EARLY INTERVENTION FOUNDATION

Evaluating early intervention programmes

Six common pitfalls, and how to avoid them

February 2018

Jack Martin, Tom McBride, Lucy Brims, Lara Doubell, Inês Pote, Aleisha Clarke

Acknowledgments

For their contributions to the preparation of this report, we are very grateful to Kirsten Asmussen (Early Intervention Foundation), Mark Ballinger (EIF), Daniel Acquah (EIF), Nick Axford (University of Plymouth), Raj Chande (Behavioural Insights Team), Liam O'Hare (Queen's University Belfast), Matthew van Poortvliet (Education Endowment Foundation), and Lizzie Poulton (Zippy's Friends).

Contents

Early Intervention Foundation
10 Salamanca Place
London SE1 7HB

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

This paper was first published in February 2018.

High-quality evidence on 'what works' plays an essential part in improving the design and delivery of public services, and ultimately outcomes for the people who use those services. Early intervention is no different: early intervention programmes should be commissioned, managed and delivered to produce the best possible results for children and young people at risk of developing long-term problems.

EIF has conducted over 100 in-depth assessments of the evidence for the effectiveness of programmes designed to improve outcomes for children. These programme assessments consider not only the *findings* of the evidence – whether the evidence suggests that a programme is effective or not – but also the *quality* of that evidence. Studies investigating the impact of programmes vary in the extent to which they are robust and have been well planned and properly carried out. Less robust and well-conducted studies are prone to produce biased results, meaning that they may overstate the effectiveness of a programme. In the worst case, less robust studies may mislead us into concluding a programme is effective when it is not effective at all. Therefore, to understand what the evidence tells us about a programme's effectiveness, it is also essential to consider the quality of the process by which that evidence has been generated.

In this guide, we identify a set of issues with evaluation design and execution that undermine our confidence in a study's results, and which we have seen repeatedly across the dozens of programme assessments we have done to date. To help address these six common pitfalls, we provide advice for those involved in planning and delivering evaluations – for evaluators and programme providers alike – which we hope will support improvements in the quality of evaluation in the UK, and in turn generate more high-quality evidence on the effectiveness of early intervention programmes in this country.

Six pitfalls, at a glance

See the following 'In detail' sections for further explanation and definition of key terms.

1 No robust comparison group

Problem: A robust comparison group is essential for concluding whether participation in a programme has caused improvements in outcomes. However, some studies do not use a comparison group at all; others use a comparison group which is not sufficiently robust, biasing the results.

Solution: Evaluators should endeavor to use a comparison group in impact evaluations. Ideally this should be generated by random assignment (as in a randomised control trial, or RCT), or through a sufficiently rigorous quasi-experimental method (as in a quasi-experimental design studies, or QED).

2 High drop-out rate

Problem: Attrition – the loss of participants during an evaluation – can introduce two problems: the study sample may become less representative of the target population, and the intervention group and control group may become less similar. These biases can result in misleading conclusions regarding programme effectiveness or the applicability of findings to the target population.

Solution: There are a range of measures to improve participants' cooperation with data collection, such as financial compensation. In addition, researchers can conduct analyses to verify the extent to which attrition has introduced bias and report any potential effects on the results.

3 Excluding participants from the analysis

Problem: Excluding participants from data collection and analysis due to low participation in the programme risks undermining the equivalence of the intervention and control groups, and so biasing the results. Bias can also arise from excluding control group participants who receive some or all of the programme that is being evaluated.

Solution: Evaluators should attempt to collect outcome data on all participants and include them in the final analysis of outcomes, regardless of how much of the programme was received. This maintains greater similarity between the intervention and control group, and so is less likely to produce bias.

4 Using inappropriate measures

Problem: Using measures which have not demonstrated validity and reliability limits our confidence in an evaluation's findings and conclusions. Validity is the extent to which a measure describes or quantifies what is intended. Reliability is the extent to which it consistently produces the same response in similar circumstances.

Solution: Researchers should use validated measures which are suitable for the intended outcomes of the programme, and appropriate for the target population.

5 Small sample size

Problem: If there are not enough participants in the study it is hard to have confidence in the results. Small sample sizes increase the probability that a genuinely positive effect will not be detected. They also make it more likely that any positive effects which are detected are erroneous. In addition, smaller sample sizes increase the probability that the intervention and control groups will not be equivalent in RCTs.

Solution: Researchers need to be realistic about the likely impact of their programme and potential attrition, and to use power calculations to identify the appropriate sample size. Use strategies to recruit the correct number of participants and retain them in the study, such as financial compensation. EIF will not consider evaluations with fewer than 20 participants in the intervention group.

6 Lack of long-term follow-up

Problem: Studies which do not assess long-term outcomes (at least one year post-intervention) – or do not assess them well – cannot tell us if short-term effects persist. Long-term outcomes are often the most important and meaningful outcomes, in terms of the ultimate goal of the programme.

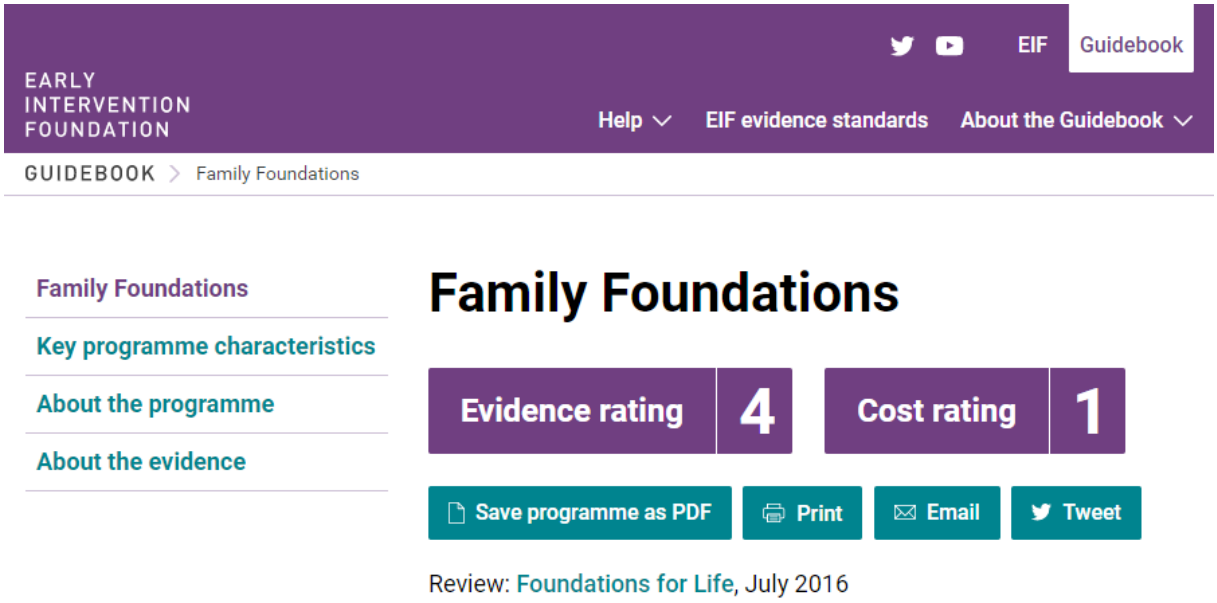
Solution: Researchers should plan data collection to capture both potential short- and long-term outcomes. Guard against problems which are particularly likely to damage the quality of long-term outcome analyses: maintain comparison groups, attempt to minimise attrition, and conduct analysis to account for attrition.

