April 2026 | **Handbook**

GUIDEBOOK HANDBOOK V 2.0

Procedures and standards for assessment of evidence for interventions

Foundations
What Works Centre for Children & Families

foundations.org.uk

The Foundations Guidebook is a free online resource that contains evidence on what works to support children, young people, and families.

Finding out what works best for children and families is not easy. There is a lot of evidence available, but it can be hard to know how reliable it is. This means it can be difficult to decide on the right kind of support for children and families.

The Guidebook aims to change this. It is free and easy to access. It is designed to help local leaders, commissioners, and practitioners, and researchers and policymakers, to use evidence when they make decisions about how best to support children and families.

Visit the Foundations Guidebook: foundations.org.uk/toolkit/guidebook
Visit the Foundations website: foundations.org.uk

This Handbook *(Guidebook Handbook v 2.0)* reflects the revised processes and standards **applicable from April 2026**.

Please see *Guidebook Handbook v 1.0* (previously the *Technical Guide*) for information on the previous standards and processes by which **interventions published before April 2026** were assessed: https://foundations.org.uk/wp-content/uploads/2026/04/guidebook-handbook-v1.0-procedures-and-standards.pdf

About Foundations, the national What Works Centre for Children & Families

Foundations, the national What Works Centre for Children & Families, believes all children should have the foundational relationships they need to thrive in life. By researching and evaluating the effectiveness of family support services and interventions, we're generating the actionable evidence needed to improve them, so more vulnerable children can live safely and happily at home with the foundations they need to reach their full potential.

Acknowledgments

We thank the members of Foundations' Evidence and Evaluation Advisory Panel who contributed to the development of the revised evidence standards: Nick Axford, Rachel Churchill, Sajid Humayun, and Stephen Morris.

We are also grateful to other expert reviewers of the standards – Jacqueline Barnes, Rhiannon Evans, Kerry Dwan and Amy Hall – and the systematic review team who piloted the standards – Lisa Jones, Menna Abdelgawad, and Emily Smith.

CONTENTS

ABOUT THE GUIDEBOOK

What is the Foundations Guidebook?

The Foundations Guidebook provides information on evidence to support decision-making across the spectrum of early intervention, services for children and families, and children's social care.

It contains information on existing research on over 130 interventions with at least preliminary evidence of improving children and families' outcomes.

The Foundations Guidebook is part of the Foundations Toolkit, which also contains the Practice Guides – these are commissioned by the Department for Education and provide evidence-based recommendations for those commissioning and delivering child and family support at the local level. The Practice Guides and Guidebook offer a complementary set of tools, allowing users to access guidance on evidence-based practices, while at the same time identifying interventions which could support the implementation of these practices.

The Foundations Guidebook was launched in May 2025, building on the Guidebook of the Early Intervention Foundation, with increased functionality and enhanced information on interventions.

About this Handbook (v 2.0)

This Handbook (*Guidebook Handbook v 2.0* (2026)) supersedes the Technical Guide published during the launch of Foundations' Guidebook in May 2025.[1] The *Technical Guide* has been renamed *Guidebook Handbook v 1.0* for consistency with this revised Handbook v 2.0.[2] This revised v 2.0 of the Handbook includes:

- A more thorough step-by-step description of Foundations' process for assessing the evidence for an intervention – this will be useful for intervention developers or providers submitting their intervention for assessment for inclusion on the Guidebook
- A description of the revised evidence standards for Level 3, which were developed in 2025/2026 – as well as for providers, these will be particularly useful for study authors supporting submissions, and for Guidebook users wanting to understand Foundations' approach to promising evidence in more depth

[1] The processes and standards published in the *Technical Guide* as part of the launch of Foundations' Guidebook in May 2025, were established in 2016 under the Early Intervention Foundation (EIF) and published on the EIF Guidebook (website).

[2] *Guidebook Handbook v 1.0* is has information on the previous standards and processes by which interventions published before April 2026 were assessed. *Guidebook Handbook v 2.0* (2026) superseded v 1.0, reflecting revised processes and standards applicable from April 2026.

- Information on features introduced in intervention assessments from 2026: in particular, enhanced presentation of evidence on how an intervention may contribute to reducing racial disparities; information about an updated cost model to underpin cost ratings; and a streamlined process for partners.

For interventions added to the Guidebook prior to this Handbook, refer to the *Guidebook Handbook v 1.0*[3] (previously the *Technical Guide* – renamed for consistency with the revised *Guidebook Handbook v 2.0*) for the evidence standards and processes used. The new standards and other features are not implemented retrospectively to interventions already listed on the Guidebook.

The updated Guidebook Evidence Standards

Historically, Foundations used its own evidence standards to appraise studies to determine the evidence rating for interventions appearing on the Guidebook. These standards were informed by other frameworks and tools (e.g. Blueprints, Cochrane's Risk of Bias 1 tool, CONSORT guidelines), and were developed in consort with other What Works Centres and with a range of expert advisors. They were formally approved by an evidence panel during the set-up phase of the Early Intervention Foundation (EIF) in 2016 and were previously available on the EIF Guidebook website. For more information about the previous version of the evidence standards and their development, see *Guidebook Handbook v 1.0.*[4]

Recently, Foundations has started awarding grants for systematic reviews to inform the development of Practice Guides, which were commissioned by the Department for Education. These systematic reviews largely use Cochrane's Risk of Bias 2 (RoB2), Risk of Bias in Cluster Randomised Trials (RoB-CRT) or Risk of Bias In Non-randomised Studies – of Interventions (ROBINS-I) tools to appraise included studies. The Practice Guides have an evidence standards framework[5] for translating the findings of the review into strength of evidence ratings for recommendations for practice.

To align more closely with the Practice Guides, and with widespread research methods and best practice, the Guidebook is now adopting a set of updated evidence standards for Level 3, or promising evidence. These standards take the form of Cochrane's tools (RoB2, RoB-CRT, and ROBINS-I) with implementation guidance developed specifically for trials in children's social care and early intervention, together with a small set of Guidebook-specific standards.

The implementation guidance for Cochrane's tools is largely informed by the standards of another toolkit in early intervention and children's social care research, the Title IV-E Prevention Services Clearinghouse, and by the previous Guidebook standards. The aim is to make Cochrane's tools:

---

[3] Available at: https://foundations.org.uk/wp-content/uploads/2026/04/guidebook-handbook-v1.0-procedures-and-standards.pdf

[4] Available at: https://foundations.org.uk/wp-content/uploads/2026/04/guidebook-handbook-v1.0-procedures-and-standards.pdf

[5] See: https://foundations.org.uk/how-to-use-the-practice-guides/

- Easier to understand, particularly for less experienced reviewers, where the Cochrane tool signalling questions and guidance might be complex
- More readily applicable to trials of children's social care and early intervention programmes, which commonly have characteristics not seen in medical and clinical literature, e.g. relatively high attrition rates; use of self-report measures; and lack of blinding of practitioners and participants from assignment condition
- More consistent in application and risk of bias judgements, through clearer thresholds for risk of bias in some domains.

The additional set of Guidebook-specific standards is needed for important risk of bias characteristics not covered in Cochrane's tools. This is because the Guidebook is based on rating individual studies, and there are often few eligible studies in this area of research. For individual studies, additional assurance is required about the quality of the sample in order to have confidence in the effects observed. Again, these are largely based on the standards of the Title IV-E Prevention Services Clearinghouse, and on previous Guidebook standards.

Existing Guidebook intervention entries

Guidebook entries were migrated to the current Foundations Guidebook in 2025, from the Early Intervention Foundation Guidebook, and relaunched in a new format designed to improve the user experience. Interventions listed on the Guidebook before the April 2026 update of Foundations' Guidebook processes and standards described in this revised Handbook, maintain their listing and will not be reassessed against the new standards or have the new entry features, unless they are reassessed as part of a new assessment round where new evidence has become available. In other words, existing Guidebook intervention entries, including evidence and cost ratings, currently remain unchanged.[6]

What is the Foundations Guidebook's approach to race and ethnicity?

Foundations has published a position statement[7] on its approach to working towards a Guidebook which promotes racial equity and inclusion through interventions and their evidence. This position statement also outlines actions which are taken as part of intervention assessment, including the process of creating an Equality, Diversity, Inclusion, and Equity (EDIE) summary for each intervention entry. These are reflected throughout this Handbook, to show how a focus on racial equity is included at each stage of intervention assessment.

---

[6] *Guidebook Handbook v 1.0* has information on the previous standards and processes by which interventions published before April 2026 were assessed. v 1.0 of the Handbook is available at: https://foundations.org.uk/wp-content/uploads/2026/04/guidebook-handbook-v1.0-procedures-and-standards.pdf

[7] Go to: https://foundations.org.uk/about-the-guidebook/edie-and-the-guidebook/

Previously, when the Foundations Guidebook was launched in May 2025, a full list of race and ethnicities in included studies was added to intervention entries for the first time, in the intervention summary and in the full evidence description. While this was a step in the right direction, a much deeper and more contextualised approach was needed. The position statement aims to acknowledge the limitations and historical shortcomings of dominant, Western evidence frameworks and approaches to evaluation, and outlines how evidence for interventions' effectiveness for minoritised groups, and the experiences of the intervention of minoritised groups will be included in the Guidebook. Anti-racism is an organisational strategic priority, and therefore this position statement focuses on promoting racial equity. However, we are also committed to expanding the Guidebook to better reflect the experiences and needs of a broader range of minoritised groups. We will continue to work to address other forms of marginalisation to ensure the Guidebook is inclusive, representative, and equitable for all.

In terms of language, the Guidebook aims to present information about race and ethnicity transparently and sensitively, reflecting how these characteristics are reported in the studies underpinning the Guidebook evidence rating, while aligning with current best practice on reporting and language.

The Guidebook preserves the terms used in studies underpinning evidence ratings wherever possible, whether broad (e.g. 'Black') or specific (e.g. 'African American'), except in cases where terms are now outdated or offensive. In such cases, terms are 'translated' to align with government and institutional style guidance (e.g. the UK Office for National Statistics' race and ethnicity categories, US, Australian, and New Zealand governmental guidelines, and Foundations' own internal style guide).

Terms that indicate nationality (e.g. 'Finnish'), regions (e.g. 'Mediterranean'), or ambiguous cultural and linguistic descriptors are excluded from the intervention summary on the Guidebook, particularly when study context does not clarify whether a term refers to ethnicity or nationality.

In study summaries, all demographic characteristics of study participants are listed (including ethnicity, race, nationality, language, and related descriptors) as recorded in the studies, with percentages if available. The only changes made in this section are replacements of outdated or offensive language. Where studies do not report race or ethnicity, the guidebook will state this clearly (e.g. 'not reported').

Our approach ensures that evidence is presented in a way that is respectful, clear, and consistent, while supporting commissioners and local area leaders to understand who interventions have worked for, and where evidence may be lacking for specific populations. It is important to note, however, that the intervention may have been delivered and evaluated successfully with other populations; only those populations involved in the studies which underpin the Guidebook evidence rating are mentioned. Also, unless specified, the intervention may have been designed to be implemented with a wide variety of different communities.

We recognise that the position statement and these actions are a starting point; our approach to equality, diversity, inclusion, and equity on the Guidebook will continue to evolve as we engage with new evidence, feedback, and best practice.

What kinds of evidence are included on the Guidebook?

Our assessments focus on impact evaluations – studies that help us understand whether an intervention has had a measurable, positive effect on child outcomes. These typically include:

- Randomised controlled trials (RCTs)
- Quasi-experimental designs (QEDs)
- Some well-designed pre–post studies.

We do not include qualitative evidence as part of our formal assessment process, as this type of evidence does not allow us to draw strong conclusions about whether an intervention has caused change. Qualitative evidence is of course useful for other purposes, and we have an organisational commitment to undertake implementation and process evaluations as part of our own funded evaluations, and to synthesise qualitative evidence to inform our Practice Guides. Currently, qualitative evidence is only included in the EDIE summaries on the Guidebook. We also only include evidence from studies which took place in high-income countries and those with a high Human Development Index, because the findings from these are most relevant and applicable to the UK context; and because evaluations from other countries tend to produce inflated effect sizes in comparison, given the different Business As Usual context.

HOW ARE INTERVENTIONS ASSESSED FOR THE GUIDEBOOK?

1. Identification and prioritisation of interventions

Identification

Foundations identifies interventions for appraisal primarily through two separate routes:

- **Interventions identified in systematic reviews underpinning Foundations' Practice Guides.** The Practice Guides are commissioned by the Department for Education and produced by Foundations. They translate the strongest available evidence into actionable recommendations to support local leaders in strengthening family services. Each Practice Guide is based on a systematic review, commissioned by Foundations, and these systematic reviews identify interventions of interest for the Guidebook which are then fed into our Guidebook assessment pipeline.
- **Interventions identified by the Guidebook team.** The Guidebook team select interventions for assessment and consideration for inclusion on the Guidebook through

open calls to intervention providers, themed assessment rounds designed to support national and local priorities, and research by the Foundations' Guidebook team.

Our methods may evolve over time to reflect emerging priorities and feedback from Guidebook users.

Definition of an intervention

The Foundations Guidebook takes a broad approach to the definition of an intervention, to ensure all interventions and approaches relevant to our users are eligible for inclusion. An intervention is **a form of support for children and/or their families which is clearly defined and replicable**. This means that there is **a manual or other documentation which describes the interventions' key components and delivery**. It is not necessary for an intervention to have a fixed set of sessions or length, or start to finish structure, in order to be included on the Guidebook; more flexible approaches are eligible, as long as they are defined and replicable.

Prioritisation and selection

Once potential interventions are identified through the steps outlined above, they are reviewed for prioritisation for evidence assessment for the Guidebook using our preliminary checklist. This is completed primarily through desk-based research and contact with the intervention's developer or provider where necessary. At this stage, each intervention must meet the following criteria:

- **A clearly defined and replicable model:** The intervention should be described in enough detail that another organisation could deliver it as intended. This typically means there is a manual, guide, or equivalent documentation outlining its core components, delivery approach, and any required materials.
- **Sufficient information about key implementation features:** Information should be available on how the intervention works in practice – for example, the content covered, the expected dosage or frequency, the delivery modality (such as group, one-to-one, in-person, or digital), who delivers it, the setting, and which elements are core versus adaptable. This ensures an understanding of how the intervention operates and what is required for effective implementation.
- **A focus on supporting children and/or their families:** The intervention must be designed to benefit children directly or indirectly through work with parents, carers, or families.

Interventions that meet these requirements then move on to the scored section of the preliminary checklist. This scoring process helps us prioritise interventions based on:

- Their alignment with Foundations' strategic aims and priority areas
- Whether they are UK based
- The presence of high-priority drivers (e.g. inclusion in a Practice Guide) or medium-priority drivers (e.g. stakeholder interest or wider sector relevance)

- EDIE-related considerations including a focus on equity and inclusion, studies involving minoritised groups, or cultural adaptations.

Guidebook Equality, Diversity, Inclusion and Equity (EDIE) commitment

Racially minoritised communities are often excluded or underrepresented in trials due to structural and systemic barriers. To help address these inequities, Foundations includes the following EDIE-focused criteria within our scored preliminary checklist:

- Designed by and/or for a racially minoritised group, or demonstrating a strong commitment to equity and inclusion: This includes interventions created in partnership with specific communities, as well as those designed to benefit all families but with clear consideration of how participants from diverse backgrounds will be supported. Targeted interventions may be developed by and/or for a particular group, incorporating tailored content, components, or delivery methods that reflect the group's needs, strengths, and lived experiences.
- Known evidence from studies involving racially minoritised groups: Interventions that have been tested with, or show promising results for, minoritised groups.
- Known variants or cultural adaptations: Interventions that have been adapted to reflect cultural norms, languages, or community contexts.

This prioritisation and selection process is taken as guidance, and there may be other reasonable justifications for choosing an intervention to assess.

Approach to adaptations

Foundations takes an inclusive approach to adaptations, reflecting real-world implementation conditions; different versions of interventions which are not substantially different adaptations are included in a single Guidebook entry.

Adaptations are an appropriate and necessary part of an intervention's development, particularly when considering scalability, or implementation in new contexts. In addition, evaluation of an intervention always highlights areas for improvement, whether through an Implementation and

Guidebook EDIE commitment

Interventions may be adapted to take into account cultural differences or language barriers. Where applicable, intervention pages will include mention of or links to equity-focused adapted models. These adaptations will be assessed separately if they are considered substantial, and together with the original intervention otherwise.

Process Evaluation or more informally. It is important not to artificially separate this body of evidence from across the evaluation pipeline for an intervention.

The Guidebook draws on the Title IV-E Prevention Services Clearinghouse's guidance for determining whether an adaptation is substantial or not substantial, as outlined in figure 1 and table 1 below.

Figure 1. Exhibit 2.4 in *Prevention Services Clearinghouse Handbook Version 2.0* (p. 23) 'Process for Assessing Substantial Adaptations' (link to long descriptive text)

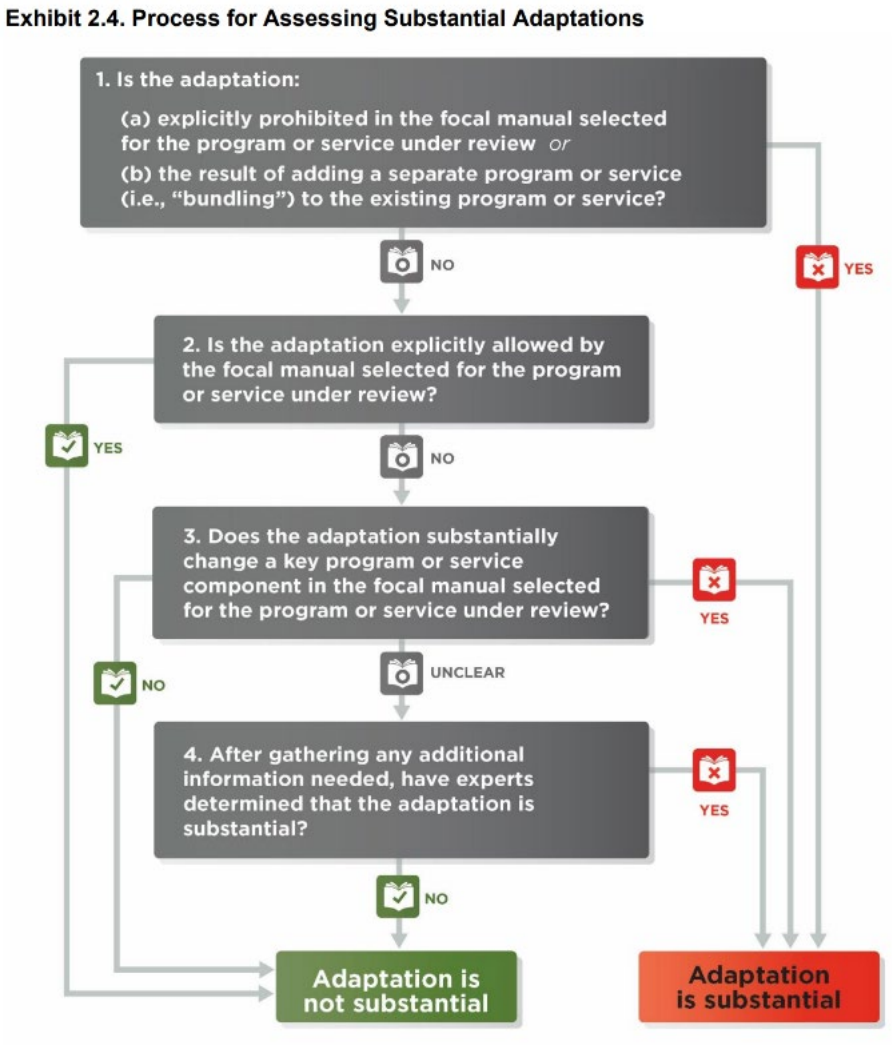


Table 1. Exhibit 2.5 in P*revention Services Clearinghouse Handbook Version 2.0* (pp. 24–25) 'Examples of Adaptations that Are Substantial and Not Substantial'

Programme or Service Key Component Domains	Adaptations that Are Not Substantial	Adaptations that Are Substantial
Dosage – The intended quantity, duration, and frequency of services to be delivered. Can include characteristics of individual sessions (e.g. session frequency and length) and of the overall programme or service (e.g. treatment duration, total sessions, total hours)	- Modestly changing session frequency (e.g. number of sessions per week or month) or session length (e.g. minutes per session), or total number of sessions. - Modestly changing the duration of treatment (i.e. the length of time from the start of the programme or service to the end). - Modestly changing the total amount of treatment contact time (e.g. number of hours of treatment services delivered over the course of the programme or service). - Accelerating or lengthening treatment without changing total contact time (e.g. switching from 12 weekly sessions to six twice-weekly sessions or vice versa but keeping 12 total hours of treatment overall). - Minor differences in session or programme or service dosage when these components are defined flexibly for the programme or service (e.g. delivering a programme in 21 sessions when the number of sessions usually ranges from 15–20 sessions; a treatment duration of 2.5 months when it is typically completed within 2 months).	- Changes in session frequency (e.g. monthly to weekly) or session length (e.g. extending sessions from one hour to four hours) that substantially change total contact time. - Other modifications that substantially change total contact time (e.g. a brief version of a therapy that reduces the total number of hours spent in therapy from 40 hours to 10 hours).

