggest that the intervention may not be feasible for the participants who would most benefit from it.	The study design should include features to increase the likelihood that participants are retained in the intervention and the study. Studies should also make use of 'intent-to-treat' designs that – where possible – include pre/post tests of all participants regardless of their level of participation in the intervention. Statistical analyses should also be conducted to determine the extent that differential attrition took place and that the groups have remained equivalent. Small levels of differential attrition can be controlled for through statistical analyses.	Attrition records were retained and comparisons of the characteristics and baseline scores of participants were made to verify for differences. Some differences were observed, demonstrating that more vulnerable families were more likely to leave the intervention conditions. The authors recognise this as a potential threat to the validity of the findings, particularly the size of the impacts.
Diffusion of treatments (contamination)	Implementation of the intervention has wider system impacts which affect the control group.	Diffusion can occur as a result of an intervention introducing practice changes throughout a service. For example, it is common for interventions to include quality assurance features (see Steps 4 & 8) that are then adopted across the entire service, reducing the relative value offered by the intervention.	This can be difficult to mitigate against, but should be taken into account when interpreting study findings.	Families were ineligible for the intervention if they were attending other interventions.

Source: Adapted from Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

#### **Appendix references**

- Gage, S. H., Jones, H. J., Burgess, S., Bowden, J., Smith, G. D., Zammit, S., & Munafò, M. R. (2017). Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample Mendelian randomization study. *Psychological medicine*, 47(5), 971–980.
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

## Appendix F Readiness tool: Promoting Alternative Thinking Strategies (PATHS®)

During programme selection, it is important to research and compare possible evidencebased prevention programmes to ensure that you are selecting the programme that is the best fit for your community and available resources. This tool is intended to help you plan for the implementation of an evidence-based programme with a commitment to quality and long-term sustainability.

### **Programme selection**

Is PATHS<sup>®</sup> a good fit for your community? While researching and comparing possible programmes, consider the following information to determine if PATHS<sup>®</sup> is the best fit for the community.

## **1.** Are these the risk and protective factors identified and/or prioritised by your community?

PATHS® Targeted Risk Factors	PATHS® Targeted Protective Factors		
Aggressive behaviour	Emotional awareness and understanding		
Impulsive behaviour	Emotional regulation		
Poor problem-solving skills	Social skills		
Poor social skills	Problem-solving		
Early initiation and persistent antisocial behaviour	School opportunities for prosocial involvement		
Poor academic performance	School rewards for prosocial involvement		
Low school commitment	Interaction with prosocial peers		
Favourable attitudes towards antisocial behaviour			

#### 2. Is PATHS® appropriate for the population you plan to target?

• PATHS<sup>®</sup> is intended to be taught to children in preschool through Grade 6 as part of the regular school day.

#### 3. What are the outcomes you intend to change in your targeted population?

PATHS® has been shown to have the following outcomes:

- Short-term:
  - improved emotion knowledge & awareness
  - improved self-control

- improved social problem-solving
- improved social & emotional skills
- improved cognitive abilities (executive functions).
- Long-term:
  - improved school readiness
  - reduced antisocial behaviour.

## **Appendix G General evaluation resources**

#### Websites

- The Magenta Book: https://www.gov.uk/government/publications/the-magenta-book
- Handbook on Impact Evaluation: Quantitative Methods and Practices: https://openknowledge.worldbank.org/ handle/10986/2693
- Proving and Improving: a quality and impact toolkit for social enterprise: http://www.nefconsulting.com/ourservices/evaluation-impact-assessment/prove-and-improve-toolkits/
- The Social Return on Investment Network: http://www.socialvalueuk.org/
- Basic Guide to Outcomes-Based Evaluation for Nonprofit Organizations with Very Limited Resources: http://managementhelp.org/evaluation/outcomes-evaluation-guide.htm
- Child Outcomes Research Consortium: https://www.corc.uk.net/outcome-experience-measures/

#### **Books and papers**

Asmussen, K. (2011). The evidence-based parenting practitioner's handbook. Routledge.

- Gottfredson, D.C., Cook, T.D., Gardner, F.E., Gorman-Smith, D., Howe, G.W., Sandler, I.N., & Zafft, K.M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926.
- Martin, J., McBride, T., Brims, L., Doubell, L. Pote, I., & Clarke, A. (2018) *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Early Intervention Foundation. Available: https://www.eif.org.uk/ resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

Patton, M.Q. (1997). Utilization-focused evaluation: The new century text, 3rd edn. Sage Publications.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Weiss, C.H. (1998). Methods for studying programs and policies. Prentice Hall.