Introduction: EIF Guidebook, programme assessments & evidence ratings

EIF's online Guidebook is a key tool for supporting the delivery of effective services.[1] The Guidebook provides independent information on the effectiveness and delivery of early intervention programmes. We assess programmes according to the strength of their evidence – the quality and quantity of data suggesting that a programme has had a positive impact on outcomes for children – and give each programme an evidence rating based on this assessment.[2]

FIGURE 1

An evidence rating on the EIF Guidebook



Source: EIF Guidebook

Our assessment is based on 34 criteria that describe the essential components of high-quality evaluation of impact, including study design, sample properties, measurement and analysis. These 34 criteria have been set by EIF, and similar standards are held by other organisations which assess evidence, known as clearinghouses.

For more information on other clearinghouses, see the list of useful resources at the end of this guide.

1 See: http://guidebook.eif.org.uk/

2 For more on the evidence ratings and what they mean, see: http://guidebook.eif.org.uk/guidebook-help/how-to-read-the-guidebook

Programmes published on the EIF Guidebook have evidence ratings ranging from 2 to 4+. An evidence rating of 3 indicates the level at which we can assert with confidence that participation in a programme has caused improvements in outcomes. We call these 'evidence-based programmes'.

However, there are over 30 programmes on our Guidebook which were eligible for a rating of 3 (based on the type or design of the evaluation study) but failed to achieve it because of issues with the way the evaluation was conducted or reported. This is a missed opportunity to add high-quality evidence to what we know about effective early intervention in the UK.

Conducting high-quality evaluation is challenging. We want to be part of the solution, by supporting a step-change in the quantity and quality of impact evaluation in the UK. In this guide, we identify six common pitfalls that we see in the planning, conducting and writing up of evaluation studies which limit our ability to determine with confidence whether a programme has been effective, and set out tactics and methods for avoiding or accounting for each of these issues.

Lastly, although these lessons are taken from our assessment of programmes designed for children and families, the lessons are applicable to impact assessment in all areas of social policy.

Identifying our six evaluation pitfalls

We have identified six common evaluation pitfalls. To those involved in evaluation, they may seem obvious. Nonetheless, these are issues that we encounter often in our assessment work and which frequently reduce our confidence in study results, culminating in a lower-than-expected evidence rating on our Guidebook.

Our six pitfalls were chosen on the basis of:

- **Frequency:** how often this issue has been encountered in our assessments.
- **Impact:** how frequently this issue has resulted in a programme receiving a lower evidence rating than it might have done had the issue been addressed satisfactorily.
- **Usefulness:** the extent to which practical guidance on addressing this issue can be provided in a short and accessible way.

Pitfalls were also selected for balance across EIF's levels: roughly half are issues that frequently cause programmes to not achieve a level 2 (or 'preliminary') evidence rating, and the remainder frequently hold programmes back from achieving level 3 and 4 ('evidence-based') ratings.

Some important considerations...

- This is not an exhaustive account of every issue that can occur with impact evaluation.
- This guidance is intended to support evaluation of relatively well-developed programmes. We do not provide guidance here to support the initial stages of programme development, when crucial decisions are still being made about the theory of change and logic model. However, some evaluation issues discussed here (such as high attrition or poor recruitment) could be related to issues in the design and delivery of the programme itself, and so should be considered at the early stages of programme piloting.
- While the pitfalls discussed here relate primarily to evaluation design and execution, we also cover associated issues with how evaluations are reported: what information is included in the write-up of the evaluation, and how transparent the researchers are about

the methods they have used and what has happened over the course of the evaluation. In practice, EIF can only assess evidence on the basis of what is presented by the researchers. This means that poor reporting can result in a programme receiving a lower evidence rating than it would have done had good reporting standards been adhered to.

For more information on reporting standards, see the list of useful resources at the end of this guide.

- We acknowledge that EIF's evidence rating and inclusion in the EIF Guidebook will often not be the primary motivation for conducting an impact evaluation. In some cases, failure to adhere to our standards will be a conscious choice, driven by consideration of costs and resources or where a programme is at in its development cycle. Nevertheless, we believe that adhering to our standards does increase the confidence that a reader can have in the findings of an evaluation. And so, in the absence of any compelling reason not to comply, our standards represent a good rule of thumb for driving up the quality of impact evaluation in the UK.
- We recognise that some of these issues are likely to arise in many cases because studies are insufficiently funded, for example, to include a robust comparison group or to conduct a long-term follow-up. Our guidance on common evaluation pitfalls is therefore pertinent not only to evaluators and programme providers, but also to funders, who can and should ensure that adequate resources are available to conduct high-quality evaluations as part of the projects and initiatives that they support.

Pitfall 1: No robust comparison group

Why does the lack of a robust comparison group undermine confidence in an evaluation's findings?

To know what impact a programme has had, we need to know the outcomes of the participants who have received the programme. But we also need to be able to estimate what would have happened to these participants if they had *not* received the programme. This is known as the 'counterfactual'.

Knowing what would have happened in the absence of a programme is not easy, because we cannot observe the same group of participants simultaneously receiving and not receiving the programme. However, we can use evaluation methods to *estimate* what would have happened in the absence of the programme, and so ascertain whether participation in a programme can be causally linked to any improvements in outcomes (HM Treasury 2011; Khandker, et al, 2009). This is done using a comparison group: a group of individuals who do not receive the programme but are otherwise very similar to those receiving the programme.

Some evaluations do not use a comparison group at all. At EIF, we see many 'one-group pre-post' studies which, instead of a comparison group, estimate impact by taking a group of participants who have received the programme and comparing their outcomes before and after the programme has been delivered. Any change between these two points in time is then attributed to the impact of the programme. This approach has advantages in terms of ease, especially where there is a need to evaluate a programme that has been rolled out nationally, in which case no obvious, robust comparison group exists.

With pre-post studies, however, it is impossible to control for all potential sources of bias so that we can attribute improved outcomes to participation in the programme. It is always possible that factors other than the programme are responsible for any observed improvements. For example, a one-group pre-post evaluation of a child behaviour programme cannot rule out the possibility that any improvement over time is simply due to children becoming more compliant as they age and mature, rather than the programme itself. Indeed, promising findings observed in these preliminary one-group pre-post studies are often not replicated when more rigorous evaluations are conducted (Deeks et al, 2003; Shadish et al, 2002).