Programme or Service Key Component Domains	Adaptations that Are Not Substantial	Adaptations that Are Substantial
Modality – The delivery setting (e.g. in-home vs office) and format (e.g. group vs individual) of the programme or service	- Delivering intervention in the home compared to office-based delivery. - Changing the delivery setting or format to meet the needs of a specific population (e.g. using talking circles instead of focus groups in tribal populations, having separate discussion groups for mothers and fathers in populations with gender-specific parenting norms). - An intervention typically delivered one-on-one between a client and a therapist is adapted to allow sessions to be completed either in-person or via video-conference without substantial changes to intervention content.	- Changing from individual to group therapy or vice versa. - Changing from synchronous (i.e. live) to asynchronous (i.e. self-paced) programme or service delivery.
Content – The subject matter, themes, activities, examples, skills, methods, and/or goals of a programme or service intervention	- Modifying examples or illustrations. - Providing intervention in a different language. - Updating activities or exercises to increase the relevancy of the programme or service in a particular cultural or contextual setting. - Adding an introductory session or a concluding/graduation session to reinforce existing content.	- Making substantial changes to content (e.g. adding a substance use prevention component to a parenting intervention that previously only had content on child anxiety; making substantive content modifications to address developmental differences).
Providers – Characteristics of the providers who are intended to implement the programme or service (e.g. education, experience, training)	- Delivering the programme or service by slightly different types of providers than described in the manual or original research on the programme or service (e.g. using M.S.W. social workers instead of master's-level licensed counsellors). - Requiring providers to have relevant language skills and cultural knowledge of and experience working with the population being served.	- Delivering the programme or service with substantially different providers than described in the manual (e.g. using untrained paraprofessionals instead of trained nurses to deliver a programme). - Changing from provider-led sessions to self-led sessions.

Note: Adaptations may affect multiple components simultaneously. Each component is reviewed independently. For example, adding two sessions may not be substantial based on dosage alone but may be substantial based on content if new subject matter is added.

Maintenance

New evidence is published all the time, and this new evidence may impact an intervention's evidence rating or provide additional information of interest to Guidebook users.

Foundations aims (funding and capacity permitting) to check in and see if there is any new evidence for each Guidebook intervention regularly, by:

- Asking intervention providers to provide us with newly published evidence
- Literature searching (automated Google Scholar searches and other bibliographic databases)
- Looking at other clearinghouses that have also assessed/described interventions that are published on the Guidebook.

Foundations prioritises which interventions are 'maintained' by considering factors including: new evidence which is likely to change the rating; new evidence from the UK; and new evidence on the intervention's efficacy in reducing racial disparity or benefiting disadvantaged or minoritised groups.

2. Involving intervention developers and providers

After an intervention is selected for Guidebook assessment, the intervention providers are contacted to inform them of the intention to assess, and to provide them with the Terms of Reference for the Foundations Guidebook. Providers are also requested to complete an initial online questionnaire, referred to as the Intervention Submission Form part 1 (ISF1), which gathers information concerning:

- The intervention's theory of change
- The outcomes and target population of the intervention
- The delivery of the intervention
- Impact evaluation evidence underpinning the intervention.

Providers also indicate whether they agree to Foundations' Terms of Reference and whether they have the full support of the intervention developer through filling in the online questionnaire.

For maintenance round assessments, providers are informed and requested to submit any additional new evidence; there is no requirement to reconsent to the Terms of Reference or recomplete the online form.

Non-consenting interventions

Foundations always attempts to seek the cooperation of intervention providers before, during, and after the assessment of their intervention. Intervention providers are a key partner in the assessment process, supporting Foundations in producing accurate descriptions of an intervention's content, delivery, and implementation, as well as helping to identify all relevant evidence.

However, sometimes it is the case that the intervention provider does not engage in the assessment process; typically this occurs where Foundations makes multiple attempts to contact the provider keep them informed of the assessment process, but there is no response or only a partial response. In these cases, the intervention is assessed and included even when the provider has not volunteered their intervention to be assessed and included. These interventions are clearly labelled on the Guidebook.

This approach is motivated by the perceived public benefit, in line with the aims of the Guidebook:

- **Flexibility to assess promising interventions or interventions of interest to users.** Interventions are often identified because they are promising or look to fill an important gap, or because users report that they would like to know more about the intervention. It is important that the Guidebook provides independent assessments of these interventions, even if they do not volunteer.
- **Identification of interventions with limited evidence or evidence of not working**. In line with being a What Works Centre, it is important to identify interventions with a less developed evidence base or with evidence of not working. This means a fuller picture of the evidence can be provided to Guidebook users. Those interventions with weaker evidence can then be further evaluated, and information about interventions likely to be ineffective or to have negative effects can be taken into account in (de)commissioning decisions.

The process for non-consenting providers is otherwise the same as the standard process; non-consenting providers receive the same emails as consenting providers and are given the same opportunities to engage with the process, and to challenge Foundations' evidence rating. The same documentation is created and stored from the assessment process, with the exception that some parts of the Guidebook entry may not be completed where they depend on information from the provider (e.g. the cost rating).

3. Literature search

The purpose of the search stage is not to conduct a systematic review or meta-analysis, but is designed to identify the most robust study or studies available relevant to an intervention. Not every evaluation of an intervention needs to be assessed, provided there is confidence that:

- The strongest available evidence suggesting positive impact has been identified
- There is a check whether strong evidence exists that would support a null/no effect or mixed rating.

Search sources

After identifying an intervention to assess, literature searches use the following sources:

- Foundations' Practice Guides, using citation searches of the underpinning systematic review
- Google Scholar
- Other evidence clearinghouses (e.g. Blueprints, CEBC)

- Academic databases where possible (e.g. Scopus).

In most cases, these sources will be sufficient to identify the main evaluation studies. Depending on how the intervention was first identified, references may already be available from the intervention's provider or Practice Guides, but supplementary searches are still carried out. At a later stage, additional studies may be requested from the evaluator if required.

Initial screening

Each study identified is briefly screened to determine whether it meets the minimum criteria to progress to the triage stage. At this stage, the following information is checked:

- **Study design:** Is the study one of the following designs: pre–post study, QED, RCT, qualitative study, mixed methods, process evaluation?
- **Is this study about the intervention of interest?** Does the study directly evaluate the intervention being assessed?
- **Is this an empirical study?** Does the study use real-world evidence, where researchers collect data through methods like surveys, interviews, experiments, or observations?
- **Does the study include insights about ethnically or racially minoritised groups or other equity-relevant subgroups for whom this intervention is particularly relevant or likely to be accessed by?** Does the study provide findings about groups who may experience different outcomes, access, or experiences related to the intervention? *(This question does not contribute to the decision on eligibility for quantitative studies but helps identify studies that may contribute qualitative EDIE insights.)*

A study reaches the eligibility criteria to progress to triage if:

- It uses an appropriate study design for its purpose (quantitative for evidence rating;

Guidebook EDIE commitment

As part of the search and screening process, we also record whether studies identified include information relevant to race and ethnicity, or other equity-relevant subgroups for whom this intervention is particularly relevant or likely to be accessed by (such as age, disability, socioeconomic status). This applies to both qualitative and quantitative evidence. The presence or absence of this information does not affect whether a study progresses to the triage stage, aside from qualitative studies. Triage decisions are based on the study design and relevance as listed above. However, the presence of EDIE information is noted at this stage to:

- Inform the EDIE narrative on the intervention summary page, which includes both qualitative and quantitative insights
- Provide context on how the intervention may work for different groups or how it may impact different groups.

This means that impact evaluations without such race and ethnicity information may still progress to triage, and qualitative studies that do not progress to triage may still be recorded and referred to at a later stage if they provide useful race and ethnicity insights.

qualitative for EDIE insights; see EDIE commitment below for further detail), and
- It is an empirical study that directly evaluates the intervention of interest.

Studies that do not meet these criteria should be excluded at this stage.

Identifying EDIE information in studies

When screening studies, information related to race or ethnicity, or information related to other identified equity-relevant subgroups is identified. At the screening stage, this information is not extracted or analysed. This information may include the following:

Quantitative evidence

- Subgroup analyses by race and/or ethnicity or other equity-relevant subgroups
- Differential impacts across racial or ethnic groups or other equity-relevant subgroups.

Qualitative evidence

- Participants experience of intervention by race or ethnicity or other equity-relevant subgroups
- Findings related to cultural relevance, adaptation, accessibility, or acceptability
- Contextual factors that may influence how the intervention is experienced by different groups.

4. Triaging evaluation studies

The purpose of the triage stage is to carry out a provisional review of study quality to prioritise which studies to take forward for full assessment. The triage stage is used to:

- Identify the most promising studies with the most potential to contribute to the highest possible Guidebook evidence rating
- Rule out lower-quality studies that are unlikely to affect the intervention's overall rating.

At triage, each study that has progressed from screening is reviewed and assigned a provisional triage rating based on our high-level Guidebook evidence criteria below. Triage ratings are provisional and are used to support prioritisation decisions rather than making final judgements about study quality.

Provisional triage ratings

Each study is assigned a provisional triage rating based on the criteria set out in table 2:

Table 2. Provisional triage rating criteria

Provisional triage rating	A study meets this rating if it:
Level 2	• Is a quantitative impact evaluation • Measures child outcomes • Uses outcome measures that appear to be validated (where this is unclear, give the benefit of the doubt at this stage unless there is an obvious issue) • Has an appropriate sample size, defined as at least 20 participants in the treatment group at the point of analysis.
Level 3	Meets all Level 2 triage criteria **AND**: • Includes a comparison group • Has an appropriate sample size, with at least 20 participants in both the control group and treatment group at the point of analysis • Where a cluster design is used (e.g. groups assigned to treatment or control), applies statistical analysis that accounts for clustering (e.g. multilevel modelling or cluster-adjusted standard errors).
Level 4	Meets all Level 3 triage criteria **AND**: • Includes long-term follow-up, defined as outcomes measures 12 months or more after the intervention • Includes at least one independent outcome measure, meaning outcome data collected independently of both the study participants and those delivering the intervention (e.g. independent observers or administrative data). Validated self-report measures may still be included but must be supplemented by at least one independent measure.

Priority ranking

Studies below Level 2 should not progress to full assessment. For all other studies, progression is based on reviewer judgement, informed by the provisional triage rating and study characteristics. A weighted priority ranking is used to support this judgement, which indicates which studies are most likely to be impactful and contribute to higher overall Guidebook ratings. The weighted score is calculated, in order, using the provisional triage rating, whether the study is UK-based, and whether the study has a racially or ethnically minoritised population focus.

The priority ranking should be used to:

- Guide the order in which studies are taken forward to full assessment
- Encourage transparent and consistent prioritisation across reviewers.

The weighting is designed to ensure that higher-quality study designs are always prioritised over lower-quality designs. For example, a Level 2 study cannot outrank a Level 3 study, even if it is UK based and includes a racially or ethnically minoritised population focus.

Table 3. Weighted scoring system

	Score
Provisional triage rating	
Contribute to Level 4	+30
Level 3	+20
Level 2	+10
NL2	0
Study location	
UK	+5
Non-UK	0
Racially or ethnically minoritised population focus	
Yes	+2
No	0

Guidebook EDIE commitment

The weighted priority ranking includes a criterion that gives additional weight to studies focusing on racially and ethnically minoritised populations. This ensures that evidence focused on minoritised groups is given additional consideration when deciding which studies progress to full assessment.

5. Extracting information from studies

The study extraction phase ensures that all relevant descriptive information from each study is captured in an organised and consistent manner. The aim of this phase is to collect granular detail on each study, including:

- Referencing details (e.g. author's last name, publication date)

- Study details (e.g. further information on the context and location in which the study was set)
- Study design (e.g. whether the study was an RCT, QED, etc.)
- Comparison group information
- Information on the intervention
- Population risk factors
- Demographics of participants in the trial
- Information on the sample in the trial
- Information on outcomes
- EDIE information (included in demographics and risk factors).

Where interventions have been identified through systematic reviews underpinning Foundations' Practice Guides, this information may be provided by the systematic review partner.

For qualitative studies relating to EDIE, some brief information is extracted including: country of study, participant characteristics, participant experiences of the intervention, barriers, and facilitators.

Guidebook EDIE commitment

In this phase, equity-related information is extracted consistently across all studies using established tools such as the PRO EDI. Any subgroup analyses or differential impacts on specific populations are extracted to inform a narrative summary on the intervention's potential to reduce disparities. Furthermore, equity-related information is extracted from implementation and qualitative studies to inform a second narrative summary focusing on implementation and the lived experiences of children and families who are receiving the intervention. This summary highlights how interventions are delivered in practice and how acceptable, relevant, and effective they are for different groups, particularly racially minoritised communities and other marginalised groups.

Information is extracted, where applicable, about intervention features that aim to promote equity. This includes approaches that were developed with or for specific communities, or that intentionally address access and inclusion. Mention of or links to adapted models are also included, and there are clear guidelines on when adaptations are assessed separately or together with the original intervention.

6. Appraising the evidence

After data extraction, studies selected in the triage process are appraised in priority order, starting with studies identified as having potential to achieve a Level 3 rating or contribute to a Level 4

intervention rating. Not all identified studies will be fully appraised; only those identified in triage as most likely to contribute to an intervention's evidence rating are taken through evidence review.

Evidence appraisal is completed by two Foundations team members, who discuss any discrepancies in assessment to come to a final internal assessment rating.

Studies are appraised as follows (see flowchart below):

1. **Does the study meet Level 2 criteria?**

   To meet Level 2 criteria a study must meet the following evidence criteria:

- Participants complete the same set of measures once shortly before participating in the programme and once again immediately afterwards.
- The sample is sufficiently large to test for the desired impact.
- For pre–post studies, overall study attrition is not higher than 40% (with at least 60% of the sample retained). For comparison group studies, overall study attrition is not higher than 65% (with at least 35% of the sample retained).
- The measures are appropriate for the intervention's anticipated outcomes and population.
- The measures are valid and reliable. This means that the measures are standardised and validated, and the methods for standardisation are published. Administrative data and observational measures might also be used to measure programme impact, but there is sufficient information to determine their validity for doing this.
- Methods used to analyse results are appropriate given the data being analysed (categorical, ordinal, ratio/parametric or non-parametric, etc.) and the purpose of the analysis.

Studies are also appraised against the following criteria, but a Level 2 rating is unlikely to be withheld on the basis of these items if necessary information is missing or unclear; reviewers are instead looking for 'red flags':

- The sample is representative of the intervention's target population in terms of age, demographics, and level of need. The sample characteristics are clearly stated.
- Measurement is independent of any measures used as part of the treatment.
- The intervention's model clearly identifies and justifies its primary and secondary outcomes and there is a statistically significant main effect of improving at least one or more of these outcomes, depending on the number of outcomes measured.

The study results must also meet the following **impact** criteria:

- There are no harmful effects
- There is evidence of a statistically significant positive impact ($p<0.05$) on at least one Foundations Guidebook child outcome
- There is consistency among the findings, resulting in few mixed results within the study.

Additionally, the following criteria are considered 'desirable', but do not drive a study rating:

- The study has clear processes for determining and reporting drop-out and dose.

- Subgroup analysis is used to verify for whom the intervention is effective and the conditions under which the effectiveness is found. (Statistically significant findings within subgroups are not treated as a replacement for a main effect.)

Studies which do not meet either the evidence standards or impact requirements of Level 2 are described as 'not Level 2' or 'NL2' and not assessed further. Studies meeting the Level 2 evidence standards assessing child outcomes but without an identified statistically significant positive child outcome **are** assessed further; if these studies meet the Level 3 standards they are given a 'no effect' or 'NE' rating.

2. **If the study meets Level 2 criteria, does it meet design specific Level 3 criteria?**

   - If a study uses a parallel RCT design, it is appraised using Cochrane's Risk of Bias 2 tool (RoB2) with Foundations' implementation guidance (see below). Studies receiving a rating of 'some concerns' or 'low risk' are then appraised against the Guidebook additional standards for RCTs.
   - If a study uses a cluster RCT design, it is appraised using Cochrane's RoB-CRT tool for cluster randomised trials with Foundations' implementation guidance. Studies receiving a rating of 'some concerns' or 'low risk' are then appraised against the Guidebook additional standards for Cluster RCTs.

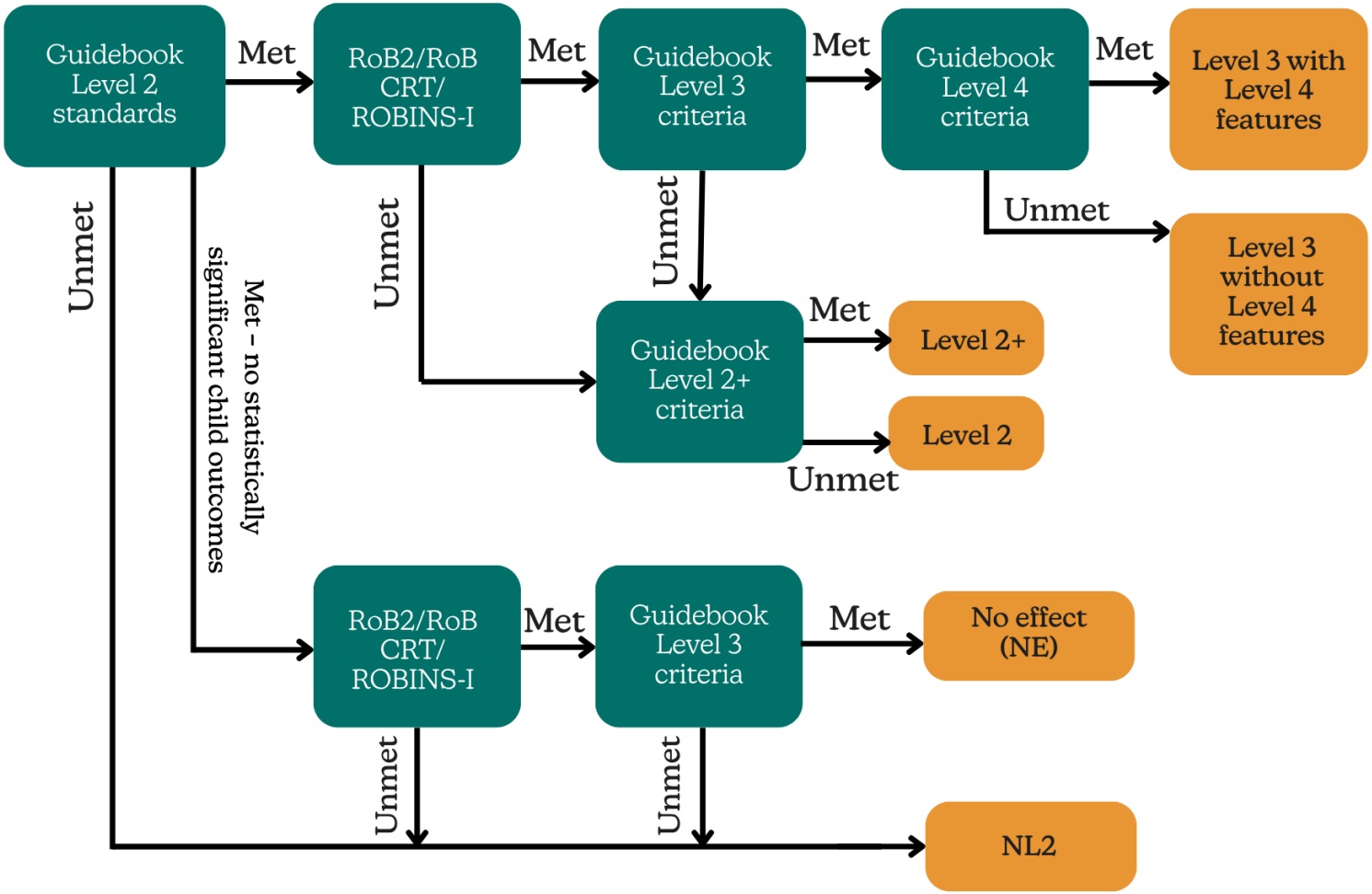
   - If a study uses a quasi-experimental design, it is currently appraised using historic Guidebook standards; these will shortly be replaced by Cochrane's Risk Of Bias In Non-randomised Studies – of Interventions, Version 2 (ROBINS-I V2) with Foundations' implementation guidance, and Guidebook additional standards for QEDs.

3. **If a study meets Level 2 criteria but not Level 3 criteria, does it meet Level 2+ criteria?** Studies meeting Level 2 criteria with a comparison group design that meet the Level 3 baseline equivalence criteria are awarded a Level 2+ rating.

4. **If the study meets Level 3 criteria, does it also meet criteria for a potential Level 4 / Level 4+ intervention rating?** Individual studies are not awarded a rating higher than Level 3; however, interventions may be eligible to receive an intervention rating of Level 4 or Level 4+ if additional criteria are met in one of the contributing studies. The following potential Level 4 features are flagged in the evidence appraisal of individual studies, ahead of the generation of an intervention rating:

   - At least one evaluation uses a form of measurement that is independent of the study participants (and also independent of those who deliver the programme). In other words, self-reports (through the use of validated instruments) might be used, but there is also assessment information independent of the study participants (e.g. an independent observer, administrative data, etc.)
   - There is evidence of a long-term outcome of 12 months or more.
   - **For Level 4+:** At least one of the effectiveness evaluations will have been conducted independently of the programme developer.
   -

Figure 2. Assigning a study evidence rating (link to long descriptive text)



Implementation guidance for Cochrane's RoB2 for randomised trials in children's social care and early intervention

This implementation guidance is intended to support the use of Cochrane's Risk of Bias tool version 2 (RoB2; Sterne et al., 2019) in the assessment of risk of bias in randomised controlled trials in children's social care or early intervention.

It is designed to be used alongside the RoB2 cribsheet and RoB2 full guidance documents.

- For some signalling questions, no additional guidance has been provided, and this is indicated.
- For some questions, this guidance provides clarification of the RoB2 guidance (e.g. simpler wording) or examples of application specific to children's social care; for these questions, the guidance begins 'following RoB2 guidance'.
- For other signalling questions, this guidance provides substantial detail of how to appraise studies, for example adding in specific thresholds for attrition or characteristics of measures, where this is underspecified or inappropriate for children's social care research in the RoB2; for these questions, this is also indicated at the start of the guidance.

Appraising contrasts

Note that the guidance assumes that reviewers follow RoB2 guidance to appraise each outcome or result from a study separately (Sterne et al., 2019). In this guidance we refer to these as **contrasts** (as per Prevention Services Clearinghouse (Wilson et al., 2024; PSC)); a contrast is defined as **"a comparison of an eligible intervention condition to an eligible comparison condition on a specific outcome for a specific posttest measurement"** (Wilson et al., 2024).