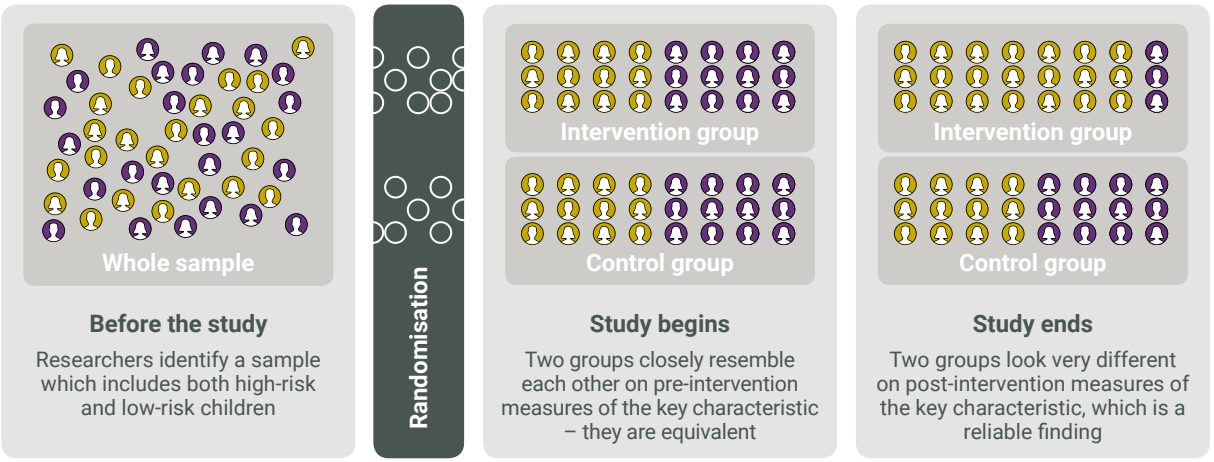
This kind of study may represent an important step in programme development and in preparing for more rigorous evaluation in the future (for example, by exploring a programme's potential impact). Ultimately, however, robust impact evaluation requires a comparison group, and so programmes that want to be considered 'evidence-based' must consider further evaluation using more robust designs once promising findings have been identified in pre-post studies.

Comparison group studies involve comparing the outcomes of an intervention group (which receives the programme) to those of a control group, which does not receive the programme but is otherwise equivalent to the intervention group in important characteristics, such as demographics.

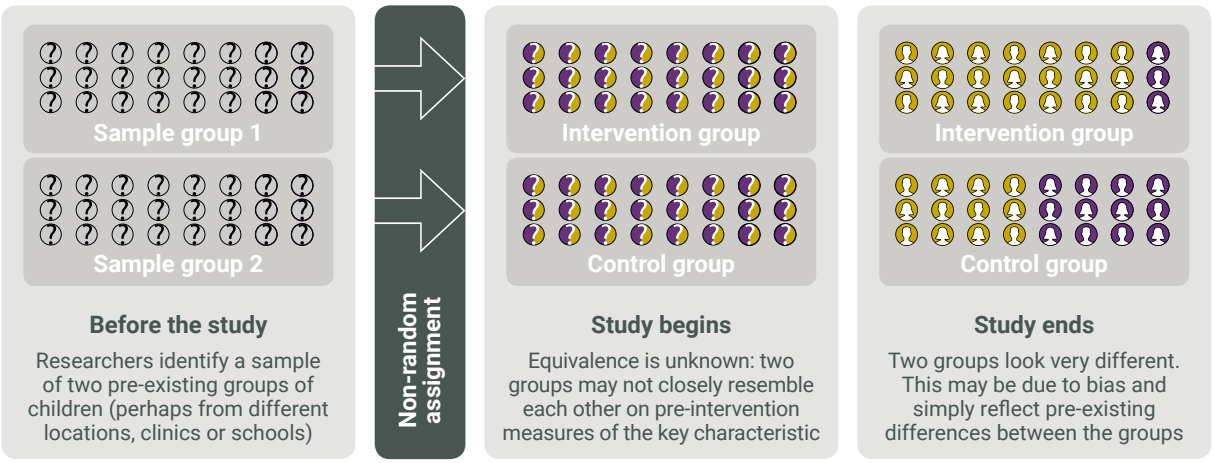
Simply having a comparison group, however, is not enough. We see many studies that make use of a comparison group but do not use sufficiently rigorous methods to generate it, meaning that the comparison group is not sufficiently similar to the intervention group – that is, it is not a robust comparison group. If the intervention and control groups are different in their characteristics, an evaluation cannot tell us if changes are due to the programme or these pre-existing differences between the two groups (Shadish et al, 2002). For example, an evaluation of an anxiety prevention programme which found that the intervention group had lower anxiety scores than the control group after the programme was delivered might lead us to conclude that the programme had worked to reduce anxiety. However, if participants in the intervention group were more likely to be living in less stressful circumstances than those in the control group, the control group's higher scores may reflect this pre-existing difference, rather than any programme impact. See **figure 2** for an illustration of this risk.

FIGURE 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

How can this issue be addressed?

Researchers should endeavor to use a good comparison group when evaluating programme impact. The best designs ensure that groups examined as part of the evaluation are equivalent on key characteristics before the programme is delivered.

There are two main methods for ensuring equivalence:

- **Randomised control trial (RCT):** RCTs ensure equivalence by randomly assigning participants to the intervention and control groups, ensuring that there is no systematic differences between the two study groups on any characteristics (although they may differ by chance) and therefore that they are equivalent in all respects except that one group receives the programme (Shadish et al, 2002). It is important that researchers report the **type** of randomisation used (such as simple randomisation or stratified randomisation), the **method** of randomisation (eg, coin toss or random number table), and **how randomisation was implemented** (eg, sequentially numbered envelopes or central randomisation service).
- **Quasi-experimental designs (QEDs):** QEDs use statistical methods such as propensity score matching to ensure that the outcomes of the group that receives the programme are compared to those of a comparison group that is equivalent on important measured characteristics. While a well conducted RCT is likely to ensure equivalence between the groups across *all* characteristics (the key benefit of randomisation), a quasi-experimental design can typically only guarantee equivalence on characteristics which are measured by the researcher and are included in the statistical procedures of the study (Shadish et al, 2002).

For more information on study design, see the list of useful resources at the end of this guide.

What are the implications for EIF's evidence rating?

- For a study to be considered robust (rated 3 or higher), it must be a comparison group study. Participants should be randomly assigned to intervention and control groups (RCT), or sufficiently rigorous methods should be used to generate an appropriately comparable sample through non-random methods (QED).
- A one-group pre-post study can achieve level 2 on our standards, which we describe as having preliminary evidence.

Case study 1: Comparison groups

Issue: EIF reviewed an evaluation of an early intervention programme which compared parents from sites running the programme to parents from sites which did not offer the programme. Although this study had a comparison group (and so was an improvement upon a one-group pre-post only design), the sites were chosen on the basis of whether or not they were providing the programme and because they were geographically close to each other. It is possible that parents who register with a site where an early intervention programme is available will be different to parents who register somewhere it is not available on important characteristics, which may lead to bias in the results. Indeed, in this case, statistical tests confirmed that the groups were different with respect to important variables such as child behaviour. Therefore the positive results that were observed could be due to pre-existing differences between the groups, rather than the effectiveness of the programme itself.

What could have been done to address the issue: To increase confidence in the results, researchers would have needed to randomly assign parents, or groups of parents, to the intervention and control groups, or use robust QED methods to construct equivalence between the two study groups.

EIF evidence rating: Because sufficiently rigorous methods were not used to generate the comparison group, this evaluation could not be considered to provide fully robust (level 3) evidence, and so this programme received an evidence rating of 2 (indicating 'preliminary evidence').

Pitfall 2:
High drop-out rate

Why does a high drop-out rate undermine confidence in an evaluation's findings?

Drop-out – or attrition – refers to the loss of participants who are recruited into a study but are not included in the final analysis, as researchers have been unable to collect data from them.

There is an important difference between someone dropping out of the study and dropping out of the programme. For example, a research participant can stop attending programme sessions (drop out of the programme) and yet still continue to provide data (not drop out of the study); on the flipside, another person may continue to receive the programme and yet refuse to provide data or participate in the study.

Attrition can occur for various reasons, including researchers losing track of study participants and participants refusing to take part in data collection.

Attrition can have two major types of consequences for the robustness of the evaluation:

- **Unrepresentativeness:** This is an issue if certain types of participants are more likely to leave the study than others, meaning that the evaluation sample becomes less representative of the programme's target population, and that the findings can only tell us about specific groups of people rather than a broader population. For example, the evaluation findings of a programme designed to support the general population would be undermined if disadvantaged participants were much more likely to drop out. The findings of this study would be based on a sample of relatively less disadvantaged participants, and so could not be generalised to apply to a more diverse group.
- **Bias and non-equivalent groups:** Random assignment or quasi-experimental techniques are designed to produce study groups that are equivalent on key demographic and outcome variables. However, attrition may undermine this if certain types of participants are more likely to leave the intervention group than the control group, or vice versa. This is problematic, as any differences between the groups on outcomes may reflect differences between the types of participants retained in each group, instead of the true impact of the programme. See **figure 3** below.

How can this issue be addressed?

Due to the risks listed above, researchers should always aim to minimise attrition as far as possible. Attrition can be lessened by using a range of measures to increase participants' cooperation with data collection and reduce logistical challenges (Brueton et al, 2011):

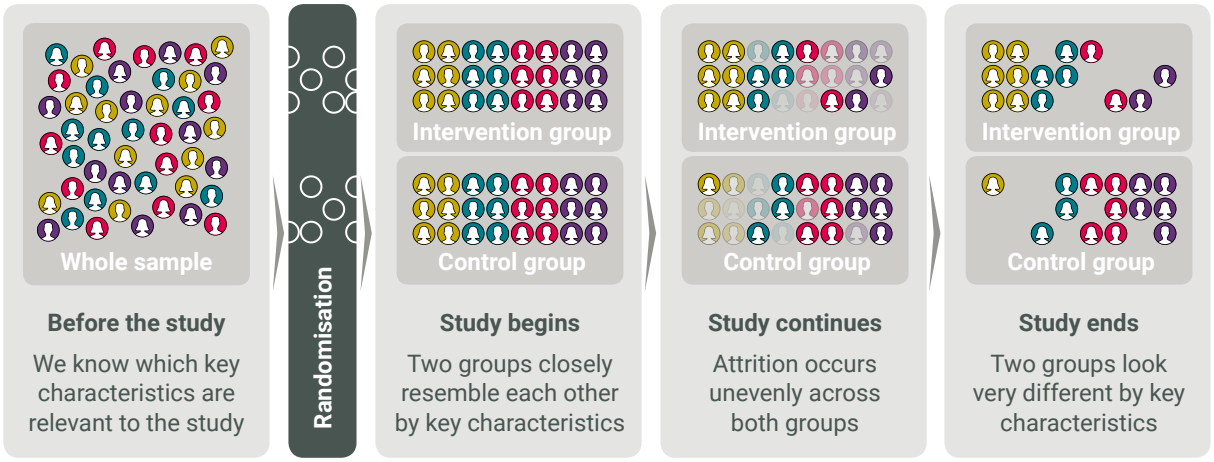
- clear communication of the benefits of taking part in the research
- case management, such as assigning research team members to follow-up with participants
- maintaining detailed contact information, to maximise the likelihood of being able to track down all participants
- compensation, such as cash, vouchers or equivalent gifts

- reminding participants, by letter, phone, email or other forms of electronic messaging
- ensuring data collection is proportional and not overly burdensome.

For more information on preventing attrition, see the list of useful resources at the end of this guide.

FIGURE 3

How attrition can introduce bias



Our hypothetical study is to evaluate the impact of an intervention designed to increase students' academic attainment. Imagine that students shown by the gold circles typically score higher than other students, and the people shown by purple circles score lower. Uneven attrition means that our final intervention group is mostly golds and greens, while our control group is mostly pinks and purples. Having more high-achieving students in the intervention group implies we would probably find a higher average test score for the intervention group than for the control group – even if the intervention was not actually effective at changing student performance. The observed effect of the intervention is biased: some of the differences in outcomes stem from differences between the intervention and control groups due to attrition.

Source: Adapted from What Works Clearinghouse 'WWC Standards Brief: Attrition'

Participant drop-out can rarely be prevented entirely. Therefore, once there has been some level of attrition in the study, researchers should examine their sample and conduct analyses on the extent to which it may have introduced bias. It is important that researchers report this information and the output of these analyses, so that study results can be interpreted in this light.

Researchers should report on the following aspects of study attrition:

- **Attrition rates (the extent of attrition):** There are two kinds of attrition that are important to consider and report: the **overall attrition rate** describes the number of participants who have left the sample overall, while the **differential attrition rate** refers to the difference in attrition between the intervention and control groups. Both are important, as the greater the extent of either, the greater the bias likely to be introduced. Researchers should report these figures, or report the underlying participant flow through the study – that is, how many participants are involved in each stage of the study, how many are retained and how many drop out – such that attrition rates can be calculated.

For more information on reporting attrition, see the list of useful resources at the end of this guide.

- **Attrition type (the nature of attrition):** Researchers should review and report:
  - *Differences between study drop-outs and completers:* To examine whether the sample has become unrepresentative, researchers should conduct analyses (statistical hypotheses tests, such as t-test or chi-square tests) comparing the baseline characteristics (such as demographic characteristics or baseline measurements of

the outcome variables) of the retained sample and those who have dropped out. This analysis of overall attrition type will tell us if a certain type of participant was more likely to leave the study, how the sample has changed due to attrition, and who the results are representative of and generalisable to.

- *Whether attrition undermined the equivalence of the study groups:* To examine whether differential attrition has potentially introduced bias, researchers should conduct analyses comparing the baseline characteristics of the retained intervention group and the retained comparison group. If any differences between the groups are identified, this should be reported, and outcome analysis of programme impact should statistically control for these differences. Most statistical software packages which might be used to analyse outcome data (such as Stata, SPSS or R) have an option to statistically control for other variables.

In addition, there are methods available for estimating missing data within an impact evaluation. This means it may be possible to reduce the impact of attrition by using statistical methods to predict plausible values for missing data points, so that individuals with incomplete data can still be included in the analysis. However, it is worth noting that these techniques are not 'magic bullets' for dealing with the problems of overall and differential attrition, and rely on assumptions which are often impossible to verify.

For more information on estimating missing data, see the list of useful resources at the end of this guide.

What are the implications for EIF's evidence rating?