> "For example, a study with one intervention condition and one comparison condition that reports findings on one outcome measured immediately after treatment has a single contrast. A study with one intervention condition and one comparison condition that reports findings on two outcomes measured immediately after treatment would have two contrasts, one for each of the comparisons between the intervention and comparison conditions on the two outcomes. A study with one intervention condition and one comparison condition that reports findings on one outcome measured at 3, 6, and 12 months after treatment would have three contrasts, one for each outcome measurement period." (Wilson et al., 2024; PSC 5.1)

In practice, reviewers often do not assess every possible contrast in a research paper, and the appropriate contrast or contrasts for review are determined by the scope and focus of the systematic review or project the risk of bias assessment is feeding into. When this is the case, it is still often useful to take note of any contrasts that are likely not to receive the same risk of bias judgement as the contrast in focus, for example if the paper reports high attrition for a later follow-up, or if a secondary outcome reported by the paper uses a measure likely to be at high risk of bias. Maintaining this awareness that not all contrasts within a study are equal(ly biased or unbiased)

will prevent reviewers from inappropriately extrapolating a risk of bias judgement from one contrast to all contrasts in a paper. This is especially important where there are mixed findings in the study, including null or negative findings, so that publication bias is not exacerbated.

Selection of contrasts to appraise depends on the purpose of the review at hand. Reviews should have a clear protocol which specifies the outcomes of interest, whether this is all outcomes (in the case of the Foundations Guidebook), or particular outcomes (typically, in the case of systematic reviews).

Arriving at a Risk of Bias judgement

This implementation guidance is provided for each RoB2 signalling question; reviewers should use this additional guidance to support signalling question judgements. They should use the algorithms provided by RoB2 (Sterne et al., 2019) to arrive at an overall risk of bias judgement for a contrast.

The RoB2 tool uses signalling questions and decision 'algorithms' to guide reviewers to a judgement of risk of bias for five separate domains. These signalling questions and algorithms lead to five domain-level judgements of either 'low risk', 'some concerns', or 'high risk'.

The contrast is considered at low risk of bias overall if "the study is judged to be at low risk of bias for all domains for this [contrast]" (Sterne et al., 2019).

The contrast is considered to have some concerns overall if "the study is judged to raise some concerns in at least one domain for this [contrast], but not to be at high risk of bias for any domain" (Sterne et al., 2019).

The contrast is considered to be at high risk of bias overall if "the study is judged to be at high risk of bias in at least one domain for this [contrast]" OR "the study is judged to have some concerns for multiple domains in a way that substantially lowers confidence in the [contrast]" (Sterne et al., 2019).

This implementation guidance specifies that if 4/5 or 5/5 domains are judged to be 'some concerns' (and the remainder 'low risk'), this results in an overall judgement of 'high risk'. If 3/5 domains are judged to be 'some concerns' (and the remainder low risk), then reviewer judgement is required as to whether this makes the overall risk of bias judgement 'some concerns' or 'high risk'; reviewers should consider the particular reasons why the domains have been given 'some concerns'. If 1/5 or 2/5 domains are judged as 'some concerns', and the rest 'low risk', then the overall judgement is 'some concerns'.

Sources of guidance

For reference, this guidance includes its sources – either Cochrane's Revised Risk of Bias tool for randomised trials (Sterne et al., 2019; RoB2), the Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures version 2.0 (Wilson et al., 2024; PSC), or previous Foundations Guidebook appraisal standards (Guidebook). Where there is a direct quotation from or reference to a section in the RoB2 full guidance, this is referenced (Higgins et al., 2019). Where a source is not indicated, the guidance has been formulated for this document in consultation with external experts.

Risk of bias arising from the randomisation process

1.1 Was the allocation sequence random?

Following RoB2 guidance, appropriate random assignment may be simple, or use block or stratification, or minimisation techniques (Sterne et al., 2019; RoB2), and is at the appropriate level (individual/family/group/centre) for the design and research questions (Guidebook).

To determine whether the level is appropriate, consider the types of effects the intervention is expected to produce: if the intervention is designed to improve outcomes for individual children, young people or families, then individual-level or family-level randomisation is typically appropriate. Cluster randomisation at a higher-level (group, Family Hub, local authority area, etc.) can also be appropriate here, depending on the implementation and delivery. This is typically the kind of intervention seen in appraising trials in children's social care and early intervention. However, if the effect of interest is at a higher level, for example an impact on a community such as crime rate, then randomisation should be at this higher, cluster level.

(NB if cluster randomisation, use the cluster randomised trial implementation guidance alongside RoB2 for cluster-randomised trials (Eldridge et al., 2021).)

Examples of non-random allocation include allocation decisions influenced by baseline assessments (e.g. where those with higher need are assigned to the intervention condition), or based on factors like participants' date of birth, or day or time of assignment.

Randomisation could also be compromised by reclassifying those who refused to participate in the intervention into the control group; assigning participants to the intervention group to achieve a target sample number; or one site refusing randomisation but still being included in the study.

Randomisation is not compromised if additionally recruited individuals are not included in analysis; or if participants are excluded because it was discovered they did not meet the inclusion criteria (Wilson et al., 2024; PSC 5.4). For example, imagine a trial of a group intervention being run in Family Hubs: some parents are included in the intervention group at the last minute as some spaces are still available and they consent to be in the trial; they are not randomised to the intervention group but added directly. If these additionally recruited families are ***excluded*** from the analysis, then randomisation is not compromised.

1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?

Following RoB2, randomisation allocation should be concealed until enrolment is complete. Participants, practitioners, and researchers should be blinded to allocation prior to participant enrolment – that is, until after consent has been obtained and eligibility confirmed.

RoB2 guidance does not specify the preferred or accepted order of enrolment (consent), baseline data collection and randomisation. While this order is preferred as lowest risk, studies are not judged as higher risk for another order, as long as there is concealment of the allocation sequence

until enrolment and assignment, and there are no other causes for concern around randomisation (see 1.3).

1.3 Did baseline differences between intervention groups suggest a problem with the randomisation process?

Following RoB2 guidance, no additional calculations are required to judge baseline differences; it is acceptable for reviewers to look at the data for baseline differences, and identify any that stand out as red flags indicating a problem with randomisation. That is, the point of this item is not to establish that there is baseline equivalence on key variables (see Domain 3), but to check that there are no differences which indicate that randomisation was compromised. For this item, both outcome variables at baseline and demographic characteristics can be considered.

RoB2 guidance references statistically significant differences in p-values (Sterne et al., 2019); it is acceptable to consider either p-values or effect sizes (for example, Cohen's *d*); when both are available, effect sizes are strongly preferred, given that significance testing with p values with an inherently random sample does not make statistical sense. The thresholds for acceptable effect sizes of baseline difference outlined in Domain 3 can be used as guides for problematic differences here, but are not prescriptive; reviewer judgement is required.

For example, consider an evaluation of a parenting intervention designed to enhance support for families with a Child in Need or otherwise engaging with a social worker, in which parents in the intervention group are on average younger and more socioeconomically disadvantaged, and score more highly on social isolation and parenting stress than those in the control group, with consistent effect sizes of the difference of d > 0.3. This suggests that randomisation may have been compromised by somehow allocating families with higher levels of need to the intervention condition. Depending on the description of the trial design and other details available, reviewers may judge that baseline differences did suggest a problem with randomisation here. However, if only one or two characteristics differ substantially at baseline, this is consistent with chance (see RoB2 guidance).

Risk of bias due to deviations from the intended interventions (effect of assignment to intervention)

2.1 Were participants aware of their assigned intervention during the trial? AND

2.2 Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?

Following RoB2, if participants (typically children, young people, and their families), carers, and people delivering the interventions (typically practitioners, such as family workers, social workers, or therapists) are all blinded to the intervention groups or conditions during the trial, the contrast is considered at low risk of bias for Part 1 of this domain. Note that this is highly unlikely in early intervention and children's social care trials, because in social interventions both participants and practitioners tend to know which intervention they are receiving and delivering. This means that a full consideration of deviations arising from the trial context is necessary (see 2.3).

2.3 Were there deviations from the intended intervention that arose because of the trial context?

Slightly different from RoB2 guidance, reviewers should look for explicit mention of deviation, or clues that deviation may have occurred. This is a slightly relaxed interpretation of the RoB2 guidance and algorithm, because it is unlikely that a study report will explicitly rule out all types of deviation if it did not occur; following the RoB2 guidance explicitly, where no information is given an NI judgement and leads to 'some concerns', could lead to unfairly penalising a study overall.

If there is nothing to suggest that deviations occurred, then it may be concluded that no deviations occurred and a PN/N judgement can be given, but reviewer judgement is required.

Such deviations include crossover, practitioner changes to the protocol and imbalance in non-protocol interventions (RoB2), but also diffusion, compensatory rivalry, and resentful demoralisation:

5. **Crossover:** control group participants receive the services the intervention group participants are assigned to (Sterne et al., 2019).
6. **Diffusion:** groups interact with each other either directly or indirectly, such that the nature of the treatment becomes known to the control group. This risks the control group sharing in the benefits of the treatment, and so contaminating the control group and biasing estimates of intervention effect; for example if intervention participants are residential care workers, and outcomes are measured in key index children, diffusion may occur if control group children share the same care home and either directly or indirectly interact with the intervention group residential care worker.
7. **Compensatory rivalry:** where the comparison group becomes aware that they are being denied an advantage or desirable intervention, and as a consequence, they increase their efforts to succeed and achieve positive outcomes; this could also include seeking alternative interventions.
8. **Resentful demoralisation:** where the comparison group become aware that they are being denied an advantage or desirable treatment, and as a consequence, they become demoralised, discouraged, and/or angry and give up (Guidebook).
9. **Practitioner changes to the protocol**, for example due to their own knowledge or concerns about implementation.
10. **Imbalance in non-protocol interventions across groups**, for example if individuals in the intervention group, through attendance at a centre for intervention participation, additionally become aware of and access other support services to a greater extent than individuals in the control group, this would introduce bias (Sterne et al., 2019).

2.4 Were these deviations likely to have affected the outcome?

Following RoB2, if there are identified deviations, consider whether these are likely to have an effect on the outcome. Examples where the outcome may be affected include:

- If a practitioner chooses to supplement an intervention with their favoured treatment approach that has not been pre-specified (e.g. a family worker adds motivational interviewing to a manualised parenting intervention which did not include this)

- If participants in a control group seek out, or are directed to, additional support that they would not have been likely to consider without knowledge of the intervention (e.g. through being recruited to a trial at a Family Hub, parents are specifically referred to an alternative BAU intervention).

2.5 Were these deviations from intended intervention balanced between groups?

Following RoB2, deviations are considered balanced between groups if the same deviation or deviations happen in both groups, and are judged to have affected the same proportion of participants in each group. This signalling question requires reviewer judgement. For example, if family workers add motivational interviewing to their support of families in both the intervention and control group, this deviation would be balanced across groups; if it is added to the intervention group only, it would not be balanced – in that case, the effect being observed is of the intervention plus motivational interviewing, and this would need to be taken into account when deciding whether to include the study in the review.

Note that if it is not clear from the paper that any deviations are balanced between groups, a NI judgement is given, and is treated in line with an N or PN judgement, according to the RoB2 algorithm Figure 2 in RoB2.[8]

2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?

Following RoB2, an intention-to-treat design is used, meaning that all participants recruited to the intervention and control arms participate in the pre–post measurement and analysis, regardless of whether or how much of the intervention they receive, even if they drop out of the intervention (this does not include dropping out of the study – which is then regarded as missing data) (Guidebook).

Some studies may use 'modified ITT / mITT', which is the same as ITT with complete case analysis or listwise deletion; this is also acceptable (Sterne et al., 2019; RoB2).

If the design is not described as 'intention-to-treat' in the study report, but it is clear from the CONSORT diagram or other reporting of the flow of participants through the trial that ITT is the approach used, the reviewer can also respond Y/PY.

2.7 Was there potential for a substantial impact (on the [contrast]) of the failure to analyse participants in the group to which they were randomised?

As it is often not possible to specify what would constitute a 'substantial impact', and in children's social care and early intervention trials, reassignment might often be linked to prognostic factors, if an ITT or mITT analysis has not been used a Y/PY/NI judgement is appropriate for this criterion, leading to a 'high risk' judgement.

---

8 See Figure 2 Algorithm for suggested judgement of risk of bias due to deviations from the intended interventions (Higgins et al., 2019 p 33)

Some research papers report both ITT and 'per protocol' analyses; in this case it may be possible to judge the degree of impact of the alternative analysis method. However, in these studies the results of the ITT analysis should be appraised for risk of bias and the per protocol results discounted.

Risk of bias due to deviations from the intended interventions (effect of adhering to intervention)

Do not use this domain; all studies should be appraised for the effect of assignment to intervention version of Domain 2 on RoB2.

Risk of bias due to missing outcome data

3.1 Were data for this outcome available for all, or nearly all, participants randomised?

This implementation guidance for Domain 3 operationalises the RoB2 principle that "the number of participants with missing outcome data is sufficiently small that their outcomes, whatever they were, could have made no important difference to the estimated effect of intervention". In children's social care and early intervention trials, there is often relatively high overall attrition, because of the challenging context interventions are implemented in, and the characteristics of the vulnerable children and families receiving interventions. Furthermore, attrition is typically linked to children and families' characteristics and circumstances, which are also predictive of the outcome.

Therefore, a more holistic approach is taken, looking at both overall and differential attrition (3.1) and its effects (3.2). The clear sliding thresholds, taken from Prevention Services Clearinghouse / What Works Clearinghouse, are used in order to facilitate consistent reviewer judgement.

For question 3.1, if levels of overall and differential attrition are within the acceptable limits described by Prevention Services Clearinghouse (Wilson et al., 2024), described below, answer 'Y'. If either differential or overall attrition is not reported, answer 'NI' for signalling question 3.1.

Attrition is calculated from randomisation; randomisation of individuals to conditions is considered to have occurred once individuals learn their assignment condition (Wilson et al., 2024; PSC).

Overall and differential attrition are within acceptable limits if the combination falls within the green area on the figure below (from the What Works Clearinghouse (WWC)); see the table following the figure for maximum acceptable differential attrition rates dependent on overall attrition rate (Wilson et al., 2024; PSC).

Overall attrition is attrition across all participants, from randomisation to when the outcome being appraised was measured. Differential attrition is the difference between attrition in the treatment group and attrition in the control group.

For example, if a trial started with 100 participants, with 50 in each group, and included 80 participants at endline, then overall attrition would be 20%. In the treatment group, 5 participants dropped out of the study, meaning 10% attrition, while in the control group, 15 participants dropped out of the study, meaning 30% attrition. Differential attrition is therefore also 20%. This

places attrition in the 'red zone', over the acceptable limits, and so the reviewer must answer N, and proceed to 3.2.

In another case, a trial started with 100 participants, with 50 in each group, and included 80 participants at endline, so overall attrition is again 20%. This time, 11 participants in the treatment group drop out of the study (22% attrition), and 9 participants in the control group (18% attrition), so differential attrition is 4%. In this case, attrition is in the 'green zone', within acceptable limits, and so the reviewer can answer Y. Following the RoB2 algorithm, this means that the whole of Domain 3 is at low risk of bias.

Figure 3. Potential bias associated with overall and differential attrition (Exhibit 5.3. from the *Prevention Services Clearinghouse Handbook v2*) (link to long descriptive text)

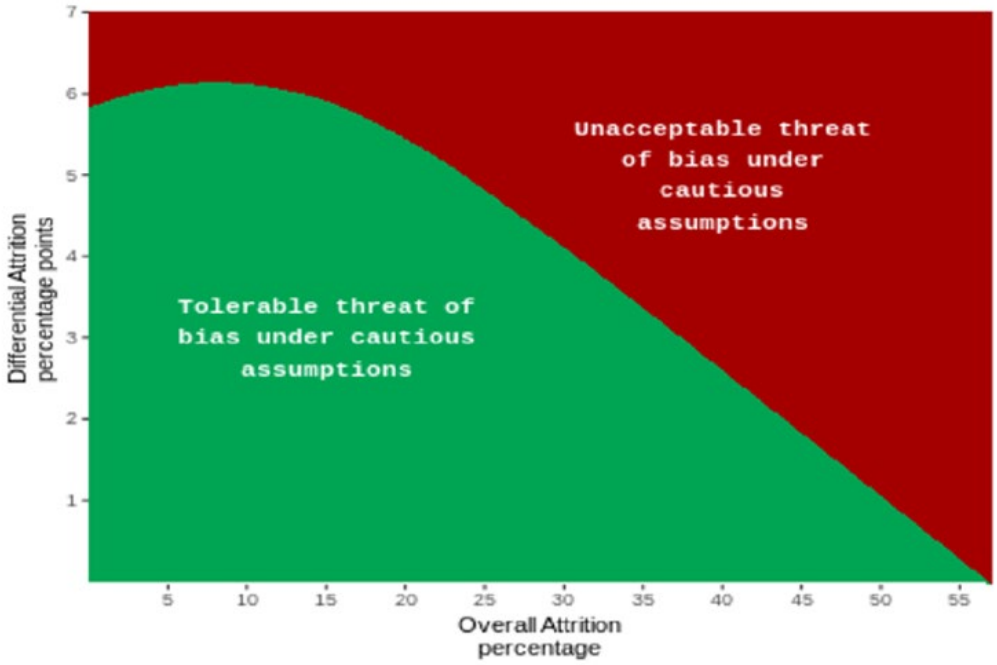


Exhibit 5.3 from Prevention Services Clearinghouse Handbook v2 p60

Table 4. Prevention Services Clearinghouse Attrition Boundaries (Exhibit 5.4. from the *Prevention Services Clearinghouse Handbook v2*) (link to long descriptive text)

Exhibit 5.4. Prevention Services Clearinghouse Attrition Boundaries

Overall Attrition	Differential Attrition	Overall Attrition	Differential Attrition	Overall Attrition	Differential Attrition
0	5.7	20	5.4	40	2.6
1	5.8	21	5.3	41	2.5
2	5.9	22	5.2	42	2.3
3	5.9	23	5.1	43	2.1
4	6.0	24	4.9	44	2.0
5	6.1	25	4.8	45	1.8
6	6.2	26	4.7	46	1.6
7	6.3	27	4.5	47	1.5
8	6.3	28	4.4	48	1.3
9	6.3	29	4.3	49	1.2
10	6.3	30	4.1	50	1.0
11	6.2	31	4.0	51	0.9
12	6.2	32	3.8	52	0.7
13	6.1	33	3.6	53	0.6
14	6.0	34	3.5	54	0.4
15	5.9	35	3.3	55	0.3
16	5.9	36	3.2	56	0.2
17	5.8	37	3.1	57	0.0
18	5.7	38	2.9		
19	5.5	39	2.8		

Source. What Works Clearinghouse (n.d.)

Note. Overall attrition rates are given as percentages. Differential attrition rates are given as percentage point differences. Attrition computations are rounded to whole numbers for determining overall attrition and to the nearest hundredth for differential attrition. For example, an overall attrition rate of 15.4% and differential attrition rate of 5.894pp would be rounded to 15% and 5.89pp, respectively. This contrast would be evaluated as low attrition because 5.89pp falls below the boundary of 5.9pp.

3.2 Is there evidence that the result was not biased by missing outcome data?

As explained in 3.1 above, guidance for 3.2 operationalises the RoB2 signalling question with clear thresholds to enable judgement of when the result is or is not likely to be biased by attrition.

Step 1: Does the outcome measure at baseline (or an appropriate proxy, see below) demonstrate baseline equivalence in the analytic sample (d <0.05)?

Baseline equivalence of a measure is defined as a difference of d<0.05. When examining baseline equivalence, we are interested in the analytic sample**;** that is, the final sample used in analysis to generate the reported results, as opposed to the randomised sample.

To establish baseline equivalence, the following can be used [in order of preference]:

1. **Direct pretest.** Defined as the same (or nearly the same) measure used for the outcome.
2. **Correlated pretest.** Defined as a measure in any eligible outcome domain (PSC 4.1.8) that has a correlation of 0.60 or higher with the outcome in the analytic sample. (A

correlation shown in the comparison condition only is also acceptable). Correlated pretests do not have to be in the same or similar domain as the outcome.

3. **Pretest alternative.** Defined as a measure in the same or similar domain as the outcome. No correlation threshold is specified – pretest alternatives are generally assumed to be correlated with the outcome by virtue of being conceptually related or are common precursors to the outcome [there are conditions under which pretest alternatives are acceptable].
4. **Sociodemographic characteristics.** Under certain conditions specified below, combinations of individual or individual and neighbourhood sociodemographic characteristics may be used to establish baseline equivalence. Eligible individual and neighbourhood characteristics are summarised below.

a. **Individual sociodemographic characteristics.** Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, (3) Household composition, and (4) Age of sample members.
b. **Neighbourhood sociodemographic characteristics.** Neighbourhood is defined as an Indices of Deprivation Lower-layer Super Output Area (LSOA); postcode; ONS census Output Area; or other small geographical area. In US, this is defined as a census tract, ZIP Code, or smaller geographic unit (or similarly sized tabulation unit for studies conducted outside of the United States). Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, and (3) Household composition (Wilson et al 2024; PSC pp. 65–66).

   Note that measures must meet all measurement standard requirements, i.e. reliability, face validity, and consistency, and must be completed before the start of the intervention.

Effect sizes can be calculated via the Effect Size Calculator – Campbell Collaboration if required.

If the effect size of the difference at pretest (or alternative) is d<0.05, answer Y/PY to signalling question 3.2. If the effect size of the difference at pretest (or alternative) is d > 0.05, proceed to step 2.

Step 2: If the outcome measure at baseline (or an appropriate proxy, see below) does not demonstrate baseline equivalence in the analytic sample, then is the difference moderate and controlled for in analysis?

If the effect size of the difference at pretest is moderate (d > 0.05 but < 0.25), then it must be controlled for in analyses, or there must be sensitivity analyses showing that there is no need to control for the difference in the analysis; if it is controlled for in analyses or sensitivity analyses are satisfactory answer Y/PY to signalling question 3.2. If the effect size is large, not controlled for in analyses or there are no sensitivity analyses, proceed to step 3.

Step 3: If the effect size of the difference at pretest is large (d > 0.25), or is moderate and not controlled for in analyses, answer N/PN to signalling question 3.2. This is evidence that the result was biased by the missing outcome data.

Example 1: In an RCT of a group parenting intervention, differential attrition puts the trial in the 'red area', with attrition over the acceptable limit. The primary outcome measure is a measure of children's behaviour, and this is measured at baseline and after the intervention. This is therefore

the preferred way of establishing whether there is baseline equivalence between the groups. The effect size of the difference between the groups is calculated for the *analytic sample* at baseline, and found to be d = 0.15. This means that there is *not* baseline equivalence, but it is possible to address an effect of this size in analysis. The outcome measure at baseline is included in the regression analysis, and therefore the reviewer can respond Y.

Example 2: In a different trial, attrition is again over the acceptable threshold, and there is no outcome measurement at baseline, because it is a perinatal intervention, with a child outcome. There is also no pretest alternative, so demographic characteristics are considered. Demographic characteristics in the analytic sample have an effect size of d = 0.15 but they are not included in the analysis, and there are no additional sensitivity analyses available. Therefore the reviewer must respond N.

3.3 Could missingness in the outcome depend on its true value?

Following RoB2, if missingness (attrition) is not within acceptable limits described in 3.1, and there is not evidence that the result is not biased (3.2), Domain 3 can still be considered low risk as a whole if there is evidence that data is missing completely at random (MCAR), for example if data is missing as a result of data file corruption that was not in any way related to the outcome. It is relatively rare and highly unlikely that missing data is unrelated to the outcome; in most cases this signalling question should be answered Y/PY.

If there is a clear explanation in the research paper to justify why outcome data can confidently be considered MCAR, this signalling question can be answered N/PN (Sterne et al., 2019).

3.4 Is it likely that missingness in the outcome depended on its true value?

In children's social care and early intervention research, if data is not missing completely at random (see 3.3), it is reasonable to assume that it is "likely that missingness in the outcome depended on the true value" (Sterne et al., 2019), that is, that the reasons behind participants dropping out of the study are linked in some way to the outcome of interest. Following this implementation guidance, this signalling question is only considered for studies in which there are high levels of missing data and a significant difference in the outcome measure at baseline (or an appropriate proxy) which is too large to be controlled for in analyses, and data is not MCAR. Always answer Y/PY to this signalling question.

Risk of bias in measurement of the outcome

4.1 Was the method of measuring the outcome inappropriate?

In children's social care and early intervention trials, adapted or bespoke measures developed for the intervention are common. It is therefore particularly important to check that valid and reliable measures have been chosen appropriately. This implementation guidance for signalling question 4.1 provides four characteristics of the outcome measure, which the outcome measure must all meet, to determine that the method of measuring the outcome was not inappropriate. These characteristics are based on previous Guidebook and Prevention Services Clearinghouses' requirements for measures.

1. The measurement is independent of the treatment.

The measure is not used as part of the intervention (for example, by the practitioner to monitor and determine the content of intervention sessions); this is sometimes known as a 'treatment inherent' measure. (Guidebook)

If there is no information available, consider this criterion met.

2. An appropriate measure was used, for the population

The measure is appropriate for the vast majority of the sample population in terms of age, level of need, and country, ethnicity, culture, and language. Reviewer judgement is required to decide which of these factors are most important to demonstrate for the measure used. It has been designed for and tested or normed with a similar sample to that in the study. (Guidebook)

For example, if a measure designed to assess externalising behaviour in 5–10-year-olds is used with 2–4-year-olds, the method of measuring the outcome is inappropriate; similarly, a measure developed for the general population may not be appropriate for a sample who has experienced complex trauma, unless the measure had been separately tested with the higher needs population.

If there is no information available, consider this criterion met.

3. The measure has high face validity

To satisfy the criterion for face validity, there must be a sufficient description of the outcome or baseline measure for the reviewer to determine that the measure is clearly defined, has a direct interpretation, and appears to measure the construct it was designed to measure (Wilson et al., 2024; PSC p. 73). In other words, at face value, the measure does what it says it does (e.g. does a measure said to measure anxiety actually measure anxiety, rather than another construct?)

This criterion requires reviewer judgement based on information in the paper or easily available; if there is no information available, consider this criterion NOT met.

4. The measure has high reliability

The outcome or baseline measure either must be a measure which is assumed to be reliable (see below) or must meet one or more of the following standards for reliability (depending on what is appropriate for the type of measure):

- Internal consistency (such as Cronbach's alpha) of 0.50 or higher.
- Test–retest reliability of 0.40 or higher
- Interrater reliability (correlation) of 0.50 or higher
- Interrater agreement (percentage agreement or kappa) of 0.80 or higher for percentage agreement and 0.60 or higher for kappa.

When required, reliability statistics may be based on an independent sample (with similar characteristics to the study sample) and/or the study sample. An independent sample is strongly preferred, as this avoids the risk of a novel tool being overfitted on the study sample; however, a demonstration of reliability from the study sample may be allowed with careful reviewer

judgement. For example, where a measurement tool is already well established with extensive validation and standardisation, and a cultural or linguistic adaptation is developed and used in the study in question, with reliability demonstrated in the study sample, this may be allowed with reviewer judgement. If there is no information in either the study under review or additional sources and the measure is not assumed to be reliable (see below), consider the criterion NOT met.

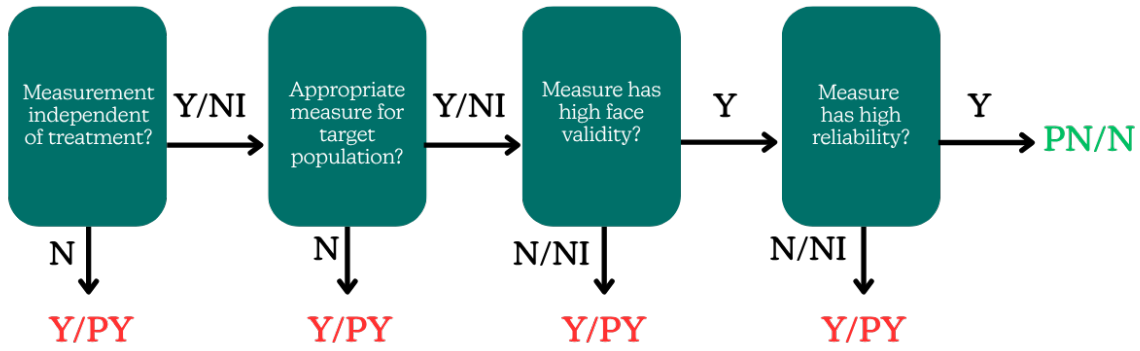
Some types of measures are normally assumed to be reliable, including:

- Demographic characteristics, such as age, race/ethnicity, education level, SES, employment status, etc.
- Medical or physical tests, such as urinalysis, weight measurement, etc.

Administrative data can often be assumed to be reliable, but can vary substantially in quality and so reviewer judgement is required to decide whether this criterion is met; shortcomings of administrative data may be described in the study, or information about the dataset can be checked. Administrative data includes records obtained from schools, child welfare or other social service agencies, hospitals, or clinics.

If all four characteristics are met, then a judgement of PN/N can be given; if at least one is not met, then a judgement of Y/PY is given.

Figure 4. Flowchart illustrating steps to judge whether the measure is not inappropriate; answers to questions lead to RoB2 judgement for signalling question 4.1 (link to long descriptive text)



4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?

Following RoB2, to determine whether measurement or ascertainment of the outcome could have differed between intervention groups, consider whether:

1. There is equivalent measurement of groups in terms of timing

The time between pretest (baseline) and posttest (outcome) does not systematically differ between intervention and comparison conditions. If there is a systematic difference in timing between groups, this criterion is not met; for example, if the treatment group received the posttest after 12 weeks, while the control group received it after 8 weeks (Wilson et al., 2024; PSC p. 74).

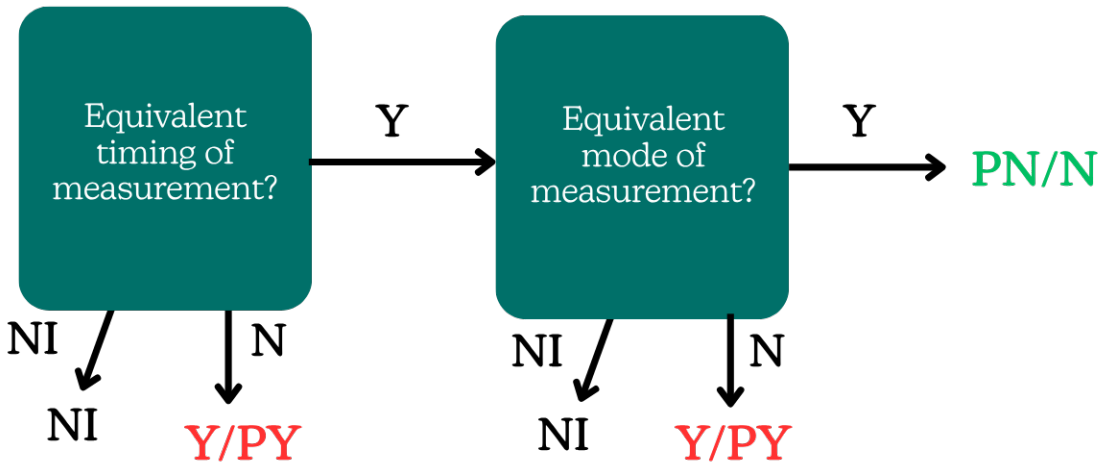
2. There is equivalent measurement of groups in terms of mode

The data collectors and data collection modes for data collected from intervention and comparison conditions either are the same, or are different in ways that would not be expected to have an effect on the measures (Wilson et al., 2024; PSC p. 74).

For example, one group being measured through a face-to-face interview and the other group through an online survey would not be acceptable; similarly if systematically different types of practitioners or staff administered a questionnaire or interview between groups, and the type of staff is likely to have affected participants' responses, this would not be acceptable.

If a measure is equivalent in timing and mode across groups, respond PN/N; if it is not equivalent on one or both characteristics, respond PY/Y.

Figure 5. Flowchart illustrating how to establish whether measurement differed between groups; answers to the questions lead to RoB2 judgement for signalling question 4.2 (link to long descriptive text)



4.3 Were outcome assessors aware of the intervention received by study participants?

No additional guidance needed.

4.4 Could assessment of the outcome have been influenced by knowledge of intervention received?

Following RoB2, consider whether the type of measure means that an unblinded assessor could have influenced the outcome. For example, if the assessor is administering an objective cognitive test or questionnaire, without seeing or hearing the participants' responses (e.g. on a computer or by giving the participant a pen-and-paper questionnaire to fill in themself), this is unlikely to have affected the outcome; if the assessor is leading and/or coding an interview with the participant, their awareness could influence the outcome.

Participant self-reports are common in children's social care and early intervention research, and for self-reports always answer Y.

4.5 Is it likely that assessment of the outcome was influenced by knowledge of intervention received?

Following RoB2 (with the exception of self-reports, see below), if outcome assessors are (or may be) aware of the intervention received, and it is possible that awareness could have biased the outcome, then judgement must be used to decide whether this is likely to have affected the participants' outcomes. As stated in the RoB2 guidance, this is more likely to be the case when there are strong levels of belief in either beneficial or harmful effects of the intervention (Wilson et al., 2024; PSC; Higgins et al., 2019; 4.5 p. 18), for example assessments of recovery by a practitioner who delivered the intervention.

Self-reports are potentially influenced by knowledge of the intervention received, because they are inherently 'unblinded', and judgement is needed to determine whether it is likely that participants' reporting of the outcome was influenced by knowledge of the intervention received. If the measure is valid and reliable, benefit of the doubt can often be given leading to a 'PN' judgement. This is a slightly relaxed implementation of RoB2, as self-reports are both widely used in children's social care research, and, being widely used, often have good psychometric properties (are valid and reliable).

Risk of bias in selection of reported result

5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalised before unblinded outcome data were available for analysis?

No additional guidance needed.

Is the numerical result being assessed likely to have been selected, on the basis of the results, from …

5.2 … multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain?

No additional guidance needed.

5.3 … multiple eligible analyses of the data?

No additional guidance needed.

RoB CRT implementation guidance for cluster randomised trials in children's social care and early intervention

This implementation guidance is intended to support the use of Cochrane's Risk of Bias tool version 2 for cluster trials (Eldridge et al., 2021; RoB2 CRT) in the assessment of risk of bias in randomised controlled trials in children's social care or early intervention.

It is designed to be used alongside the RoB CRT cribsheet and RoB CRT full guidance documents.

- For some signalling questions, no additional guidance has been provided, and this is indicated.
- For some questions, this guidance provides clarification of the RoB CRT guidance (e.g. simpler wording) or examples of application specific to children's social care; for these questions, the guidance begins 'following RoB CRT guidance'.
- For other signalling questions, this guidance provides substantial detail of how to appraise studies, for example adding in specific thresholds for attrition or characteristics of measures, where this is underspecified or inappropriate for children's social care research in the RoB CRT; for these questions, this is also indicated at the start of the guidance.

Cluster designs

For trials where groups of individuals are randomised, rather than individuals, the trial has a cluster design. Clusters may include local authorities or geographic areas, Family Hubs, clinical settings, intervention providers, or families. A trial may not be explicitly referred to as a cluster randomised trial in the report, but if randomisation is not at the individual level, it should be treated as a CRT.

Appraising contrasts

Note that the guidance assumes that reviewers follow RoB2 CRT guidance to appraise each outcome or result from a study separately (Eldridge et al., 2021). In this guidance we refer to these as **contrasts** (as per Prevention Services Clearinghouse (Wilson et al., 2024; PSC)); a contrast is defined as **"a comparison of an eligible intervention condition to an eligible comparison condition on a specific outcome for a specific posttest measurement"** (Wilson et al., 2024).

> "For example, a study with one intervention condition and one comparison condition that reports findings on one outcome measured immediately after treatment has a single contrast. A study with one intervention condition and one comparison condition that reports findings on two outcomes measured immediately after treatment would have two contrasts, one for each of the comparisons between the intervention and comparison conditions on the two outcomes. A study with one intervention condition and one comparison

> condition that reports findings on one outcome measured at 3, 6, and 12 months after treatment would have three contrasts, one for each outcome measurement period." (Wilson et al., 2024; PSC 5.1)

In practice, reviewers often do not assess every possible contrast in a research paper, and the appropriate contrast or contrasts for review are determined by the scope and focus of the systematic review or project the risk of bias assessment is feeding into. When this is the case, it is still often useful to take note of any contrasts that are likely not to receive the same risk of bias judgement as the contrast in focus, for example if the paper reports high attrition for a later follow-up, or if a secondary outcome reported by the paper uses a measure likely to be at high risk of bias. Maintaining this awareness that not all contrasts within a study are equal(ly biased or unbiased) will prevent reviewers from inappropriately extrapolating a risk of bias judgement from one contrast to all contrasts in a paper. This is especially important where there are mixed findings in the study, including null or negative findings, so that publication bias is not exacerbated.

Selection of contrasts to appraise depends on the purpose of the review at hand. Reviews should have a clear protocol which specifies the outcomes of interest, whether this is all outcomes (in the case of the Foundations Guidebook), or particular outcomes (typically, in the case of systematic reviews).

Arriving at a Risk of Bias judgement

This implementation guidance is provided for each RoB2 CRT signalling question; reviewers should use this additional guidance to support signalling question judgements. They should use the algorithms provided by RoB2 CRT (Eldridge et al., 2021) to arrive at an overall risk of bias judgement for a contrast.

The RoB2 tool uses signalling questions and decision 'algorithms' to guide reviewers to a judgement of risk of bias for five separate domains. These signalling questions and algorithms lead to five domain-level judgements of either 'low risk', 'some concerns', or 'high risk'.

The contrast is considered at low **risk of bias overall** if "the study is judged to be at **low risk of bias for all domains** for this [contrast]" (Higgins et al., 2019).

The contrast is considered to have **some concerns overall** if "the study is judged to raise **some concerns** in at least one domain for this [contrast], but not to be at high risk of bias for any domain" (Higgins et al., 2019).

The contrast is considered to be at **high risk of bias overall** if "the study is judged to be at **high risk of bias** in at least one domain for this [contrast]" OR "the study is judged to have **some concerns** for **multiple domains** in a way that substantially lowers confidence in the [contrast]" (Higgins et al., 2019).

This implementation guidance specifies that if 4/5 or 5/5 domains are judged to be 'some concerns' (and the remainder 'low risk'), this results in an overall judgement of 'high risk'. If 3/5 domains are judged to be 'some concerns' (and the remainder' low risk), then reviewer judgement is required as to whether this makes the overall risk of bias judgement 'some concerns' or 'high risk'; reviewers should consider the particular reasons why the domains have been given 'some concerns'. If 1/5 or 2/5

domains are judged as 'some concerns', and the rest 'low risk', then the overall judgement is 'some concerns'.

Note the particular RoB CRT algorithm for Domain 1: if 1a and/or 1b have a 'some concerns' rating, Domain 1 is 'some concerns'; if either 1a or 1b have a 'high risk' rating, then Domain 1 is 'high risk'.

Sources of guidance

For reference, this guidance includes its sources – either Cochrane's Revised Risk of Bias tool for cluster randomised trials (Eldridge et al., 20121; RoB CRT), Cochrane's Revised Risk of Bias tool for randomised trials (Sterneet al., 2019; RoB2), the Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures version 2.0 (Wilson et al., 2024; PSC), or previous Foundations Guidebook appraisal standards (Guidebook). Where there is a direct quotation from or reference to a section in the RoB2 full guidance, this is referenced (Higgins et al., 2019). Where a source is not indicated, the guidance has been formulated for this document in consultation with external experts.

Risk of bias arising from the randomisation process

1a.1 Was the allocation sequence random?

Following RoB CRT guidance, appropriate random assignment may be simple, or use block or stratification, or minimisation techniques (Sterne et al., 2019; RoB2), and is at the appropriate level (individual/family/group/centre) for the design and research questions (Guidebook).

To determine whether the level is appropriate, consider the types of effects the intervention is expected to produce: if the intervention is designed to improve outcomes for individual children, young people or families, then individual-level or family-level randomisation is typically appropriate. Cluster randomisation at a higher-level (group, Family Hub, Local Authority area, etc) can often also be appropriate here, depending on the implementation and delivery. This is typically the kind of intervention seen in appraising trials in children's social care and early intervention. However, if the effect of interest is at a higher level, for example an impact on a community such as crime rate, then randomisation should be at this higher, cluster level.

(NB if individual level randomisation, use the parallel RCT implementation guidance alongside RoB2 for individually-randomised, parallel-group trials (Sterne et al., 2019))

Examples of non-random allocation of clusters also include allocation decisions influenced by baseline assessments (e.g. where clusters such as Family Hubs with higher proportions of participants with higher needs demonstrated at baseline are assigned to the intervention condition), and methods based on local-area demographics (e.g. clusters with higher levels of need in the community allocated to intervtervention; in contrast, stratification based on demographics is acceptable).

Randomisation could also be compromised by reclassifying clusters who refused to participate in the intervention into the control group; assigning clusters to the intervention group to achieve a target sample number; or one site/cluster refusing randomisation but still being included in the study.

Randomisation is not compromised if clusters recruited after randomisation are not included in analysis; or if clusters or participants are excluded because it was discovered they did not meet the inclusion criteria, and this exclusion is applied to both intervention and control groups (Wilson et al., 2024; PSC 5.4). For example, imagine a trial of a group intervention being run in Family Hubs, which are randomised at the cluster level: some additional Family Hubs hear about the trial and ask to be included so they can run the intervention, after randomisation has taken place. They consent to be in the trial, but are not randomised but added to the intervention group directly. If these additionally recruited Family Hubs are *excluded* from the analysis, then randomisation is not compromised.

1a.2 Was the allocation sequence concealed until clusters were enrolled and assigned to interventions?

Following RoB CRT, randomisation allocation should be concealed until enrolment of clusters is complete. Clusters (e.g. directors of Children's Services in local authorities, managers of Family Hubs, lead practitioners of delivery partners), participants (if identified and recruited before cluster assignment), pracitioners, and researchers should be blinded to allocation until clusters are enrolled. In cluster trials, individual participants may be identified and/or recruited after cluster assignment, in which case, concealment of allocation for individuals may not be relevant to this question. In this scenario, it is unlikely that individuals would become aware of potential allocation before enrolment and assignment for clusters, but further risks this entails are covered in 1b.1–3 below.

1a.3 Did baseline differences between intervention groups suggest a problem with the randomisation process?

Following RoB CRT guidance, no additional calculations are required to judge baseline differences; it is acceptable for reviewers to look at the data for baseline differences and identify any that stand out as red flags indicating a problem with randomisation. That is, the point of this item is not to establish that there is baseline equivalence on key variables (see Domain 3), but to check that there are no differences which indicate that randomisation was compromised. Note that for this item, cluster numbers, cluster characteristics, and individuals' outcome variables at baseline can be considered.

RoB2 guidance references statistically significant differences in p-values (Sterne et al., 2019); it is acceptable to consider either p-values or effect sizes (for example, Cohen's *d*), and when both are available, effect sizes are strongly preferred, given that significance testing with p values with an inherently random sample does not make statistical sense. The thresholds for acceptable effect sizes of baseline difference outlined in Domain 3 can be used as guides for problematic differences here, but are not prescriptive; reviewer judgement is required.

For example, consider an evaluation of a parenting intervention designed to enhance support for families with a Child in Need or otherwise engaging with a social worker, in which parents in the intervention group are on average younger and more socioeconomically disadvantaged, and score more highly on social isolation and parenting stress than those in the control group, with consistent effect sizes of the difference of $d > 0.3$. This suggests that randomisation may have been compromised, for example by allocating groups in more disadvantaged areas to the intervention

condition. Depending on the description of the trial design and other details available, reviewers may judge that baseline differences did suggest a problem with randomisation here. However, if only one or two characteristics differ substantially at baseline, this is consistent with chance (see RoB2 guidance).