- For a comparison group study to be considered as providing preliminary evidence (an evidence rating of 2), the overall attrition rate must be less than 65% – that is, no more than 65% of the initially assigned sample can leave and be excluded from final outcome analyses. For a pre-post study, the overall attrition rate must be less than 40%.
- For a study to be considered robust (rating of 3 or higher), researchers must report clearly on the extent of both overall and differential attrition. If overall attrition is greater than 10%, then researchers must report on the differences between study drop-outs and completers, and perform analyses demonstrating that study attrition did not undermine the equivalence of the intervention and comparison groups.

Case study 2: High drop-out rate

Issue: EIF reviewed an evaluation of a therapy-based programme for children. The researchers clearly reported overall and differential attrition at each key point. At post-intervention, overall attrition was over 20%. However, the researchers did not report analyses investigating the extent to which attrition had undermined the equivalence of the intervention and control groups (whether certain types of participant were more likely to drop out of one group than the other). Therefore, it is unclear to us whether the results reflect differences between the groups introduced by attrition or the true impact of the programme.

What could have been done to address the issue: The study would have benefited from a clear reporting of attrition analyses. The researchers could have statistically controlled for any differences those analyses may have identified, provided that those differences were not too extreme.

EIF evidence rating: Because there was the potential for uncontrolled biases, this evaluation could not be considered to provide fully robust (level 3) evidence, and so this programme received an EIF evidence rating of 2 (indicating preliminary evidence).

Pitfall 3: Excluding participants from the analysis

Why does excluding participants from the analysis undermine confidence in an evaluation's findings?

In an RCT, random assignment means that, on average, the intervention and comparison groups are similar in all respects, except that one receives the intervention and the other does not. However, if participants are excluded from the analysis because they have not received a 'sufficient' amount of the programme then equivalence may be undermined. They may be excluded either by not collecting their data, or by not including them in analyses (which is known as **per-protocol analysis**).

It may seem surprising in a study estimating the effects of a programme to include in the analysis participants who have not received any or much of the programme. However, it's important that the intervention and control groups are as similar as possible, and excluding participants from analysis can undermine this. For example, participants attending more sessions of an obesity reduction programme may be more motivated to lose weight than those who attend fewer sessions. Removing participants who have received a lower dose of the programme – that is, who attended fewer sessions – may result in an intervention group that includes only the most motivated participants. This could make the programme look effective by comparison with the control group, but this, in part, could be due to differences in motivation between the groups, rather than the effectiveness of the programme itself.

Conversely, those assigned to a control group should be included in analyses and analysed *as* control group members regardless of whether they participated in or benefited in some way from the programme received by the intervention group. Again, this may seem surprising, but it is essential to ensure that the control and intervention groups are as similar as possible.

How can this issue be addressed?

Researchers should conduct intent-to-treat (ITT) analysis. In an ITT design, the evaluator attempts to collect *all* participant outcome data *regardless* of how much of the programme has been received by individual study participants, and to include each participant in analyses as part of the groups they were originally assigned to.

Intent-to-treat analysis requires:

- **Comprehensive data collection:** Evaluators must attempt, as far as possible, to collect data on all assigned participants. ITT relies on participants who attended few programme sessions (or dropped out of the programme entirely) still providing data.

This means that evaluators need to be able to collect data from participants even if they do not physically attend the programme, either by visiting them at home or collecting data remotely.[3]

- **Correct analysis:** Evaluators should always analyse the data they collect using ITT principles: analyse each participant according to the group to which they were originally assigned, regardless of their level of participation. Evaluators may *also* conduct and report per-protocol analyses, to come to an estimate of the effect for the sub-group who received more of or participated more fully in the programme – but this should be viewed as additional and supplementary. ITT analyses should always be reported, as this provides an unbiased estimate of impact on the target population.

For more on complier average causal effect (CACE) analysis, see the list of useful resources at the end of this guide.

What are the implications for EIF's evidence rating?

- For a study to be considered robust (rating of 3 or higher), researchers must use intent-to-treat analysis.

Case study 3: Excluding participants from the analysis

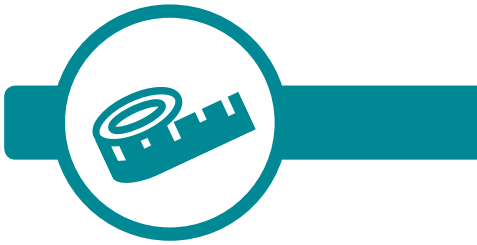
Issue: EIF reviewed an evaluation of a programme designed to improve social and emotional skills in children. We discovered that the researchers had collected data for children in the intervention group at post-test only if they had received at least half of the available programme sessions. This per-protocol design is likely to have introduced biases into the findings.

What could have been done to address the issue: The study would have benefited from attempting to collect data on all participants, and including these participants in the analysis, regardless of the amount of the programme each participant had received.

EIF evidence rating: Because the per-protocol design may have introduced biases, this evaluation cannot be considered to provide fully robust (level 3) evidence, and so this programme received EIF evidence rating of 2 (indicating preliminary evidence).

[3] See pitfall 2, on drop-out rate, for more on methods for improving participation in measurement and preventing attrition.

Pitfall 4: Using inappropriate measures



Why does using inappropriate measures undermine confidence in an evaluation's findings?

Measurement of outcomes is a crucial aspect of impact evaluation. It is important that programme developers are able to clearly articulate the outcomes they are seeking to achieve and that evaluators have a clear sense of the best way to measure changes in those outcomes.

Logic models can be a helpful tool in providing this clarity. A logic model is a statement of what a programme or service consists of (in terms of inputs, activities and outputs) and what a programme intends to achieve (short- and long-term outcomes) alongside a statement of theory – how and why a programme is expected to achieve its desired effects. Logic models are commonly expressed in diagram form, showing the connections between the various components of the programme.

For more information on logic models, see the list of useful resources at the end of this guide.

Appropriate measures should demonstrate both validity and reliability:

- **Measurement validity** refers to the extent to which a measure is successful in describing or quantifying what it is intended to measure. There are a number of different types of validity. One example is convergent validity, which describes the extent to which scores on a measure correlate with scores on another measure which one would expect to be related. For example, a self-report measure of empathy may have strong convergent validity if it correlates strongly with observations of actual charitable behaviour (Souza et al., 2017). We can conclude very little from the results of a study which uses a non-validated measure as we cannot have confidence it is properly measuring what is intended.
- **Measurement reliability** refers to the consistency of a measure. If a measure is reliable, it will elicit the same response under the same or similar contexts. For example, if you assess a young person's empathy using a self-report measure, and then repeat that measure a couple of days later, you would expect – all else being equal – for the measurement to suggest similar levels of empathy. If there are dramatic differences, and there is no reason to expect that the young person's levels of empathy have changed, then this may be caused by random errors (Souza et al., 2017). This example refers to **test-retest reliability**, which describes the extent to which the outcomes of an assessment are stable over time. There are a number of other different types of reliability, including **interrater reliability**, which is the extent to which different observers making assessments come to similar conclusions. When measurement is unreliable, it will be unclear whether a given result reflects a true effect or is the product of measurement error.