Risk of bias arising from the timing of identification or recruitment of participants in a cluster-randomised trial

The signalling questions in this domain are concerned with the risk of differential recruitment. When individuals are recruited into a study after the clusters of which they are a part have been randomised, it is possible for those recruiting individuals into the study to be aware of the outcome of randomisation and for this to influence recruitment, compromising randomisation. For example, if a local authority is randomised to receive an intervention, and practitioners then preferentially select participants that they believe will benefit from the intervention knowing that the cluster is in the intervention group, differential recruitment has occurred. Differential recruitment or identification of participants cannot happen if participants are identified or recruited before randomisation. If participants are recruited after randomisation, reviewers are required to assess whether it is likely that differential recruitment occurred based on descriptions of randomisation and recruitment processes in the study and appraisal of baseline demographic data for 'red flags' indicating differential recruitment.

1b.1 Were all the individual participants identified and recruited (if appropriate) before randomisation of clusters?

This could happen if participants are actively recruited in the trial but this happens before the cluster level is randomised, or if participants are not actively recruited in the trial but they are identified before randomisation takes place.

1b.2 Is it likely that selection of individual participants was affected by knowledge of the intervention assigned to the cluster?

As it is difficult for a reviewer to judge whether knowledge of intervention assignment is likely to have influenced selection in any given case, answer Y/PY if it is possible that knowledge of intervention assignment could have influenced selection.

Only answer N/PN to this signalling question if all of the following conditions apply:

- Recruiters do not know the cluster allocation (whether participants will receive the intervention or not).
- Participants do not know the cluster allocation before recruitment (e.g. they don't know if a centre offers the intervention being trialled).
- People identifying potential participants for future recruitment do not know the cluster allocation. (RoB2c 1b.2)

In short, the following must not know the cluster allocation:

- Those identifying participants
- Those identifying potential participants

- Recruiters
- Potential participants themselves. (RoB CRT 1b.2)

For example, in an intervention to support parents where a child has a Child in Need status, local authorities are randomised to intervention or control group at the start of the trial period. Families are recruited to the intervention on a rolling basis, as children receive a Child in Need status and are referred. Practitioners across the local authority know they are in a trial and know which arm they are in. This is likely to affect the selection of individual participants into the trial.

1b.3 Were there baseline imbalances that suggest differential identification or recruitment of individual participants between intervention groups?

This signalling question is concerned with 'red flags' and does not require baseline equivalence to be demonstrated for all measures; baseline imbalances suggest differential identification or recruitment if:

- There are significant and consistent differences in the numbers of participants recruited to each condition (differing from the ration specified in the trial protocol), indicating preferential recruitment to either the control or intervention condition
- There are significant and consistent imbalances in baseline demographics or pretest measures between the conditions, indicating intervention group recruitment may have been targeted at a subset of the population in a way that differs from control group recruitment.

In the example in 1b.2, randomisation at the cluster level had a 1:1 ratio with randomisation stratified by local authority size. Approximately the same number of families would therefore be expected at baseline in each arm. However, 540 families were recruited to the intervention in the intervention condition, and 410 in the control condition. This suggests a difference in recruitment to the sample between conditions.

Risk of bias due to deviations from the intended interventions (effect of assignment to intervention)

2.1a Were participants aware that they were in a trial?

No additional guidance needed.

2.1b Were participants aware of their assigned intervention during the trial? AND

2.2 Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?

Following RoB CRT participants (typically children, young people, and their families), carers, and people delivering the interventions (typically practitioners, such as family workers, social workers, or therapists) are all blinded to the intervention groups or conditions during the trial (or unaware they are in a trial), the contrast is considered at low risk of bias for Part 1 of this domain. Note that this is highly unlikely in early intervention and children's social care trials, because in social interventions both participants and practitioners tend to know which intervention they are

receiving and delivering. This means that a full consideration of deviations arising from the trial context is necessary (see 2.3).

2.3 Were there deviations from the intended intervention that arose because of the trial context?

Slightly different from RoB2 guidance, reviewers should look for explicit mention of deviation, or clues that deviation may have occurred. This is a slightly relaxed interpretation of the RoB2 guidance and algorithm, because it is unlikely that a study report will explicitly rule out all types of deviation if it did not occur; following the RoB2 guidance explicitly, where no information is given an NI judgement and leads to 'some concerns', could lead to unfairly penalising a study overall.

If there is nothing to suggest that deviations occurred, then it may be concluded that no deviations occurred and a PN/N judgement can be given, but reviewer judgement is required.

Such deviations include crossover, practitioner changes to the protocol and imbalance in non-protocol interventions (RoB2), but also diffusion, compensatory rivalry, and resentful demoralisation:

5. **Crossover:** control group participants receive the services the intervention group participants are assigned to (Sterne et al., 2019).
6. **Diffusion:** groups interact with each other either directly or indirectly, such that the nature of the treatment becomes known to the control group. This risks the control group sharing in the benefits of the treatment, and so contaminating the control group and biasing estimates of intervention effect; for example if intervention participants are residential care workers, and outcomes are measured in key index children, diffusion may occur if control group children share the same care home and either directly or indirectly interact with the intervention group residential care worker.
7. **Compensatory rivalry:** where the comparison group becomes aware that they are being denied an advantage or desirable intervention, and as a consequence, they increase their efforts to succeed and achieve positive outcomes; this could also include seeking alternative interventions.
8. **Resentful demoralisation:** where the comparison group become aware that they are being denied an advantage or desirable treatment, as a consequence, they become demoralised, discouraged, and/or angry and give up (Guidebook).
9. **Practitioner changes to the protocol**, for example due to their own knowledge or concerns about implementation.
10. **Imbalance in non-protocol interventions across groups**, for example if individuals in the intervention group, through attendance at a centre for intervention participation, additionally become aware of and access other support services to a greater extent than individuals in the control group, this would introduce bias (Sterne et al., 2019).

Look for explicit mention of deviation, or clues that deviation may have occurred. It is unlikely that a study report will explicitly rule out all types of deviation if it did not occur. If there is nothing to suggest that deviations occurred, then it may be concluded that no deviations occurred, but reviewer judgement is required. Note this is a relaxed interpretation of the RoB2 algorithm, where 'NI' leads to 'some concerns'.

2.4 Were these deviations likely to have affected the outcome?

Following RoB CRT, if there are identified deviations, consider whether these are likely to have an effect on the outcome. Examples where the outcome may be affected include:

- If a practitioner chooses to supplement an intervention with their favoured treatment approach that has not been pre-specified (e.g. a family worker adds motivational interviewing to a manualised parenting intervention which did not include this)
- If participants in a control group seek out, or are directed to, additional support that they would not have been likely to consider without knowledge of the intervention (e.g. through being recruited to a trial at a Family Hub, parents are specifically referred to an alternative BAU intervention).

2.5 Were these deviations from intended intervention balanced between groups?

Following RoB CRT, deviations are considered balanced between groups if the same deviation or deviations happen in both groups, and are judged to have affected the same proportion of participants in each group. This signalling question requires reviewer judgement.

For example, if family workers add motivational interviewing to their support of families in both the intervention and control group, this deviation would be balanced across groups; if it is added to the intervention group only, it would not be balanced – in that case, the effect being observed is of the intervention plus motivational interviewing, and this would need to be taken into account when deciding whether to include the study in the review.

Note that if it is not clear from the paper that any deviations are balanced between groups, a NI judgement is given, and is treated in line with an N or PN judgement, according to the RoB CRT algorithm.

2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?

Following RoB CRT, an 'intent-to-treat' design is used, meaning that all clusters and participants recruited to the intervention and control arms participate in the pre–post measurement and analysis, regardless of whether or how much of the intervention they receive, even if they drop out of the intervention (this does not include dropping out of the study – which is then regarded as missing data) (Guidebook).

Some studies may use 'modified ITT / mITT', which is the same as ITT with complete case analysis or listwise deletion; this is also acceptable (Sterne et al., 2019; RoB2).

If the design is not described as 'intention-to-treat' in the study report, but it is clear from the CONSORT diagram or other reporting of the flow of participants through the trial that ITT is the approach used, the reviewer can also respond Y/PY.

2.7 Was there potential for a substantial impact (on the [contrast]) of the failure to analyse participants in the group to which they were randomised?

As it is often not possible to specify what would constitute a 'substantial impact', and in children's social care and early intervention trials, reassignment might often be linked to prognostic factors, if an ITT or mITT analysis has not been used a Y/PY/NI judgement is appropriate for this criterion, leading to a 'high risk' judgement.

Some research papers report both ITT and 'per protocol' analyses; in this case it may be possible to judge the degree of impact of the alternative analysis method, however in these studies the results of the ITT analysis should be appraised for risk of bias and the per protocol results discounted.

Risk of bias due to deviations from the intended interventions (effect of adhering to intervention)

Do not use this domain; all studies should be appraised for the effect of assignment to intervention version of Domain 2 on RoB2.

Risk of bias due to missing outcome data

3.1a Were data for this outcome available for all clusters that recruited participants?

This implementation guidance for Domain 3 operationalises the RoB2 principle that "the number of participants with missing outcome data is sufficiently small that their outcomes, whatever they were, could have made no important difference to the estimated effect of intervention". In children's social care and early intervention trials, there is often relatively high overall attrition, because of the challenging context interventions are implemented in, and the characteristics of the vulnerable children and families receiving interventions. Furthermore, attrition is typically linked to children and families' characteristics and circumstances, which are also predictive of the outcome.

Therefore, a more holistic approach is taken, looking at both overall and differential attrition (3.1) and its effects (3.2). The clear sliding thresholds, taken from Prevention Services Clearinghouse / What Works Clearinghouse, are used in order to facilitate consistent reviewer judgement.

For cluster RCTs, both attrition of clusters and attrition of individuals are considered.

For 3.1a, if levels of attrition of clusters are within the acceptable limits described by Prevention Services Clearinghouse (Wilson et al., 2024), described below, answer 'Y'.

If either differential or overall attrition is not reported at cluster level, answer 'NI' for signalling question 3.1a.

3.1b Were data for this outcome available for all, or nearly all, participants within clusters?

If levels of attrition of individuals are within the acceptable limits described by Prevention Services Clearinghouse (Wilson et al., 2024), described below, answer 'Y' for signalling question 3.1a.

If either differential or overall attrition is not reported at an individual level, answer 'NI' for signalling question 3.1b.

3.1a and 3.1b attrition guidance

Attrition is calculated from randomisation; randomisation of clusters to conditions is considered to have occurred once cluster trial staff learn their assignment condition.

For cluster RCTs, both attrition of clusters and attrition of individuals are considered; individual level attrition is calculated only in non-attrited clusters (see formulae below (Wilson et al., 2024; PSC p. 64)).

Where the information is not available to calculate individual level attrition only in non-attrited clusters, attrition may be calculated using the whole sample size at randomisation. This is likely to be necessary in studies in which clustering occurs at the family level.

Figure 6. Calculation of overall and differential attrition in cluster trials (from *Prevention Sciences Clearinghouse Handbook v2* 5.6.2 p. 64) (link to long descriptive text)

$$Overall\ Attrition = \frac{N\ of\ individuals\ without\ posttest\ outcome\ data}{N\ of\ individuals\ in\ nonattrited\ clusters\ at\ randomization}$$

$$Differential\ Attrition = \left| \left(\frac{N\ of\ intervention\ condition\ members\ without\ posttest\ outcome\ data}{N\ of\ intervention\ condition\ members\ in\ nonattrited\ clusters\ at\ randomization} \right) - \left(\frac{N\ of\ comparison\ condition\ members\ without\ posttest\ outcome\ data}{N\ of\ comparison\ condition\ members\ in\ nonattrited\ clusters\ at\ randomization} \right) \right|$$

Overall and differential attrition for both clusters and individuals are within acceptable limits if the combination falls within the green area on the figure below (from the What Works Clearinghouse (WWC)); see the table following the figure for maximum acceptable differential attrition rates dependent on overall attrition rate (Wilson et al., 2024; PSC).

Overall attrition is attrition across all participants, from randomisation to when the outcome being appraised was measured. Differential attrition is the difference between attrition in the treatment group and attrition in the control group.

For example, if a trial started with 100 participants, with 50 in each group, and included 80 participants at endline, then overall attrition would be 20%. In the treatment group, 5 participants dropped out of the study, meaning 10% attrition, while in the control group, 15 participants dropped out of the study, meaning 30% attrition. Differential attrition is therefore also 20%. This places attrition in the 'red zone', over the acceptable limits, and so the reviewer must answer N, and proceed to 3.2.

In another case, a trial started with 100 participants, with 50 in each group, and included 80 participants at endline, so overall attrition is again 20%. This time, 11 participants in the treatment group drop out of the study (22% attrition), and 9 participants in the control group (18% attrition), so differential attrition is 4%. In this case, attrition is in the 'green zone', within acceptable limits, and so the reviewer can answer Y. Following the RoB2 algorithm, this means that the whole of Domain 3 is at low risk of bias.

Figure 7. Potential bias associated with overall and differential attrition (Exhibit 5.3. from the *Prevention Services Clearinghouse Handbook v2*) (link to long descriptive text)

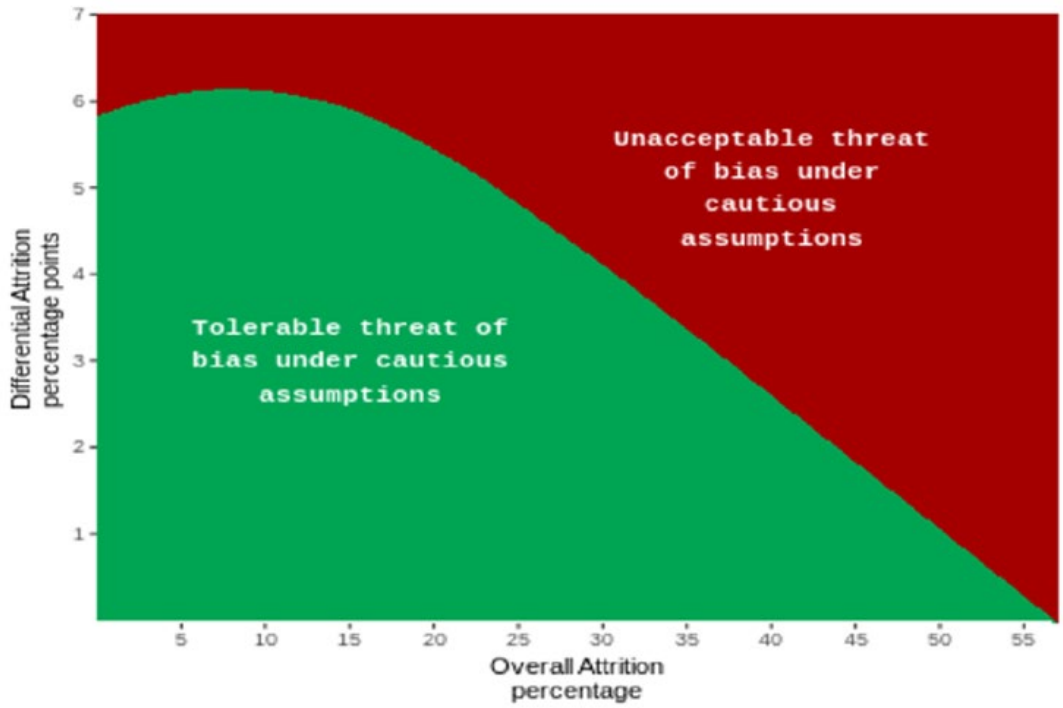


Exhibit 5.3 from Prevention Services Clearinghouse Handbook v2 p60

Table 5. Attrition Boundaries (Exhibit 5.4. from *Prevention Services Clearinghouse Handbook v2*)

Exhibit 5.4. Prevention Services Clearinghouse Attrition Boundaries

Overall Attrition	Differential Attrition	Overall Attrition	Differential Attrition	Overall Attrition	Differential Attrition
0	5.7	20	5.4	40	2.6
1	5.8	21	5.3	41	2.5
2	5.9	22	5.2	42	2.3
3	5.9	23	5.1	43	2.1
4	6.0	24	4.9	44	2.0
5	6.1	25	4.8	45	1.8
6	6.2	26	4.7	46	1.6
7	6.3	27	4.5	47	1.5
8	6.3	28	4.4	48	1.3
9	6.3	29	4.3	49	1.2
10	6.3	30	4.1	50	1.0
11	6.2	31	4.0	51	0.9
12	6.2	32	3.8	52	0.7
13	6.1	33	3.6	53	0.6
14	6.0	34	3.5	54	0.4
15	5.9	35	3.3	55	0.3
16	5.9	36	3.2	56	0.2
17	5.8	37	3.1	57	0.0
18	5.7	38	2.9		
19	5.5	39	2.8		

Source. What Works Clearinghouse (n.d.)

Note. Overall attrition rates are given as percentages. Differential attrition rates are given as percentage point differences. Attrition computations are rounded to whole numbers for determining overall attrition and to the nearest hundredth for differential attrition. For example, an overall attrition rate of 15.4% and differential attrition rate of 5.894pp would be rounded to 15% and 5.89pp, respectively. This contrast would be evaluated as low attrition because 5.89pp falls below the boundary of 5.9pp.

3.2 Is there evidence that the result was not biased by missing outcome data?

As explained in 3.1 above, guidance for 3.2 operationalises the RoB CRT signalling question with clear thresholds to enable judgement of when the result is or is not likely to be biased by attrition.

Step 1: Does the outcome measure at baseline (or an appropriate proxy, see below) demonstrate baseline equivalence in the analytic sample (d <0.05)?

Baseline equivalence of a measure is defined as a difference of d<0.05. When examining baseline equivalence, we are interested in the analytic sample; that is, the final sample used in analysis to generate the reported results, as opposed to the randomised sample.

To establish baseline equivalence, the following can be used [in order of preference]:

1. **Direct pretest.** Defined as the same (or nearly the same) measure used for the outcome. Note that even if problematic attrition is at the cluster level, if the outcome measure in question is at the individual level this is the preferred measure.
2. **Correlated pretest.** Defined as a measure in any eligible outcome domain (PSC 4.1.8) that has a correlation of 0.60 or higher with the outcome in the analytic sample. (A

correlation shown in the comparison condition only is also acceptable). Correlated pretests do not have to be in the same or similar domain as the outcome.

3. **Pretest alternative.** Defined as a measure in the same or similar domain as the outcome. No correlation threshold is specified – pretest alternatives are generally assumed to be correlated with the outcome by virtue of being conceptually related or are common precursors to the outcome [there are conditions under which pretest alternatives are acceptable].
4. **Sociodemographic characteristics.** Under certain conditions specified below, combinations of individual or individual and neighbourhood sociodemographic characteristics may be used to establish baseline equivalence. Eligible individual and neighbourhood characteristics are summarised below.
    a. **Individual sociodemographic characteristics.** Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, (3) Household composition, and (4) Age of sample members.
    b. **Neighbourhood sociodemographic characteristics.** Neighbourhood is defined as an Indices of Deprivation Lower-layer Super Output Area (LSOA); postcode; ONS census Output Area; or other small geographical area. In US, this is defined as a census tract, ZIP Code, or smaller geographic unit (or similarly sized tabulation unit for studies conducted outside of the United States). Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, and (3) Household composition (Wilson et al 2024; PSC pp. 65–66).

       Note that measures must meet all measurement standard requirements, i.e. reliability, face validity, and consistency, and must be completed before the start of the intervention.

Effect sizes can be calculated via the Effect Size Calculator – Campbell Collaboration if required.

If the effect size of the difference at pretest (or alternative) is d<0.05, answer Y/PY to signalling question 3.2(a/b). If the effect size of the difference at pretest (or alternative) is d > 0.05, proceed to step 2.

Step 2: If the effect size of the difference at pretest is moderate (d > 0.05 but < 025), then it must be controlled for in analyses, or there must be sensitivity analyses showing that there is no need to control for the difference in the analysis; if it is controlled for in analyses answer Y/PY to signalling question 3.2(a/b). If the effect size is large, not controlled for in analyses or there are no sensitivity analyses, proceed to step 3.

Step 3: If the effect size of the difference at pretest is large (d > 0.25), or is moderate and not controlled for in analyses, answer N/PN to signalling question 3.2(a/b). This is evidence that the result was biased by the missing outcome data.

Example 1: In an RCT of a group parenting intervention, differential attrition puts the trial in the 'red area', with attrition over the acceptable limit. The primary outcome measure is a measure of children's behaviour, and this is measured at baseline and after the intervention. This is therefore the preferred way of establishing whether there is baseline equivalence between the groups. The effect size of the difference between the groups is calculated for the *analytic sample* at baseline,

and found to be d = 0.15. This means that there is *not* baseline equivalence, but it is possible to address an effect of this size in analysis. The outcome measure at baseline is included in the regression analysis, and therefore the reviewer can respond Y.

Example 2: In a different trial, attrition is again over the acceptable threshold, and there is no outcome measurement at baseline, because it is a perinatal intervention, with a child outcome. There is also no pretest alternative, so demographic characteristics are considered. Demographic characteristics in the analytic sample have an effect size of d = 0.15, but they are not included in the analysis, and there are no additional sensitivity analyses available. Therefore the reviewer must respond N.

3.3 Could missingness in the outcome depend on its true value?

Following RoB CRT, if missingness (attrition) is not within acceptable limits described in 3.1, and there is not evidence that the result is not biased (3.2a and b), Domain 3 can still be considered low risk as a whole if there is evidence that data is missing completely at random (MCAR), for example if data is missing as a result of data file corruption that was not in any way related to the outcome. It is relatively rare and highly unlikely that missing data is unrelated to the outcome; in most cases this signalling question should be answered Y/PY.

If there is a clear explanation in the research paper to justify why outcome data can confidently be considered MCAR, this signalling question can be answered N/PN (Sterne et al., 2019).

3.4 Is it likely that missingness in the outcome depended on its true value?

In children's social care and early intervention research, if data is not missing completely at random (see 3.3), it is reasonable to assume that it is "likely that missingness in the outcome depended on the true value" (Higgins et al., 2019), that is, that the reasons behind participants dropping out of the study are linked in some way to the outcome of interest. Following this implementation guidance, this signalling question is only considered for studies in which there are high levels of missing data and a significant difference in the outcome measure at baseline (or an appropriate proxy) which is too large to be controlled for in analyses, and data is not MCAR. Always answer Y/PY to this signalling question.

Risk of bias in measurement of the outcome

4.1 Was the method of measuring the outcome inappropriate?

In children's social care and early intervention trials, adapted or bespoke measures developed for the intervention are common. It is therefore particularly important to check that valid and reliable measures have been chosen appropriately. This implementation guidance for signalling question 4.1 provides four characteristics of the outcome measure, which the outcome measure must all meet, to determine that the method of measuring the outcome was not inappropriate. These characteristics are based on previous Guidebook and Prevention Services Clearinghouses' requirements for measures.

1. The measurement is independent of the treatment

The measure is not used as part of the intervention (for example, by the practitioner to monitor and determine the content of intervention sessions); this is sometimes known as a 'treatment inherent' measure. (Guidebook)

If there is no information available, consider this criterion met.

2. An appropriate measure was used, for the population

The measure is appropriate for the vast majority of the sample population in terms of age, level of need, and country, ethnicity, culture, and language. Reviewer judgement is required to decide which of these factors are most important to demonstrate for the measure used. It has been designed for and tested or normed with a similar sample to that in the study. (Guidebook)

For example, if a measure designed to assess externalising behaviour in 5–10-year-olds is used with 2–4-year-olds, the method of measuring the outcome is inappropriate; similarly, a measure developed for the general population may not be appropriate for a sample who has experienced complex trauma, unless the measure had been separately tested with the higher needs population.

If there is no information available, consider this criterion met.

3. The measure has high face validity

To satisfy the criterion for face validity, there must be a **sufficient description** of the outcome or baseline measure for the **reviewer to determine** that the measure is clearly defined, has a direct interpretation, and appears to measure the construct it was designed to measure. (Wilson et al., 2024; p. 73) In other words, at face value, the measure does what it says it does (e.g. does a measure said to measure anxiety actually measure anxiety, rather than another construct?)

This criterion requires reviewer judgement based on information in the paper or easily available; if there is no information available, consider this criterion NOT met.

4. The measure has high reliability

The outcome or baseline measure either must be a measure which is assumed to be reliable (see below) or must meet one or more of the following standards for reliability (depending on what is appropriate for the type of measure):

- Internal consistency (such as Cronbach's alpha) of 0.50 or higher
- Test–retest reliability of 0.40 or higher
- Interrater reliability (correlation) of 0.50 or higher
- Interrater agreement (percentage agreement or kappa) of 0.80 or higher for percentage agreement and 0.60 or higher for kappa.

When required, reliability statistics may be based on an independent sample (with similar characteristics to the study sample) and/or the study sample. An independent sample is strongly preferred, as this avoids the risk of a novel tool being overfitted on the study sample; however, a demonstration of reliability from the study sample may be allowed with careful reviewer judgement. For example, where a measurement tool is already well established with extensive

validation and standardisation, and a cultural or linguistic adaptation is developed and used in the study in question, with reliability demonstrated in the study sample, this may be allowed with reviewer judgement. If there is no information in either the study under review or additional sources and the measure is not assumed to be reliable (see below), consider the criterion NOT met.

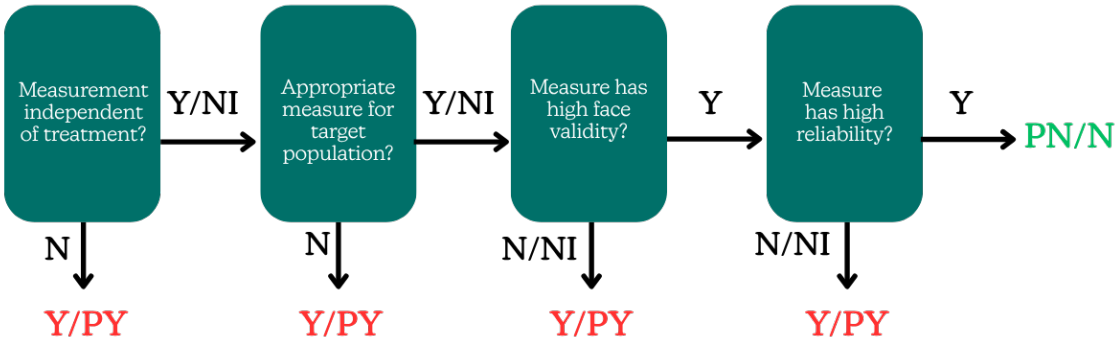
Some types of measures are normally assumed to be reliable, including:

- Demographic characteristics, such as age, race/ethnicity, education level, SES, employment status, etc.
- Medical or physical tests, such as urinalysis, weight measurement, etc.

Administrative data can often be assumed to be reliable, but can vary substantially in quality and so reviewer judgement is required to decide whether this criterion is met; shortcomings of administrative data may be described in the study, or information about the dataset can be checked. Administrative data includes records obtained from schools, child welfare or other social service agencies, hospitals, or clinics.

If all four characteristics are met, then a judgement of PN/N can be given; if at least one is not met, then a judgement of Y/PY is given.

Figure 8. Flowchart illustrating steps to judge whether the measure is not inappropriate; answers to questions lead to RoB2 judgement for signalling question 4.1 (link to long descriptive text)



4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?

Following RoB CRT, to determine whether measurement or ascertainment of the outcome could have differed between intervention groups, consider whether:

1. There is equivalent measurement of groups in terms of timing

The time between pretest (baseline) and posttest (outcome) does not systematically differ between intervention and comparison conditions. If there is a systematic difference in timing between

groups, this criterion is not met; for example if the treatment group received the posttest after 12 weeks, while the control group received it after 8 weeks (Wilson et al., 2024; PSC p. 74).

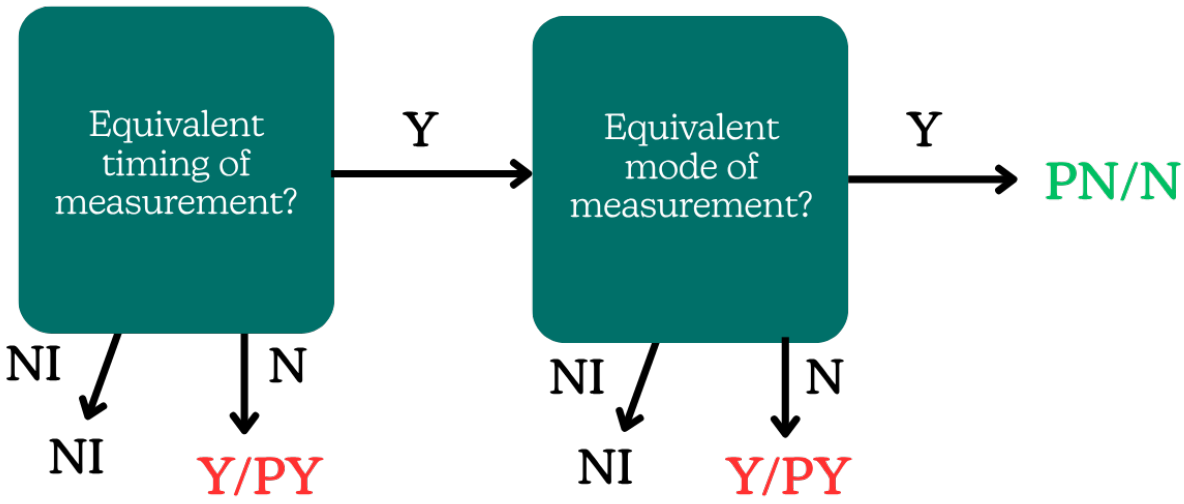
2. There is equivalent measurement of groups in terms of mode

The data collectors and data collection modes for data collected from intervention and comparison conditions either are the same, or are different in ways that would not be expected to have an effect on the measures. (Wilson et al., 2024; PSC p.74).

For example, one group being measured through a face-to-face interview and the other group through an online survey would not be acceptable; similarly if systematically different types of practitioners or staff administered a questionnaire or interview between groups, and the type of staff is likely to have affected participants' responses.

If a measure is equivalent in timing and mode across groups, respond PN/N; if it is not equivalent on one or both characteristics, respond PY/Y.

Figure 9. Flowchart illustrating how to establish whether measurement differed between groups; answers to the questions lead to RoB2 judgement for signalling question 4.2 (link to long descriptive text)



4.3a Were outcome assessors aware that a trial was taking place?

No additional guidance needed.

4.3b Were outcome assessors aware of the intervention received by study participants?

No additional guidance needed.

4.4 Could assessment of the outcome have been influenced by knowledge of intervention received?

Following RoB CRT, consider whether the type of measure means that an unblinded assessor could have influenced the outcome. For example, if the assessor is administering an objective cognitive test or questionnaire, without seeing or hearing the participants' responses (e.g. on a computer by giving the participant a pen-and-paper questionnaire to fill in themself), this is unlikely to have affected the outcome; if the assessor is leading and/or coding an interview with the participant, their awareness could influence the outcome.

Participant self-reports are common in children's social care and early intervention research, and for self-reports, always answer Y.

4.5 Is it likely that assessment of the outcome was influenced by knowledge of intervention received?

Following RoB CRT (with the exception of self-reports, see below), if outcome assessors are (or may be) aware of the intervention received, and it is possible that awareness could have biased the outcome, then judgement must be used to decide whether this is **likely** to have affected the participants' outcomes. As stated in the RoB2 guidance, this is more likely to be the case when there are strong levels of belief in either beneficial or harmful effects of the intervention (Wilson et al., 2024; PSC; Higgins et al., 2019; 4.5 p. 18), for example assessments of recovery by a practitioner who delivered the intervention.

Self-reports are potentially influenced by knowledge of the intervention received, because they are inherently 'unblinded', and judgement is needed to determine whether it is likely that participants' reporting of the outcome was influenced by knowledge of the intervention received. If the measure is valid and reliable, benefit of the doubt can often be given. This is a slightly relaxed implementation of RoB2, as self-reports are both widely used in children's social care research, and, being widely used, often have good psychometric properties (are valid and reliable).

Risk of bias in selection of reported result

5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalised before unblinded outcome data were available for analysis?

No additional guidance needed.

Is the numerical result being assessed likely to have been selected, on the basis of the results, from ...

5.2 ... multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain?

No additional guidance needed.

5.3 … multiple eligible analyses of the data?

No additional guidance needed.

Additional criteria for assessment of randomised controlled trial for guidebook assessment

This short Guidebook 'add-on' set of standards covers important risk of bias and other characteristics of trials not covered in Cochrane's tools.

Why are additional criteria needed?

These criteria are required because the Guidebook is based on rating individual studies, not pooling of many effect sizes in a meta-analysis or synthesising the weight of evidence in a systematic review. There are often few eligible studies in this area of research, and an evidence rating for an intervention is frequently based on just one or two studies. Therefore, additional assurance is required about the quality of the sample and the analysis in order to have confidence in the effects observed and results reported.

When are the additional criteria used?

Studies selected for Guidebook appraisal will initially be assessed using Cochrane's Risk of Bias 2 tool, using Foundations' implementation guidance to support application of the tool to children's social care or early intervention trials. Studies with contrasts receiving a 'low risk' or 'some concerns' overall RoB2 rating are then assessed against the following additional criteria. Contrasts which receive either a low risk or some concerns RoB2 rating and meet **all** of the additional criteria are awarded Level 3; the intervention is considered to have a promising evidence base.

As for RoB2, contrasts are appraised individually. A contrast is a single outcome or result from a study – see the Implementation Guidance for RoB2 for more information on assessing contrasts.

1. If randomisation is compromised or unclear, criterion 3 (below) must be met, regardless of attrition rates.

RoB2 allows a study to be awarded 'some concerns' in the absence of appropriate random allocation, or if allocation sequence was random but baseline imbalances suggest a problem, provided the allocation sequence has been concealed.

To achieve a Level 3 rating on the Guidebook the allocation sequence must be random and not compromised; if randomisation is compromised or there is no information regarding the nature of randomisation (an N or NI for RoB2 signalling question 1.1), criterion 3 must be met either by baseline equivalence being demonstrated or baseline differences (if moderate) being controlled for, regardless of the level of attrition (Guidebook).

2. Randomisation is concealed until enrolment. If concealment is unclear, criterion 3 (below) must be met, regardless of attrition rates.

RoB2 allows a study to be awarded 'some concerns' in the absence of clear information regarding allocation sequence concealment, provided there are no baseline imbalances suggesing a problem, or if there is no information about both allocation sequence concealment and baseline imbalances.

To meet Level 3 for the Guidebook, in the absence of clear evidence of randomisation concealment (an NI for RoB2 signalling question 1.2), criterion 3 must be met either by baseline equivalence being demonstrated or baseline differences (if moderate) being controlled for, regardless of the level of attrition (Guidebook).

3. Baseline equivalence is demonstrated if required by criteria 1 or 2 (above) or the difference is moderate and is controlled for in analysis.

If randomisation is compromised or unclear, or not concealed until enrolment, then baseline equivalence must be demonstrated in the analytic sample, or the difference must be moderate and controlled for in analysis (Guidebook, following PSC, Wilson et al., 2024), or there must be sensitivity analyses which show the difference does not affect the outcome (RoB2).

Step 1: Does the outcome measure at baseline (or an appropriate proxy, see below) demonstrate baseline equivalence in the analytic sample (d <0.05)?

Baseline equivalence of a measure is defined as a difference of d<0.05. When examining baseline equivalence, we are interested in the analytic sample; that is, the final sample used in analysis to generate the reported results, as opposed to the randomised sample.

To establish baseline equivalence, the following can be used [in order of preference]:

1. **Direct pretest.** Defined as the same (or nearly the same) measure used for the outcome.
2. **Correlated pretest.** Defined as a measure in any eligible outcome domain (Section 4.1.8) that has a correlation of 0.60 or higher with the outcome in the analytic sample. (A correlation shown in the comparison condition only is also acceptable). Correlated pretests do not have to be in the same or similar domain as the outcome.
3. **Pretest alternative.** Defined as a measure in the same or similar domain as the outcome. No correlation threshold is specified – pretest alternatives are generally assumed to be correlated with the outcome by virtue of being conceptually related or are common precursors to the outcome [there are conditions under which pretest alternatives are acceptable].
4. **Sociodemographic characteristics.** Under certain conditions specified below, combinations of individual or individual and neighbourhood sociodemographic characteristics may be used to establish baseline equivalence. Eligible individual and neighbourhood characteristics are summarised below.
    a. **Individual sociodemographic characteristics.** Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, (3) Household composition, and (4) Age of sample members.
    b. **Neighbourhood sociodemographic characteristics.** Neighbourhood is defined as an Indices of Deprivation Lower-layer Super Output Area (LSOA);

> postcode; ONS census Output Area; or other small geographical area. In US, this is defined as a census tract, ZIP Code, or smaller geographic unit (or similarly sized tabulation unit for studies conducted outside of the United States). Eligible measures are: (1) Race or ethnicity, (2) Socioeconomic status, and (3) Household composition (PSC, Wilson et al., 2024, pp. 65–66).

Note that measures must meet all measurement standard requirements, i.e. reliability, face validity, and consistency, and must be completed before the start of the intervention. If the outcome measure or proxy has a baseline difference of d < 0.05 in the analytic sample, this criterion is met.

Step 2: If the effect size of the difference at pretest is moderate, is it controlled for in analyses, or are there sensitivity analyses showing that there is no need to control for the difference?

A moderate effect size is defined as d > 0.05 but < 0.25. If the outcome measure or proxy has a baseline difference of d > 005 and d < 0.25, and this is controlled for in the analysis, or demonstrated through sensitivity analyses to not have an effect, this criterion is met.

Step 3: Does the outcome measure at baseline have an effect size of d > 0.25?
If the baseline imbalance has an effect size of d > 0.25, this criterion is NOT met.

If no information is available to make a judgement, consider this criterion NOT met.

4. There are no systematic confounds.

A systematic confound is an aspect of a study that is always present for members of one group and never present for members of the other group (and is not part of the intervention). If a confounding factor is present, it is impossible to separate the effect of the confounding factor from the effect of the intervention.

For all trial types, this includes: n=1 person-provider confound: “When all individuals in the intervention condition or all individuals in the comparison condition receive intervention or comparison services from a single provider” (e.g. a single therapist or a single doctor) the treatment effect is confounded with the skills of the provider (PSC, Wilson et al., 2024).

For example, a group of parents were recruited from a school to a group parenting intervention, and a single therapist delivered the intervention to those randomised to the intervention condition, while they had no interaction with parents randomised to the control group; in this case, this criterion would not be met. In another case, a social worker delivers an intervention on top of business as usual to half their clients randomised to the intervention group, and continues business as usual with the other half; in this case, this criterion would be met (though other confounds like cross-over would need to be considered (RoB2 / RoB CRT 2.3).

For cluster trials, this also includes n=1 administrative unit provider confound:

> “When all individuals in the intervention condition or all individuals in the comparison condition receive intervention or comparison services in a single administrative unit (e.g. clinic, hospital, organisation, or community) the treatment effect may be confounded with the capacity of that administrative unit

> to produce better outcomes or with characteristics of the administrative unit not directly associated with delivery of the intervention" (PSC, Wilson et al., 2024).

Note that 'organisation' here does not include the delivery organisation, as this is often the sole or main provider of an intervention. For example, if all participants receiving an intervention are in one local authority or Family Hub, and those in the control group in another, this confound would be present (i.e. a cluster trial with only two clusters; this could likely be assessed using the quasi-experimental design standards); if all participants in the intervention group receive an intervention from one provider who does not provide any services to participants in the control group, but all other administrative units are shared across the intervention and control groups, this confound would NOT be present.

> "When individuals in the intervention condition receive services from a single administrative unit and individuals in the comparison condition receive no services from the administrative unit (e.g. a waitlist comparison condition that does not receive treatment as usual during the waiting period), this confound is present. However, if a single provider location serves at least some participants in both the intervention condition and the comparison condition, this confound is not present" (PSC, Wilson et al., 2024).

If there is a systematic confound of either type, this criterion is not met. If there is no information available to indicate the presence of a systematic confound, consider this criterion met.

5: Clustering is accounted for in the analysis of cluster randomised trials (cluster RCTs only)

The following situations are considered acceptable when randomisation is at the cluster level:

- **a.** The unit of analysis is also at the cluster level.
- **b.** The unit of analysis is at the level of individual participants, AND clustering is taken into account in the analysis, for example using hierarchical modelling, mixed-models with cluster as random effects, or cluster-robust standard errors with general linear modelling.

If there are a small number of clusters, extra care must be taken; for example, when using OLS (Ordinary Least Squares) regression models, then using heteroskedasticity-consistent 2 or 3 (HC2 or HC3) robust standard errors would be appropriate; with multilevel models, restricted maximum likelihood, Kenward-Roger adjustments or bootstrapping could be options.

If clustering is taken into account in the analysis, this criterion is met. If clustering is not taken into account in analyses, this criterion is not met.

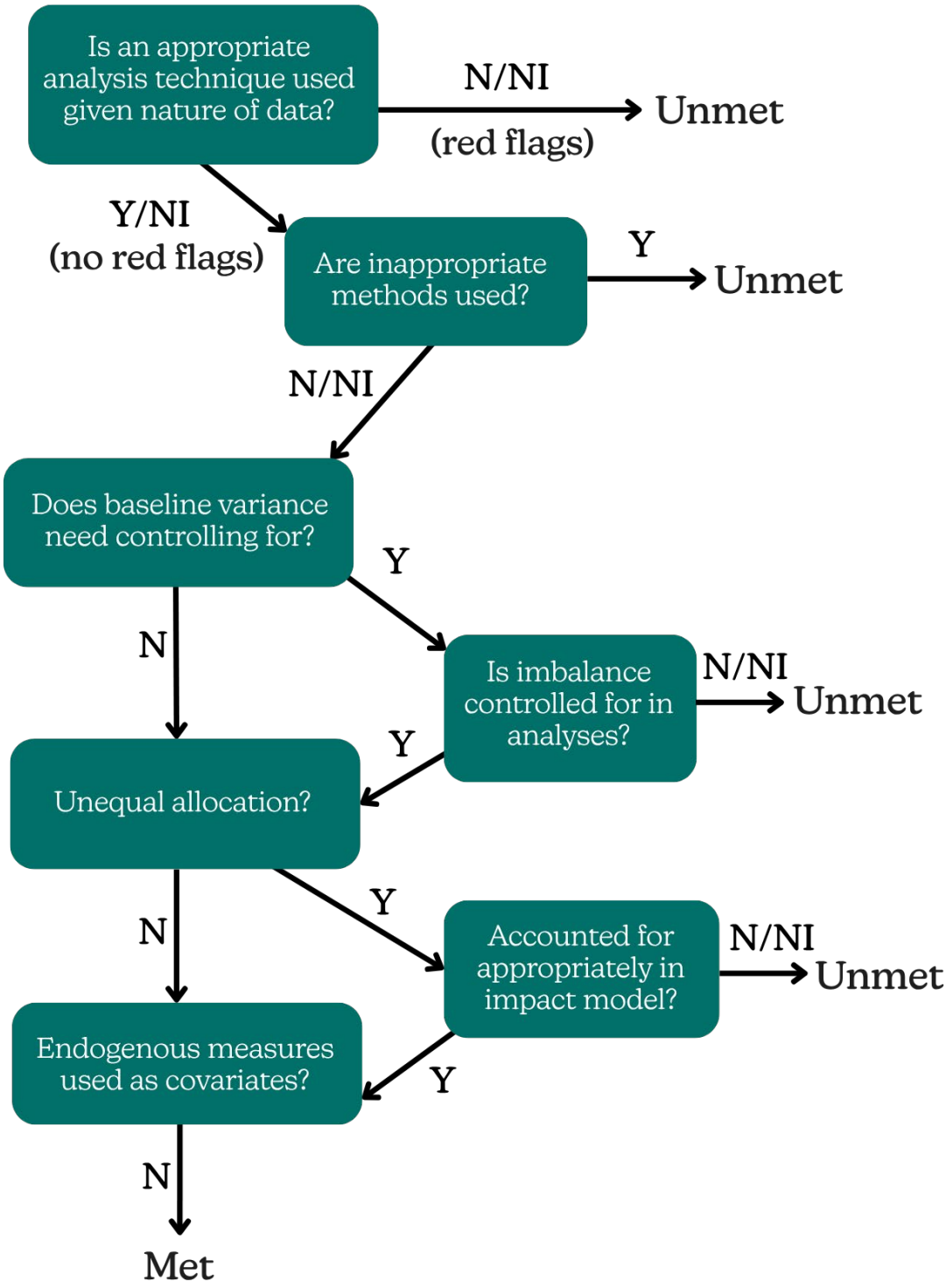
6: An appropriate statistical analysis is used given the data being analysed and the purpose of the analysis.