Measures should be identified and used correctly:

- **Validation** is the process by which the validity and reliability of measures are established. Crucially, this must be conducted independently of the evaluation in which a measure is being used – that is, validation must be achieved through a separate sample in a study specifically designed to establish the validity and reliability of the measure.
- Furthermore, validated measures should also be used in their **entirety**. Modifying, adding or removing items from a previous validated measure means there is no guarantee that the modified measure will be valid and reliable (for example, adding and removing questions from a validated questionnaire about parenting styles designed to be administered to parents of young children; or modifying the wording of existing questions). In some cases, it may be possible to use parts or subscales of a validated measure in a manner that maintains validity and reliability, depending on the measure and how it was initially validated. However, this must be assessed on a case-by-case basis and will require expert advice.

For more information on validity and reliability, see the list of useful resources at the end of this guide.

Measures should also be an **appropriate measure of the programme's anticipated outcomes in the target population**. This means:

- Measures should have been validated for use with a population which resembles the study sample, particularly with respect to age, demographics and level of need. Even well-validated measures are not necessarily appropriate in *any* given case with *any* given population. For example, a measure which has been validated for use with normally developing teenage boys is likely to be entirely inappropriate for assessing preschool-aged girls with serious developmental problems – we cannot expect this measure to be appropriate for this group unless the measure has been revalidated for this particular population.
- The measure should clearly assess outcomes which correspond to the intended effects. For example, if a programme that intends to improve child mental health uses only parenting measures in an evaluation, then it cannot be concluded that it has direct benefits for the child. While it is possible that an improvement in parenting outcomes will have knock-on effects on a child's outcomes, we cannot be sure until this has been tested directly using another, more appropriate measure or measures.

The use of unreliable measures which are not **valid, reliable, appropriate** for the population or **used in their entirety** will make it unclear whether any apparent improvement in outcomes reflects a true effect or is a product of measurement error.

How can this issue be addressed?

Researchers should use validated measures that will directly assess the impact of the programme on the programme's intended outcomes. These measures should have been validated independently of the study.

A number of resources are available online to guide researchers in their measurement decisions (see table 1 below).

TABLE 1

Sources of information on validated measures

Source	What does it contain	Link
California Evidence-Based Clearinghouse: List of reviewed measures	A list of measures related to child welfare (mental health needs and family attributes) containing assessments of how well validated they are.	http://www.cebc4cw.org/assessment-tools/measurement-tools-highlighted-on-the-cebc/
Early Intervention Foundation (EIF): *Foundations for Life* report	Tables of validated measures used in evaluations of early years programmes: related to children's attachment (page 65), children's behavior (p93), children's cognitive development (p131), and family and the home.	http://www.eif.org.uk/publication/foundations-for-life-what-works-to-support-parent-child-interaction-in-the-early-years/
Early Intervention Foundation (EIF): *Commissioner Guide: Reducing the impact of interparental conflict on children*	List of validated measures used in evaluations of programmes with a component addressing the interparental relationship that have been assessed by EIF.	https://www.eif.org.uk/files/pdf/cg-rpc-3-3-examples-validated-measures.pdf
Education Endowment Foundation (EEF): Early Years Measures database	A database of measures of language, literacy, numeracy, and social and emotional skills for children aged 0–6, containing assessments of how well validated they are.	https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/evaluator-resources/early-years-measure-database/early-years-measures-database/
EEF: Spectrum database	A database of measures of non-academic and essential skills (social and emotional skills), containing assessments of how well validated they are.	https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/measuring-essential-skills/spectrum-database/
ETS Test Collection	A searchable database of measures and information on their development, reliability and validity.	https://www.ets.org/test_link/about
Deighton et al review (2014)	A review of self-report mental health and wellbeing measures for children and adolescents.	https://capmh.biomedcentral.com/articles/10.1186/1753-2000-8-14
Denham & Hamre review (2010)	A review of social and emotional learning measures for children from preschool through to elementary school age.	https://pdfs.semanticscholar.org/e55c/3929969c5b1ffb35966ead41a08b1e040aaf.pdf

What are the implications for EIF's evidence rating?

- For a study to be considered as providing at least preliminary evidence (an evidence rating of 2 or higher), at least one significant child outcome must be identified on a measure which is valid, reliable and appropriate for the anticipated outcomes and population. Administrative data (rates of arrest, exam results etc) do not need to have established validity and reliability, but it is important that the data are described in detail and their sources are indicated. In very specific circumstances we hold to less strict standards of validation measures which assess phenomena which are directly observable or recollectable, such as self-reports of substance use.

Case study 4: Using inappropriate measures

We have often come across problematic measurement in trials that have otherwise been well designed and conducted.

Issue 1 – Non-validated measures: In an RCT designed to test whether a programme improved social and emotional skills in young people, researchers correctly identified well-validated measures of the key outcomes of interest. However, individual items were taken from the measures and then recombined and altered to create a new measure. This undermined our confidence in the findings and conclusions of the study, as the validity and reliability of this new measure had not been independently established.

> **What could have been done to address the issue:** The study would have benefited from the use of each of the measures in their entirety, rather than in altered forms. To meet their objective of keeping questionnaires short, the researchers could have used fewer measures, or more concise forms of validated measures.

Issue 2 – Validated measures used inappropriately: In an evaluation designed to test whether a programme improved behaviour in younger children, researchers used a well-validated parent report measure that assesses children's behavioural problems to measure impact. However, this measure was designed to assess the behaviour of *older* children and had been validated with respect to samples of older children. Therefore, a measure which had not been validated for the study sample had been used, and confidence in the findings was undermined.

> **What could have been done to address the issue:** The researchers could have reviewed a range of measures and selected one or more which had been validated independently with a sample of younger children that resembled more closely the target population of the evaluated programme and the study sample.

EIF evidence rating: Because the validity and reliability of the measures had not been established (or established with respect to a population which resembled the study sample), these evaluations could not be considered to provide preliminary (level 2) evidence, and so these programmes received an EIF evidence rating of NL2 (indicating 'not level 2').

Pitfall 5:
Small sample size

Why does a small sample size undermine confidence in an evaluation's findings?

Sample size refers to the number of participants involved in an evaluation study at any given point. What is critical is the size of the final sample upon which the outcome analysis is conducted, and whether it is sufficiently large for the study to produce robust results.

There are several problems intrinsic to small sample sizes which limit the strength and reliability of the conclusions that can be drawn from a study. Note that the concerns around study drop-out and attrition (namely, bias due to missing data), are distinct from the concerns around small sample size (though attrition will exacerbate small-sample issues by making the sample smaller). Issues around small sample sizes include:

- **False negative results (or type II errors):** Small sample studies are more likely to conclude that a result is not statistically significant when, in fact, the programme did have an effect (a 'false negative result'). 'Statistical power' refers to the probability that a study will not make a false negative error, and an 'underpowered' study is one where false negatives are likely: that is, where there is an insufficiently large sample size for statistical analyses to detect a statistically significant programme effect in a case where the programme *did* have a true effect. For example, if a study was set up and conducted such that it had 20% statistical power, it would be expected to correctly identify only 20 out of every 100 true programme effects (Button et al, 2013). All else being equal, a smaller sample size will reduce the likelihood of finding a statistically significant effect.
- **False discoveries:** Insufficiently large samples also reduce the likelihood that the statistically significant effects which *are* identified actually reflect a true effect. 'Positive predictive value' is the probability that an apparently positive, statistically significant finding reflects a true effect (Button et al, 2013). In underpowered studies, this value is low, which means that the false discovery rate is high: in other words, low power increases the chance that any statistically significant finding is false. Therefore, even if a study *does* identify statistically significant results, the reader's confidence in these results may be undermined.
- **Chance bias in estimating programme effect:** In a sufficiently large trial, randomisation should ensure that the study groups are equivalent. In small trials, however, differences between the intervention and control groups due to chance can and do occur (Torgerson & Torgerson, 2003). By increasing the likelihood that the intervention and control groups are not equivalent, this can bias the results of outcome analyses (sometimes referred to as 'chance bias') and result in type I errors: finding statistically significant effects when, in fact, the result was due to chance (Torgerson & Torgerson, 2003).

These issues mean that small sample studies may be uninformative and inconclusive. If they show no difference between the intervention and control, we do not know if this is because there was no effect, or because the sample was not large enough to detect any effects. Equally, any apparently positive results may be due to chance or differences between the study groups rather than genuine impact.

How can this issue be addressed?

All else being equal, larger sample sizes reduce the severity of the problems described above. Therefore, researchers should attempt to recruit and retain appropriately large sample sizes. A sufficiently large sample size depends on a range of considerations, including the design of the study and the size of the effect the researcher is attempting to identify (the 'minimal detectable effect size', typically based on an estimate of what the programme might be expected to achieve). This can be used to inform power calculations,[4] which tell the researcher how big the sample needs to be to adequately answer the research question. It is essential that researchers are realistic about the likely impact of the programme being evaluated and potential attrition, and use power calculations to identify the appropriate sample size.

For more information on conducting power calculations, see the list of useful resources at the end of this guide.

Pilot studies (small-scale versions of the planned study) are a useful mechanism for estimating recruitment and attrition rates: how many individuals (or schools or clinics etc) need to be approached so that a certain number of participants will be retained in the trial. For example, if we need 100 participants for analysis, and estimate 20% attrition, we need to increase the sample size by 25%, to 125.[5]

It is also important that researchers report the size of their sample at each point in the study (at randomisation/assignment, at baseline, post-test, at follow-up etc), and their power calculations, to provide assurance on the rigour of the approach.

For more information on reporting sample sizes, see the list of useful resources at the end of this guide.

What are the implications for EIF's evidence rating?

- For a pre-post or comparison group study to be considered as providing at least preliminary evidence (evidence rating of 2 or higher), the final sample after attrition must contain at least 20 participants. For comparison group studies, this means at least 20 in the intervention group.
- For a comparison study to be considered robust (rating of 3 or higher), the final sample must contain at least 20 participants in the intervention group *and* the control group.
- Note that while this is our minimum requirement, we would advise evaluators to recruit sample sizes larger than this, because a study of this size would only be able to reliably detect very large effects. For example, a study attempting to identify a medium reduction (d=.50) in children's conduct problems would have only approximately 35% power to identify this effect, meaning that a study of this size would correctly detect the desired effect only 35 out of 100 times. To identify a small reduction (d=.20), this study would only have approximately 10% power to identify the effect, meaning it would detect the desired effect only 10 out of 100 times. And even if a positive effect is detected by a study of this size, the reader would have reasons to doubt the legitimacy of this result, for the reasons described above (higher false discovery rate, higher probability of chance bias etc).

---

[4] Online power calculators are available. Alternatively, dedicated pieces of software can conduct power analyses, such as G*Power or Optimal Design. In addition, power calculations can be conducted using standard statistical packages such as Stata and R.

[5] See pitfall 2, on drop-out rate, for more on methods for improving participation in measurement and preventing attrition.

Case study 5: Small sample size

Issue: EIF reviewed an evaluation of a programme for adolescents. The study involved randomly assigning 16 adolescents to an intervention group and a control group (eight in each). No power calculations were presented, and the researchers themselves noted that their small sample size meant that confidence in their findings would be reduced.

What could have been done to address issue: The researchers should have recruited a sufficiently large sample at the outset of the study, such that a sufficiently large sample would be maintained even if some participants had dropped out. Ideally, the researchers would have provided power calculations to verify that the sample was sufficiently large.

EIF evidence review: Because the sample size did not meet our minimum level, this evaluation could not be considered to provide preliminary (level 2) evidence.

Pitfall 6: Lack of long-term follow-up

Why does a lack of long-term follow-up undermine confidence in an evaluation's findings?

According to EIF's criteria, a study assesses long-term follow-up if it measures outcomes 12 months or more post-intervention (after the programme has stopped being delivered). Studies which solely measure outcomes immediately after the programme has been delivered can only inform us about a programme's short-term impact.

While this is important, it is only part of the picture, for several reasons.

- It is possible for a programme's short-term impact to reduce or 'fade out' over time. Therefore, studies which identify significant positive findings at 12 months (or more) post-intervention are important, as they can help us to identify programmes which produce lasting effects (Bailey et al, 2017).
- There may be a delay between a programme being delivered and its effects materialising. For example, it may take time for parents to translate skills learnt in a programme to their everyday lives, and longer still for this to translate into outcomes for their children. Such benefits may not be immediately apparent, and a study focusing only on short-term outcomes could fail to capture the full impact of the programme.
- Short-term outcomes are often largely considered important insofar as they are steps towards a more highly prized long-term outcome, which is the ultimate goal of the programme. Studies which do not attempt to measure long-term outcomes cannot fully tell us the extent to which a programme has achieved its most important goals. This is particularly important in the field of early intervention programmes, where the objective is often to reduce the long-term social and economic costs of children and young people experiencing significant difficulties later in life.

How can this issue be addressed?

Although it is costly and complex, researchers should incorporate long-term measurement into their study design and plan data collection to coincide with the points where long-term impacts are expected to occur.

In addition to collecting data at these time points, the researchers should endeavour to maintain the rigour of their short-term analyses at long-term follow-up points. Particular issues to consider include:

- **Maintaining the comparison group:** It is common for studies to drop their comparison group at long-term follow-up points, often because those in the control group were on a waiting list and have since started to receive the programme. Instead, these studies sometimes assess how outcomes have changed *within* the intervention group from post-intervention to long-term follow-up. If there are no significant differences between the two,

they conclude that the intervention effect has been 'maintained'. However, these estimates of long-term impact suffer from the same biases as one-group pre-post studies,[6] meaning that the extent to which the intervention effect is truly maintained is not robustly demonstrated. Evaluators should avoid using waiting-list designs if they are interested in long-term follow-up.

- **High attrition:** Attrition increases over time, and it can be particularly challenging to collect data on study participants many years after their involvement. Subsequently, attrition poses a serious threat to the quality of long-term follow-up. As described in the chapter on attrition, it is important that researchers take steps to minimise drop-out and improve participation in measurement, and conduct analyses to alleviate concerns about the effect of attrition within a study, which are critical for long-term follow-up studies. In addition, researchers could make use of existing comprehensive administrative data where this is available (for example, schools are mandated by government to collect data for the National Pupil Database, meaning that there is a rich and comprehensive set of data on pupils which researchers could use instead of collecting their own data).