If multiple analyses have been run on the same outcome, select the analysis which satisfies the most criteria; e.g. if both ITT and non ITT analyses are reported, or analyses are run with and without required adjustments, apply the criteria to the most appropriate analysis and make a note

in the justification column. That is, the theoretical or statistical justification should underpin the selection of the analysis, not the outcomes of the analysis.

1. An appropriate statistical technique must be used, given the type and nature of the data being analysed (continuous, ordinal or categorical data; the distribution of the data). Benefit of the doubt may be given if information is missing in the reported study and there are not red flags that model assumptions are not met. *Note this is assessed for L2*.
2. No methods judged inappropriate can be used; methods which are not appropriate include separate group paired analysis (effectively two pre–post analyses), and one-tailed tests (unless negative effects are ruled out or you can calculate two-tailed p values). *Note this is assessed for L2*.
3. Where necessary (as determined by criterion 3), baseline variance is controlled for appropriately, using: regression models with baseline measures as covariates; gain score models where dependent variable in the regression is a difference score equal to the outcome minus the pretest; repeated measures ANOVAs; DiD models; models with fixed effects for individuals (PSC section 5.8).
4. If there is any unequal allocation of participants to groups across randomisation blocks, the impact model must account for this by either using dummy variables, reweighting observations or conducting separate impact analyses within each block and averaging the impacts (PSC section 5.9.1).
5. Impact models cannot include endogenous measures as covariates (one that is measured or obtained after baseline and that could have been influenced by the intervention) (PSC section 5.9.1).

Figure 10. Flowchart to show the steps to check that an appropriate analysis has been used and Criterion 6 is met (link to long descriptive text)



7: There is appropriate treatment of missing data.

Note that this additional criterion is in place until Cochrane update their missing data guidance; this criterion may then no longer be necessary.

Unless data is missing completely at random (MCAR), appropriate methods must be used to account for missing data in analyses.

Methods considered appropriate include:

- Listwise deletion/complete case analysis
- Multiple imputation, in which all covariates in the impact model and the outcome are included in the imputation
- Maximum likelihood estimation (PSC)
- Non-response weights (PSC/RoB2).

Methods considered inappropriate (unless there is clear justification) include:

- Last observation carried forward
- Other single imputation approaches (e.g. single mean imputation)
- Multiple imputation based only on intervention group (RoB2).

Note that this is separate to controlling for biases introduced through attrition.

If appropriate methods are used; this criterion is met; if there is no information available, consider this criterion met; if inappropriate methods are used, this criterion is not met.

7. Generating provisional evidence ratings at study and intervention level

Generation of a provisional study rating

The Guidebook evidence standards are applied to individual contrasts in a study, meaning that one study may have multiple contrasts receiving different evidence ratings. A study receives the same rating as its highest rated contrasts; contrasts within the study not achieving this rating are reported separately in the Guidebook entry, making it clear that the evidence underpinning those findings is at a lower level. Study ratings remain provisional until discussed in and confirmed by an independent external expert review.

Generation of a provisional intervention evidence rating

Interventions are awarded provisional intervention evidence ratings based on the highest-rated study evidence available.

Figure 11. Guidebook evidence standards summary (link to long descriptive text)

4 | EFFECTIVENESS

Evidence of a long-term positive impact from at least two rigorous evaluations. The evidence may include adaptations to meet the needs of different target populations.

3 | EFFICACY

Evidence of a short-term positive impact on a child outcome from at least one rigorous evaluation.

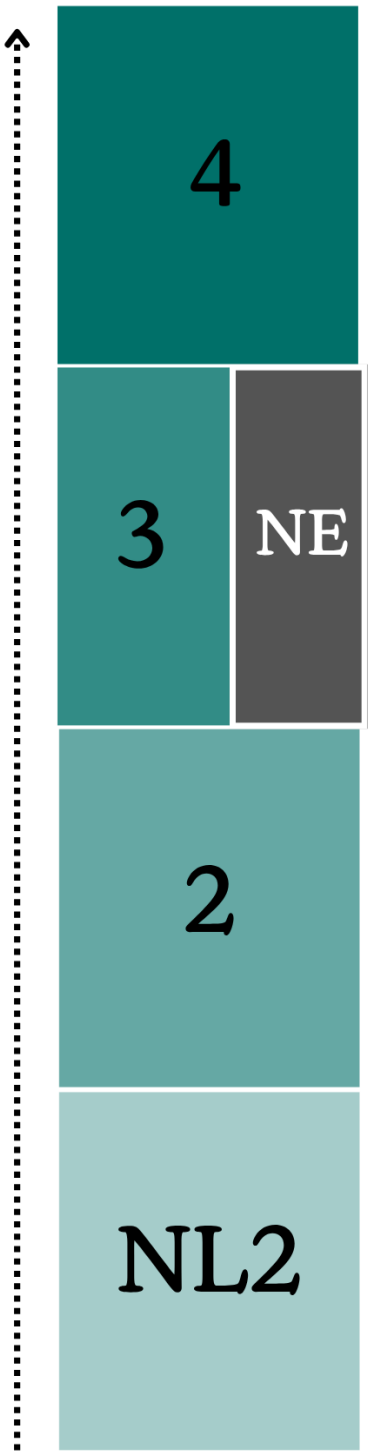
No Effect: The most rigorous evaluation, meeting Level 3 criteria, does not find evidence of improving any child outcomes.

2 | PRELIMINARY EVIDENCE

Evidence of improving a child outcome from an evaluation with at least 20 participants, using a valid measure; it's not possible to confirm that the intervention caused the improvement.

NL2 | DEVELOPING

Development of a Theory of Change and testing of impact may be underway, but evidence does not yet meet Level 2 criteria.



Level 4: Effectiveness

Level 4 recognises interventions with evidence of a long-term positive impact through multiple rigorous evaluation studies. At least one of these studies must have evidence of improving a child outcome lasting a year or longer. The evidence may include significant adaptations to meet the needs of different target populations.

The evidence must meet the following requirements:

- The intervention has demonstrated consistent significant positive child outcomes in two rigorous studies meeting all criteria required for Level 3.
- At least one study uses a form of measurement that is independent of the study participants (and also independent of those who deliver the intervention). In other words, self-reports (through the use of validated instruments) might be used, but there is also assessment information independent of the study participants (e.g. an independent observer, or administrative data).
- In at least one study there is evidence of a long-term outcome of 12 months or more.

Level 4+

To achieve a 4+ rating:

- All of the criteria for Level 4 must be met.
- At least one of the Level 4 studies will have been conducted independently of the programme developer.
- The intervention must have evidence of improving Foundations' child outcomes from three or more rigorously conducted studies meeting all criteria required for Level 3 and conducted within real-world settings.

Level 3: Efficacy

Level 3 recognises interventions with evidence of a short-term positive impact from at least one rigorous evaluation study – that is, where a judgement about causality can be made. The evaluation should demonstrate a statistically significant positive impact on at least one child outcome.

The evidence must meet the requirements for Level 2 and Level 3.

Level 3+

To achieve a 3+ rating:

- The intervention will have obtained evidence of a significant positive child outcome through a Level 3 efficacy study but also has additional consistent positive evidence from a Level 2 or Level 2+ study with a comparison group design (occurring under ideal circumstances or real-world settings).

No Effect (NE)

The intervention has a rigorous study, meeting Level 3 criteria, which is also the most rigorous impact evaluation, and which has not found evidence of improving a child outcome.

The evidence must meet the following requirements:

- The study must meet the requirements for Level 3.
- It fails to confirm any statistically significant benefits with respect to at least one Foundations child outcome.

This rating should not be interpreted to mean that the intervention will never work, but it does suggest that there may be key aspects of the intervention's logic model which require respecification and re-evaluation.

Level 2: Preliminary evidence

Level 2 recognises interventions with preliminary evidence of improving a child outcome, but where a conclusion of causal impact cannot be drawn.

The evidence must meet the requirements for Level 2.

Level 2+

To achieve a 2+ rating:

- There is a significant positive child outcome in an evaluation study meeting all the criteria for a Level 2 study but also involving a treatment and comparison group.
- There is baseline equivalence between the treatment and comparison-group participants, meeting the Level 3 baseline equivalence requirements.

Not Level 2 (NL2)

The intervention is judged to not meet the Level 2 threshold for a variety of methodological issues with the underpinning studies. The Guidebook currently does not include entries for interventions rated below Level 2.

Interventions falling into this category are typically at earlier stages of their development, with important foundational work being carried out. This might include developing a theory of change or logic model, or carrying out feasibility, implementation, or pilot evaluation studies.

Mixed findings

Sometimes, different studies underpinning an intervention's evidence rating may have mixed findings: that is, there are studies suggesting positive impact alongside studies that on balance indicate no effect or negative impact. These are indicated by an asterisk.

- **Level 3* rating:** If an intervention has strong evidence of impact from a single robust study at Level 3, but also has strong evidence of not having achieved impact from another robust study at Level 3, the intervention will receive a Level 3* rating.

- **Level 4* rating:** If an intervention has strong evidence of impact from multiple robust studies, meeting Level 4 requirements, but also has strong evidence of not having achieved impact from other robust studies, the intervention will receive a Level 4* rating.

8. Reviewing the evidence and deciding on a final rating

Following the decision on provisional evidence ratings for an intervention and its contributing studies by two Foundations evidence team members, a review or independent assessment by an external expert is obtained. The external experts appraise the studies against Foundations' standards and reach their own judgement on appropriate ratings for the studies and intervention. Any discrepancies in ratings between the independent expert and the internal Foundations team are discussed until consensus is reached.

9. Appeal and moderation

Providers then receive a final rating decision. This confirms the evidence rating agreed upon by the Foundations Guidebook team and the external expert(s). The provider also receives the technical appendix, which explains in detail the rating and outlines the reasons for including or excluding studies in that assessment round.

As set out in the Terms of Reference, providers may request a reassessment of their evidence rating following this rating decision if they believe that there are substantive grounds for doing so. Requests cannot be based on new studies that were not previously submitted, nor on broader concerns about the evidence standards themselves. However, requests are considered where a provider believes that the evidence standards, as described in the summary document, were misapplied to the studies included in the original assessment.

Requests for reassessment due to an appeal must be submitted through an online form. Providers can use this form to clearly explain how they believe the standards were misapplied or to highlight any important and relevant inaccuracies, in no more than 500 words.

Appeals are reviewed by a Foundations evidence team member who was not involved in the original assessment and an independent external expert reviewer. Together, they will consider the request and determine whether the rating should be revised. The outcome, along with an explanation of the decision, is then communicated to the provider.

10. Recording impact estimates

In the Guidebook, the impact of interventions rated Level 3 or higher is presented as an *improvement index*. This is a number between 0 and 50 that shows how much an intervention benefits children and families, expressed in a way that makes different interventions easy to compare.

Simply put, the improvement index shows how far a 'typical' child or family might move up relative to similar peers after participating in the intervention. Because it's based on percentile movement, it's sometimes referred to as 'percentile growth' or 'percentile rank improvement'.

The improvement index is related to a statistical effect size (such as Cohen's *d*). Effect sizes help researchers understand the magnitude of an intervention's impact, but they're often difficult for non-researchers to interpret because they rely on standard deviations and other statistical concepts. The improvement index addresses this by converting the effect size into a percentile-based number. This translation puts results from different studies (even those using different measures or scales) onto the same 0–50 scale.

More technically, the improvement index represents the difference between:

- The percentile rank corresponding to the mean value of the outcome for the intervention group
- The percentile rank corresponding to the mean value of the outcome for the comparison group distribution.

For example:

- An improvement index score of 25 means we would expect the average participant in the comparison group who did not receive the intervention (for whom 50% of their peers have better outcomes and 50% have worse outcomes) to improve to the point where they would have better outcomes than 75% of their peers, and worse outcomes than 25% of their peers, if they had received the intervention.
- An improvement index score of 50 means we would expect the average participant in the comparison group who did not receive the intervention to improve to the point where they would have better outcomes than 100% of their peers, and worse outcomes than 0% of their peers, if they had received the intervention. In other words, they would have the very best outcome relative to their peers.
- An improvement index score of 0 means that there is no improvement. The average participant in the comparison group who did not receive the intervention would maintain this ranking if they had received the intervention.

The improvement index is calculated by converting an effect size reported in the original study (such as Cohen's *d*, an odds ratio, or another standardised effect size) into percentile ranks. An effect size can be treated like a z-score, which is just a way of placing a number on a normal distribution (i.e. a standard bell-shaped curve). Once it is known where that number falls on the curve, it can be ascertained what percentile of people would fall below it. That percentile becomes the improvement index.

Should the effect size not be reported in the study, Foundations reviewers will calculate an effect size using data from the study and subsequently create an improvement index score.

If an improvement index is not shown, this is either because a) the study and intervention in question did not receive an evidence rating of Level 3 or above, or b) insufficient information is

reported in the original evaluation studies. Some assessment rounds historically also did not calculate the improvement index for outcomes.

In the full evidence descriptions of all studies, the effect sizes of child and parent outcomes are also presented in an outcomes table, as they are reported in the studies underpinning the evidence rating. These are reported using the unit of effect size used in the original studies (e.g. Cohen's *d*, Hedges' *g*, Odds Ratio or Risk Ratio).

Why does the Guidebook report impact in two different ways?

Effects as they were originally measured in the study tell us something useful about the nature of the improvement that an intervention has generated: a 20% reduction in smoking is easily understood, and we can learn what a five-point improvement on the Problem Behaviour Scale means in practical terms, even if we are initially unfamiliar with the scale.

However, there are limitations to this information. In particular, effects described as they are originally measured will not always be directly comparable. For example, child behaviour problems might be measured in one study on a scale of 1–5 using the Problem Behaviour Scale, and in another study on a scale of 1–12 using the Externalising Problems Inventory. A three-point change on one of these scales may mean something very different from a three-point change on the other scale, and it may be unclear, at a glance, which effect is larger or more meaningful.

Improvement index scores tell us something useful about the relative size of improvement, compared with improvements measured using other scales. This is because the improvement index score is based on a standardised measure of the size of effects, which allows us to compare the relative size of effects, and to compare effects across interventions that may have evaluated improvements on a given outcome using different scales. It also means that a larger improvement index value always, relatively speaking, indicates a larger effect.

11. Assigning cost ratings

The Guidebook aims to provide a cost rating for each intervention given an evidence rating. The cost rating is an estimate of how relatively costly an intervention is per person receiving the intervention, based on the resources required to set up and deliver it. The estimate is based on characteristics such as the time it takes to deliver the intervention, how many families it is typically delivered to (cohort size), and any staff training or qualifications required.

The cost rating is **not** an actual price or fee. Instead, it is based on unit costs, which is an estimated cost for delivering the intervention to one person, a family, or a group (e.g. such as a classroom). The rating reflects a range within which Foundations expect the real cost per recipient (or unit) to fall. The aim is to give a reasonable, indicative basis for comparing interventions.

The cost rating is also **not** the same as the market price for an intervention. It is an assessment of the resources an intervention requires in order to be delivered fully, and the relative cost per recipient, not what it will cost to buy or commission the service. The actual market price typically

includes commercially sensitive information that is not routinely available, so will in practice need to be negotiated between provider and commissioner, and can vary.

Cost information comes directly from providers through an online form known as the ISF2 (Intervention Submission Form Part 2). They are asked for detailed estimates about what the intervention requires, how it is delivered, and any training or materials involved.

The estimated cost per recipient is translated to a scale from 1 to 5, where 1 indicates the least resource-intensive interventions and 5 the most resource-intensive per recipient. The rating helps commissioners and decision-makers compare the likely costs of different interventions, based on the resources they require.

The updated cost model

In 2025–2026 the model underpinning the cost ratings was updated. The revised approach has simplified the model and takes into account inflation and increases in costs. They key differences between the previous model and the revised model are:

- The revised approach takes an economic modelling perspective, aiming to understand as much as possible the different components of costs – this is arguably better when the dataset is small; the previous approach was purely statistical, involving estimating the weights assigned to different components.
- The revised approach therefore only requires a single regression analysis to take place to establish the model, rather than a hybrid regression analysis after an iterated grid-search.
- The revised approach simplifies the parameters included, by subsuming components of costs into a single 'indicative unit cost variable', and streamlining requirements for the supervision variable.
- The revised approach includes further streamlining, by not requiring a 'level of need variable' or whether the level of external supervision is 'intensive' or not.

The input variables to the revised model are:

- Intervention duration
- Average number of intervention participants per cohort
- Whether a licence is required
- Number of staff required to deliver the intervention
- Skill level of staff (RQF level)
- Delivery duration for each staff member
- Cost of materials in first year
- Ongoing cost of materials (years 2 to 5)
- Intervention training fee
- Whether booster training sessions are provided
- Number of staff receiving training

- Training time
- Number of internal supervisors, their qualification level, and training time
- Number of external supervisors, and their qualification level.

The new cost model is not retrospectively applied to existing Guidebook intervention entries but to new and updated intervention entries from 2026 onwards. A technical paper detailing the development and specification of the model is available on request.

12. Publishing the Guidebook intervention entry

Once an intervention has completed the appraisal process, it can be listed on the Foundations Guidebook.

The top-level Guidebook entry includes a brief description of the intervention, the evidence rating, cost rating, positive child outcomes with evidence of improvement, and whether the intervention is UK available or UK tested, as well as the characteristics of the population the intervention was evaluated in, and characteristics of the intervention model, including type (e.g. group), setting and workforce. This top-level summary is followed by a more detailed model description, target population information, and the intervention's theory of change and implementation requirements.

Which child outcomes and other outcomes appear on the Guidebook?

The Guidebook assesses the strength of evidence for interventions across a set of seven clearly defined child outcome domains. These reflect key priorities for improving the lives of children and young people:

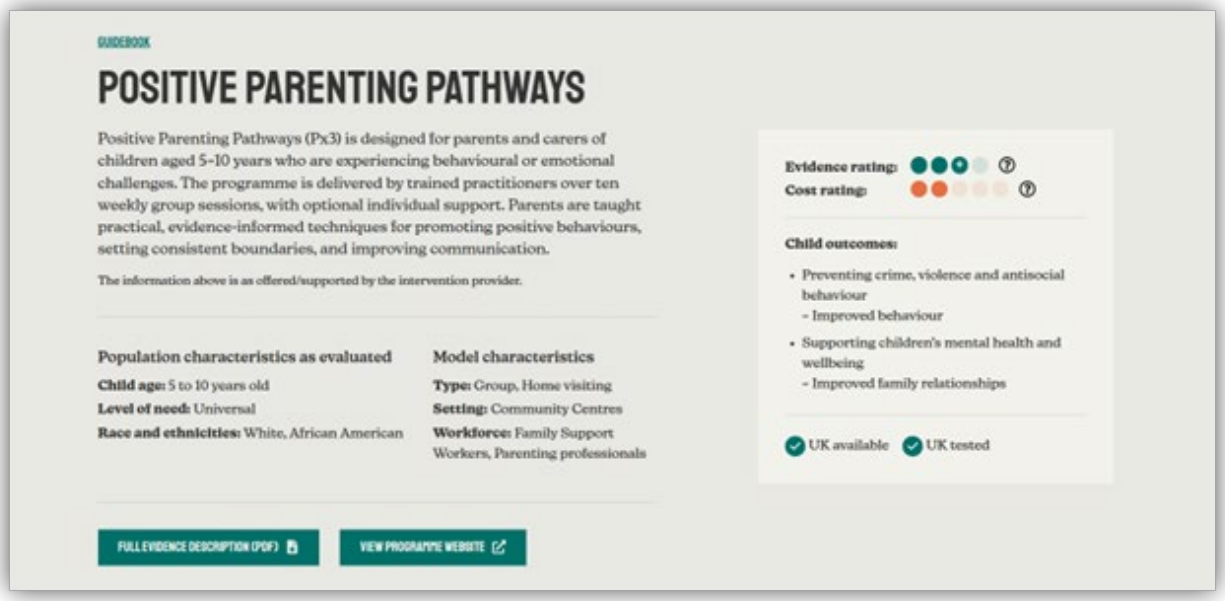
1. Supporting children's mental health and wellbeing
2. Preventing child maltreatment
3. Enhancing school achievement and employment
4. Preventing crime, violence, and antisocial behaviour
5. Preventing substance abuse
6. Preventing risky sexual behaviour and teen pregnancy
7. Preventing obesity and promoting physical healthy development.

Positive outcomes are grouped under the seven child outcome domains, for example: 'Supporting children's mental health and wellbeing' includes among other outcomes 'improved emotional wellbeing', 'improved social behaviour', and 'reduced anxiety'. These plain language outcomes are presented under the seven outcome domains in the intervention summary. They help users compare evidence across different interventions using a consistent framework.