What are the implications for EIF's evidence rating?

- A programme with two or more robust studies (rating of 3) may receive an overall evidence rating of 4 provided one of these studies provides robust evidence of long-term positive findings.

Case study 6: Lack of long-term follow-up

The EIF Guidebook includes relatively few programmes that have robustly demonstrated long-term impact and therefore achieved an evidence rating of 4.

Issue 1 – No control group at follow-up: An RCT designed to test whether a programme improved antisocial behaviour in children assessed all study participants at pre- and post-intervention. At the 12-month follow-up, however, only children in the intervention group were assessed.

> **What could have been done to address the issue:** The study would have benefited from continuing to measure both the comparison group and intervention group at follow-up (as far as possible), to provide a robust counterfactual upon which to base any long-term findings.

Issue 2 – High attrition at follow-up: EIF reviewed an RCT designed to test whether a programme improved behavioural and emotional problems in children. At post-intervention, approximately 20% of participants had dropped out from the study. At the two-year follow-up, far fewer participants completed the assessment, and the overall attrition rate grew to just over 65%, which exceeds our overall attrition standard.

> **What could have been done to address the issue:** As far as possible, the researchers should have taken steps to maintain contact with study participants in the years following the trial's initial stages, to reduce attrition at the long-term follow-up point.

EIF evidence rating: Because the long-term findings in these studies were not as robust as their short-term findings, these programmes could not receive an EIF evidence rating of 4.

[6] See pitfall 1, on robust comparison groups.

References

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 10(1), 7-39.

Brueton, V., Tierney, J., Stenning, S., Nazareth, I., Meredith, S., Harding, S., & Rait, G. (2011). Strategies to reduce attrition in randomised trials. *Trials*, 12(1), A128.

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.

Deeks, J. J., Dinnes, J., D'amico, R., Sowden, A. J., Sakarovitch, C., Song, F., ... & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health technology assessment*, 7(27), iii-x.

Greenberg, M. T. & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness,* 1, 40–67.

HM Treasury (2011). *The Magenta Book: Guidance for evaluation*.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2009). *Handbook on impact evaluation: quantitative methods and practices*. World Bank Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth.

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.

Torgerson, D. J., & Torgerson, C. J. (2003). Avoiding bias in randomised controlled trials in educational research. *The British Journal of Educational Studies,* 51(1), 36-45.

Useful resources

Introduction: EIF Guidebook, programme assessments & evidence ratings

EIF evidence reviews

Foundations for Life: What works to support parent child interaction in the early years? (2016)	http://www.eif.org.uk/publication/foundations-for-life-what-works-to-support-parent-child-interaction-in-the-early-years/
What works to enhance inter-parental relationships and improve outcomes for children? (2016)	http://www.eif.org.uk/publication/what-works-to-enhance-inter-parental-relationships-and-improve-outcomes-for-children-3/
Social and emotional learning: Skills for life and work (2015)	http://www.eif.org.uk/publication/social-and-emotional-learning-skills-for-life-and-work/

Other clearinghouses

Blueprints for Healthy Youth Development	http://www.blueprintsprograms.com/criteria
California Evidence-Based Clearinghouse for Child Welfare	http://www.cebc4cw.org/ratings/scientific-rating-scale/
Campbell Review Criteria	http://www.campbellcollaboration.org/lib/project/328/
Centre for Analysis of Youth Transitions (IFS)	http://www.ifs.org.uk/caytpubs/cayt_impact_school_health.pdf
Cochrane Review Criteria	http://handbook.cochrane.org/
CrimeSolutions.gov (Office of Justice Programs)	https://www.crimesolutions.gov/pdfs/ratinginstrument_part2.pdf
National Registry of Evidence-based Programmes and Practices (SAMHSA)	http://nrepp.samhsa.gov/04e_reviews_program.aspx
Project Oracle	http://project-oracle.com/support/validation/
Promising Practices Network (RAND)	http://www.promisingpractices.net/criteria.asp
Teen Pregnancy Prevention Program (Office of Adolescent Health)	http://tppevidencereview.aspe.hhs.gov/ReviewProtocol.aspx
Top Tier Evidence Initiative (Coalition for Evidence-Based Policy)	http://toptierevidence.org/solicitationreview-process
What Works Clearinghouse (Institute of Education Sciences)	http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Reporting standards

Consort guidelines on reporting trials	http://www.consort-statement.org/consort-2010
Spirit guidelines on reporting trial protocols	http://www.spirit-statement.org/

Pitfall 1: No robust comparison group

RCTs and QEDs

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

HM Treasury (2011). *The Magenta Book: Guidance for evaluation*.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2009). *Handbook on impact evaluation: quantitative methods and practices*. World Bank Publications.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth.

Torgerson, D. (2008). *Designing randomised trials in health, education and the social sciences: an introduction*. Springer.

Pitfall 2: High drop-out rate

Preventing attrition

Brueton, V., Tierney, J., Stenning, S., Nazareth, I., Meredith, S., Harding, S., & Rait, G. (2011). Strategies to reduce attrition in randomised trials. *Trials*, 12(1), A128.

In a school context: Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., ... & Datalab, E. (2017). *Evaluation of Complex Whole-School Interventions: Methodological and Practical Considerations*.

Reporting attrition

Consort guidelines on reporting trials: http://www.consort-statement.org/consort-2010

Dumville, J. C., Torgerson, D. J., & Hewitt, C. E. (2006). Research methods: reporting attrition in randomised controlled trials. *BMJ: British Medical Journal*, 332(7547), 969.

Estimating missing data

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med*, 86(3), 343-358.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to Do when Data Are Missing in Group Randomized Controlled Trials*. NCEE 2009-0049. National Center for Education Evaluation and Regional Assistance.

Pitfall 3: Excluding participants from the analysis

Conducting CACE analysis

Dunn G, Bentall R. (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatment). *Statistics in Medicine*. 26, 4719-4745.

Hewitt C, Torgerson D, Miles J. (2006). Is there another way to take account of noncompliance in randomized trials? *Canadian Medical Association Journal*. 175(4), 347-348.

Pitfall 4: Using inappropriate measures

Logic models

Axford, N. (2016). From miracles to logic models [blog post]. http://fnp.rubbaglove.co.uk/blogs/from-miracles-to-logic-models/

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). Chapter 2. In *Using Randomised Controlled Trials in Education.* SAGE.

Measurement validity and reliability

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.

Pitfall 5: Small sample size

Conducting power analysis

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Djimeu, E.W., & Houndolo, D-G. (2016). *Power calculation for causal inference in social science: sample size and minimum detectable effect determination*, 3ie impact evaluation manual (Vol. 26). 3ie Working Paper. http://www.3ieimpact.org/media/filer_public/2016/07/08/wp26-power-calculation.pdf

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149-1160.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S.W. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software, Version 3.0. William T. Grant Foundation. http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od

Reporting sample sizes

Consort guidance: http://www.consort-statement.org/consort-statement/flow-diagram