The Guidebook includes information about parent and other outcomes, such as improved positive parenting, reduced parental stress, improved parental mental health, or improved teaching

strategies. Parent and other outcomes are listed in study summaries. They are only included when reported in studies which also included child outcomes and which have been assessed for the Guidebook. While parent and other outcomes can provide valuable context, they do not contribute to the evidence rating and have not been assessed with the same level of scrutiny. If a study only reports parent outcomes, with no accompanying child outcomes, it is not included on the Guidebook at all.

The terminology used to describe outcomes may vary between research and practice. In the intervention summary, outcomes are presented in terms more familiar to practitioners and local leaders. However, within the full evidence description, outcomes are described using the language found in the original studies. This may sometimes include terms that are outdated or considered no longer appropriate; these reflect the original wording by the study authors and do not represent Foundations' preferred terminology.



Model description

The model description is drafted by the Foundations Guidebook team, based on information about the intervention supplied by the provider or developer, together with descriptions of the intervention in the studies which underpin the evidence rating. It provides a summary of the type of intervention, its target population and target outcomes, followed by further detail about the delivery and content of the intervention. This might include the kinds of approaches taken by practitioners during intervention sessions or the techniques used, what parents or children are learning or developing through the intervention, and the structure of the intervention.

EDIE summary

The Guidebook intervention listing also includes two narrative EDIE summaries; evidence of impact on disparities and evidence on implementation and family experience. The evidence of

impact on disparities summary follows a standardised template and draws from Level 3 studies that report on subgroup analyses or impacts on specific populations, using information gathered at the extraction stage, while the evidence on implementation and family experience summary highlights how interventions are delivered in practice and how acceptable, relevant, and effective they are for different groups, particularly racially minoritised communities and other marginalised groups. This summary draws on a broad range of evidence sources for the specific intervention, including implementation and qualitative studies, as well as Level 3, Level 2, and NL2 studies.

Guidebook EDIE commitment

Narrative summaries provide a fuller picture of the evidence of impact on specific populations, and the lived experiences of children and families receiving the intervention, using established tools such as the PRO EDI tool. Evidence on implementation and family experience draws on a broad range of evidence to reflect the reality that studies focused on marginalised or racially minoritised groups often face systemic barriers to achieving a Level 3 rating under our current standards, and yet often contain valuable insights into how interventions are experienced by families and practitioners, particularly within marginalised or minoritised communities.

Intervention pages identify and describe design features that aim to promote equity, within the description of the intervention model. This includes approaches that were developed with or for specific communities, or that intentionally address access and inclusion. These details are documented using the same EDIE frameworks to ensure consistency, using information from intervention developers or providers. Listings also include mention of or links to adaptations of models for specific populations.

Theory of change

A theory of change describes the evidence-based assumptions behind an intervention, the need for the intervention, and how the intervention works to achieve its intended positive outcomes for children and young people. Theories of change are useful for those commissioning and implementing interventions, because they help achieve a shared understanding of the intervention, its aims and rationale, and point to the existing evidence base that underpins the need for and design of the intervention. They are useful for evaluators as they can shape what outcomes are being evaluated (the research questions) and provide explanations for any effects which are observed. Reflecting on a theory of change may also highlight gaps in assumptions or uncertainties in plans to implement an intervention.

On the Foundations Guidebook, an intervention's theory of change is based on information provided by the intervention developer or provider, together with descriptions of the intervention found in studies and on the intervention website. For parenting interventions' theories of change, interventions which are based on similar theories of child development or which work in a similar

way are now described in a similar way too, making it easier to see how 'families' of interventions are similar or differ.

The Foundations Guidebook also makes the theory of change transparent, by highlighting different aspects of it: the science-based assumption, who the intervention targets, what the intervention does, and what the short-, medium-, and long-term outcomes are intended to be. While the theories of change on the Guidebook include established scientific theories, this does not necessarily mean that the theories of change themselves have been rigorously tested for particular interventions.

Implementation requirements

The implementation requirements summarises key pieces of information about how the intervention is delivered, including:

- Eligibility
- Delivery intensity and duration (how many sessions, for how long)
- Content of the intervention
- Practitioner requirements, training and supervision
- Systems for maintaining fidelity to the intervention model
- Contact details for the intervention provider.

This information is supplied to Foundations by the provider as part of the assessment process.

Evidence summary

The evidence summary includes a high level overview of the strength of evidence underpinning the entry, including the number, type and location of the contributing studies as well as their evidence rating and a plain language explanation of the rating and the significant outcomes reported in the studies.

The evidence summary is also where improvement indexes for individual child outcomes are reported, where available, with an interpretation.

The study summaries are published underneath the evidence summary, and include more detail, with key study features including race, ethnicities and nationalities of the study population and specific outcomes evidenced in each study.

Full evidence description

Each intervention page includes a downloadable PDF with a full evidence description. This includes an outcomes table, which shows the full results of studies underpinning the intervention. These tables report all outcome measures, their statistical significance, timepoint, and, where available, the number of participants (n) and effect sizes. Not all studies report sample sizes at every timepoint or provide effect sizes. Outcomes which do not contribute to an intervention's rating (for example, because they were not a statistically significant positive outcome), appear only on the outcomes table, and not in the study or evidence summaries on the intervention webpage.

The full evidence description PDF also includes new information about how studies were conducted, characteristics of those who participated, how many participants stayed in the study, and which measurement tools were used. This helps build a fuller picture of how reliable and relevant the evidence is to the local context.

13. Publishing Not Level 2 evidence assessment findings

Following assessment, some interventions may receive an NL2 rating. This rating indicates that the intervention does not yet have the direct or robust evidence needed for Foundations to confidently assess the scale and nature of its impact on children and families' outcomes. Because of this, interventions rated NL2 will not appear in our standard Guidebook search results. However, they will be available on a dedicated NL2 list on our website, so that users can still access key information about them, including the primary reason for the rating (for example, no quantitative impact evaluation to date, or only theory-building studies available).

Importantly, an NL2 rating **does not** mean the intervention is ineffective. Instead, it can reflect that the current evidence base is still emerging. Interventions in this category are often at an earlier stage of development, where essential groundwork is being carried out. This may include:

- Developing a theory of change or logic model
- Conducting feasibility or implementation studies
- Running early pilot evaluations.

With the right support and resources for further evaluation, many NL2 interventions have the potential to become evidence-based interventions in the future. They are included because it is important to highlight promising work in progress and to encourage continued evaluation and theory-building that may demonstrate future impact.

Guidebook EDIE commitment

Interventions designed for, or predominantly used by, minoritised groups may be less likely to meet the criteria for higher evidence ratings. This can be due to a variety of methodological constraints, including challenges around sample representativeness, limited statistical power, or the availability and validation of culturally appropriate measures and methods. Importantly, these limitations do not mean that such interventions are ineffective.

Publishing a list of NL2 studies creates a transparent and accessible resource that highlights emerging and contextually relevant work for minoritised groups. The intention is that this supports practitioners and researchers in identifying possible approaches, while also encouraging further evaluation, refinement, and theory-building to strengthen the evidence base for these interventions.

REFERENCES

Eldridge, S., Campbell, M. K., Campbell, M. J., Drahota, A. K., Giraudeau, B., Reeves, B. C., ... & Higgins, J. P. T. (2021) *Revised Cochrane risk of bias tool for randomized trials (RoB 2): Additional considerations for cluster-randomized trials (RoB 2 CRT).* https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool

Higgins, J. P., Savović, J., Page, M. J. & Sterne, J. A. (2019) Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). *Cochrane handbook for systematic reviews of interventions*, *28*, 366–438. Accessed at https://www.riskofbias.info/welcome/rob-2-0-tool [April 2026]

Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., ... & Higgins, J. P. (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials. *bmj*, *366*.

Wilson, S. J., Brown, S. R., Kerns, S. E. U., Dastrup, S. D., Hedberg, E., Schachtner, R., Jackson, C., Norvell, J., Campbell, W. & Wall, A. (2024) *Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures, Version 2.0*, OPRE Report # 2024-127, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

APPENDIX 1

List of interventions assessed as NL2

According to the Foundations standards of evidence, interventions that do not meet the threshold for a level 2 rating do not yet have preliminary evidence of achieving outcomes for children.

Interventions listed on this page have previously been assessed and at the time of assessment were found not to meet the threshold for inclusion in the Guidebook, because they are rated 'not level 2' (or NL2).

This might have been due to: methodological limitations in quantitative impact studies; because no child outcomes were measured in the evaluation; or because the interventions are at an earlier stage of development, developing a theory of change or logic model, or carrying out feasibility, implementation, or pilot evaluation studies.

NL2 programmes:

- Active Parenting
- Assertive Outreach Model, including Baby Express
- Baby Express
- Baby Steps
- Bookstart Baby
- Bookstart Corner
- Born to Move
- Circle of Security (home visiting)
- Enhancing Adoptive Parenting
- Enhancing Parenting Skills programme (EPAS)
- Families and Schools Together (FAST) Baby
- Family Action's Perinatal Support Project (evolved from Newpin)
- Go-Givers Make A Difference Challenge (MADC)
- It Takes Two to Talk
- Kaleidoscope Play & Learn
- Learning Together Programme – Early PEEP: 1s Level
- Learning Together Programme – Early PEEP: 2s Level
- Learning Together Programme – Early PEEP: Baby PEEP
- Listening to children (LTC)
- Mellow Babies
- Mellow Bumps
- Modified Interaction Guidance
- My Baby's Brain
- Parent Infant Project (PIP)

- Parenting Wisely
- Parents 1st Community Parent Volunteer Peer Support Programme
- Parents as Partners (formerly known as Supporting Father Involvement)
- Second Step Middle School
- Sing & Grow Programme
- Strengthening Families Program
- TalkAbility
- Target Word
- The Newborn Behavioral Observations (NBO) System
- Triple P Primary Care
- Video-feedback Intervention to Promote Positive Parenting – Sensitive Discipline (VIPP-SD)
- Video-feedback Intervention to Promote Positive Parenting (VIPP).

APPENDIX 2

Accessibility text

Figure 1. Exhibit 2.4 in *Prevention Services Clearinghouse Handbook Version 2.0* (p. 23) 'Process for Assessing Substantial Adaptations'

The chart uses a series of yes/no decision points, shown in dark grey boxes, to determine whether an adaptation to a programme or service is substantial or not substantial. The two possible outcomes are shown in coloured boxes at the bottom: green for "Adaptation is not substantial" and red for "Adaptation is substantial."

Question 1 asks: Is the adaptation (a) explicitly prohibited in the focal manual selected for the program or service under review, or (b) the result of adding a separate program or service (i.e., "bundling") to the existing program or service? If Yes, the outcome is Adaptation is substantial. If No, proceed to Question 2.

Question 2 asks: Is the adaptation explicitly allowed by the focal manual selected for the program or service under review? If Yes, the outcome is Adaptation is not substantial. If No, proceed to Question 3.

Question 3 asks: Does the adaptation substantially change a key program or service component in the focal manual selected for the program or service under review? If Yes, the outcome is Adaptation is substantial. If Unclear, proceed to Question 4. If No, the outcome is Adaptation is not substantial.

Question 4 asks: After gathering any additional information needed, have experts determined that the adaptation is substantial? If Yes, the outcome is Adaptation is substantial. If No, the outcome is Adaptation is not substantial.

Go back

Figure 2. Assigning a study evidence rating

A flowchart showing the decision-making process used to classify the evidence level of a study. The chart uses boxes and arrows, and moves left to right across three main pathways.

The top pathway begins if Level 2 standards are met. It proceeds first to a RoB2 / RoB2 CRT/ ROBINS-I assessment, then to Guidebook Level 3 criteria, and then to Level 4 criteria. If all are met, the study is classified as Level 3 with Level 4 features. If Level 4 criteria are not met, the classification is Level 3 without Level 4 features.

If Guidebook Level 3 criteria are not met in the top pathway, the study is assessed against Level 2+ criteria. If these are met, the classification is Level 2+. If not, the classification is Level 2.

If RoB2/RoB CRT/ROBINS-I is not met in the top pathway (described as "Evidence Met without Impact"), the study moves into the lower pathway.

The second pathway begins if L2 standards are not met – in this case, the study is classified as NL2.

The third pathway begins if Level 2 standards are met but there is no evidence of impact. In this case, the study then proceeds to assessment with RoB2 /RoB CRT / ROBINS-I. If this is not met, it goes to NL2. If this is met, it proceeds to Guidebook Level 3 criteria; if these are not met, it is classified as NL2; if these are met, it is classified as NE (No Effect).

Go back

Figure 3. Potential bias associated with overall and differential attrition (Exhibit 5.3. from the *Prevention Services Clearinghouse Handbook v2*)

The chart shows the relationship between overall attrition (shown on the horizontal axis, as a percentage) and differential attrition (shown on the vertical axis, as percentage points), and whether their combination represents a tolerable or unacceptable threat of bias.

- The horizontal axis shows overall attrition percentage, ranging from 5 to 55
- The vertical axis shows differential attrition in percentage points, ranging from 1 to 7
- The chart is divided into two colour-coded regions separated by a curved boundary line that slopes downward from left to right
- The green region, labelled "Tolerable threat of bias under cautious assumptions," occupies the lower-left portion of the chart. This represents combinations of overall and differential attrition that are considered acceptable
- The red region, labelled "Unacceptable threat of bias under cautious assumptions," occupies the upper-right portion of the chart. This represents combinations where the level of attrition poses an unacceptable risk of bias
- The boundary between the two regions indicates that as overall attrition increases, the tolerable threshold for differential attrition decreases.

Go back

Figure 4. Flowchart illustrating steps to judge whether the measure is not inappropriate; answers to questions lead to RoB2 judgement for signalling question 4.1

The chart uses boxes and arrows to show a sequence of four questions assessed in order. At each question, answering No (N) leads to an immediate outcome of Y/PY (shown in red), while answering Yes or No Information (Y/NI) leads to the next question. If all four questions are answered Y/NI, the final outcome is PN/N (shown in green).

Question 1 asks: Is the measurement independent of treatment? If No, the outcome is Y/PY. If Yes or No Information, proceed to Question 2.

Question 2 asks: Is the measure appropriate for the target population? If No, the outcome is Y/PY. If Yes or No Information, proceed to Question 3.

Question 3 asks: Does the measure have high face validity? If No, the outcome is Y/PY. If Yes, proceed to Question 4.

Question 4 asks: Does the measure have high reliability? If No, the outcome is Y/PY. If Yes, the final outcome is PN/N.

Go back

Figure 5. Flowchart illustrating how to establish whether measurement differed between groups; answers to the questions lead to RoB2 judgement for this signalling question (Y/PY/NI/PN/N) (4.2)

The chart uses boxes and arrows to show a sequence of two questions. Answering No (N) at either question leads to an outcome of Y/PY (shown in red). Answering No Information (NI) at either question leads to an outcome of NI (shown in orange). Answering Yes (Y) to both questions leads to a final outcome of PN/N (shown in green).

Question 1 asks: Is the timing of measurement equivalent across groups? If No, the outcome is Y/PY. If No Information, the outcome is NI. If Yes, proceed to Question 2.

Question 2 asks: Is the mode of measurement equivalent across groups? If No, the outcome is Y/PY. If No Information, the outcome is NI. If Yes, the final outcome is PN/N.

Go back

Figure 6. Calculation of overall and differential attrition in cluster trials (from *Prevention Sciences Clearinghouse Handbook v2* 5.6.2 p. 64)

This image consists of two formulas.

The first and top-most formula begins with the phrase “Overall Attrition” equals “Number of individuals without post-test outcome data” over “N of individuals in non-attrited clusters at randomisation.”

The second formula appears below the first. It begins with the phrase “Differential Attrition” followed by an equals sign. To the right of the equals sign is a more complex expression enclosed within tall brackets, indicating an absolute value.

Inside the absolute value brackets are two fractions, separated by a minus sign. The entire expression is visually larger and more complex than the first formula because it contains two stacked fractions.

Left Fraction (Intervention Group Attrition):

- Numerator: “N of intervention condition members without post-test outcome data.”
- Denominator: “N of intervention condition members in non-attrited clusters at randomisation.”

Right Fraction (Comparison Group Attrition):

- Numerator: “N of comparison condition members without post-test outcome data.”
- Denominator: “N of comparison condition members in non-attrited clusters at randomisation.”

Go back

Figure 7. Potential bias associated with overall and differential attrition (Exhibit 5.3. from the *Prevention Services Clearinghouse Handbook v2*)

The chart shows the relationship between overall attrition (shown on the horizontal axis, as a percentage) and differential attrition (shown on the vertical axis, as percentage points), and whether their combination represents a tolerable or unacceptable threat of bias.

- The horizontal axis shows overall attrition percentage, ranging from 5 to 55
- The vertical axis shows differential attrition in percentage points, ranging from 1 to 7
- The chart is divided into two colour-coded regions separated by a curved boundary line that slopes downward from left to right

- The green region, labelled "Tolerable threat of bias under cautious assumptions," occupies the lower-left portion of the chart. This represents combinations of overall and differential attrition that are considered acceptable
- The red region, labelled "Unacceptable threat of bias under cautious assumptions," occupies the upper-right portion of the chart. This represents combinations where the level of attrition poses an unacceptable risk of bias
- The boundary between the two regions indicates that as overall attrition increases, the tolerable threshold for differential attrition decreases.

Go back

Figure 8. Flowchart illustrating steps to judge whether the measure is not inappropriate; answers to questions lead to RoB2 judgement for signalling question 4.1

The chart uses boxes and arrows to show a sequence of four questions assessed in order. At each question, answering No (N) leads to an immediate outcome of Y/PY (shown in red), while answering Yes or No Information (Y/NI) leads to the next question. If all four questions are answered Y/NI, the final outcome is PN/N (shown in green).

Question 1 asks: Is the measurement independent of treatment? If No, the outcome is Y/PY. If Yes or No Information, proceed to Question 2.

Question 2 asks: Is the measure appropriate for the target population? If No, the outcome is Y/PY. If Yes or No Information, proceed to Question 3.

Question 3 asks: Does the measure have high face validity? If No, the outcome is Y/PY. If Yes, proceed to Question 4.

Question 4 asks: Does the measure have high reliability? If No, the outcome is Y/PY. If Yes, the final outcome is PN/N.

Go back

Figure 9. Flowchart to show the steps to check that an appropriate analysis has been used and Criterion 6 is met

The chart uses boxes and arrows to show a sequence of two questions. Answering No (N) at either question leads to an outcome of Y/PY (shown in red). Answering No Information (NI) at either question leads to an outcome of NI (shown in orange). Answering Yes (Y) to both questions leads to a final outcome of PN/N (shown in green).

Question 1 asks: Is the timing of measurement equivalent across groups? If No, the outcome is Y/PY. If No Information, the outcome is NI. If Yes, proceed to Question 2.

Question 2 asks: Is the mode of measurement equivalent across groups? If No, the outcome is Y/PY. If No Information, the outcome is NI. If Yes, the final outcome is PN/N.

Go back

Figure 10. Flowchart to show the steps to check that an appropriate analysis has been used and Criterion 6 is met

The image is a vertically oriented flowchart composed of teal rectangular boxes connected by black arrows. The chart visually represents a step-by-step decision process for determining whether a study's analysis requirements are *met* or *unmet*. Each box contains a question about analytical methods, and each arrow leads to the next decision point or to a final judgment.

First Box – First Question: "Is an appropriate analysis technique used given nature of data?" Two arrows extend from this box: One arrow is labelled "N/NI" (meaning No or Not Indicated) and has a "red flags" note, this arrow leads to a final outcome labelled "Unmet." The other arrow is labelled "Y/NI" (Yes or Not Indicated). This arrow continues the main flow downward to the next question.

Second Box: "Are inappropriate methods used?" Two arrows extend from this box. A **"**Y" arrow branches sideways to a final outcome labelled "Unmet." An "N/NI" arrow continues downward to the next question.

Third Box: "Does baseline variance need controlling for?" Two arrows extend to two different boxes with different questions. The "Y" arrow extends to the fourth box with the question "Is imbalance controlled for in analyses?" And the "N" arrow extends down to the fifth box with the question "Unequal allocation?"

Fourth Box: "Is imbalance controlled for in analyses?" two arrows extend: a "N/NI" arrow branches sideways to a final outcome labelled "Unmet"; a "Y" arrow continues downward to the next fifth box/question.

Fifth Box: "Unequal allocation?" The "Y" arrow extends to the sixth box with the question "Accounted for appropriately in impact model?" And the "N" arrow extends down to the seventh (final) box with the question "Endogenous measures used as covariates?"

Sixth Box: "Accounted for appropriately in impact model?" Two arrows extend from this: A "N/NI" arrow branches sideways to a final outcome labelled "Unmet", A "Y" arrow continues downward to the final box and question.

Seventh Box and final question: "Endogenous measures used as covariates?" One arrow extends labelled "N" leads to the final outcome labelled "Met".

Go back

Figure 51. Guidebook evidence standards summary

A graphic summarising Foundations' Evidence Standards, showing five evidence levels represented as stacked rectangles in different shades of teal. The levels are arranged vertically from bottom (the lightest shade of teal) to top (the darkest shade of teal), increasing in rigour. The 'No Effect' finding.

- NL2 (Not at Level 2) is the lowest level, shown as the lightest shade of teal. Key elements of the logic model are being confirmed and verified in relation to practice and the underpinning scientific evidence. Testing of impact is underway but evidence of impact at Level 2 has not yet been achieved.
- Level 2 – Preliminary Evidence is the next level up, shown in a slightly darker (but still quite pale) shade of teal. This requires evidence of improving a child outcome from a study involving at least 20 participants, representing 60% of the sample, using validated instruments.
- Level 3 – Efficacy is shown above Level 2, in medium teal shade. This requires evidence from at least one rigorously conducted evaluation demonstrating a statistically significant positive impact on at least one child outcome.
- NE (No Effect) is shown as a grey area within the Level 3 rectangle. This represents a finding of no effect on measured child outcomes in a high-quality impact evaluation (Level 3).
- Level 4 – Effectiveness is the highest level, shown at the top in the darkest shade of teal. This requires evidence from at least two high-quality evaluations demonstrating positive impacts across populations and environments lasting a year or longer. The evidence may include significant adaptations to meet the needs of different target populations.

Go back