YOUTH
ENDOWMENT
FUND

NATIONAL
CHILDREN'S
BUREAU

Technical Guide

May 2025

About National Children’s Bureau

This technical Guide has been revised by the National Children’s Bureau on behalf of the Youth Endowment Fund. The National Children’s Bureau works collaboratively across the issues affecting children to influence policy and get services working together to deliver a better childhood. They were commissioned by the Youth Endowment Fund (YEF) as their Toolkit Partner 2023-2026.

About Youth Endowment Fund

The Youth Endowment Fund’s mission is to prevent children and young people becoming involved in violence. They do this by finding out what works and building a movement to put this knowledge into practice. The fund was established in March 2019 by children’s charity Impetus, with a £200m endowment and ten-year mandate from the Home Office. For more information, please visit www.youthendowmentfund.org.uk.

Table of contents

List of tables

List of figures

1. About this Technical Guide

According to the World Health Organization violence is *"the intentional use of physical force or power, threatened or actual, […] that either results in, or has a high likelihood of, resulting in injury, death, psychological harm, maldevelopment or deprivation" (Krug et al, 2002, p.5).* The YEF Toolkit aims to summarise evidence-based approaches which could prevent children and young people (CYP) becoming involved in violence, or, for those already involved, reduce their engagement in violence.

The Toolkit aims to make research findings:

1. Accessible: research is written in plain language, without jargon.
2. Applicable: research included is relevant to both youth violence and the approach.
3. Available: it aims to bring together research from various journals/websites and make the findings readily available.
4. Accurate: the methodology underpinning the 'best bets' is reliable and valid, based on research and consultations with experts to find the best available methodology.
5. Actionable: it focuses on how the research findings have a practical influence on working with CYP at risk of, or engaging in, violence.

Aims

The Technical Guide underlies the Technical Reports written for each of the approaches in the Toolkit. Assessing different approaches allows us to see what works best for reducing violence involving CYP. To enable this, the Technical Guide provides a clear overview of the methodologies used to:

1. Identify the estimated impact of each approach on violence.
2. Calculate the evidence quality of each approach.
3. Consider how equal, diverse, inclusive and equitable the approach is.
4. Assess how best to implement the approach.
5. Narratively summarise the cost-effectiveness of the approach.

6. Make the EGM a ‘living resource’ which underpins the Toolkit and Technical Guides.

Statistical methodologies, like any research, are continually growing in knowledge, being challenged and changed. As such, this updated version of the Technical Guide is based on in-depth research on statistical methodologies, consultations with methodological experts, and reviews by an end user group. For a summary of updates, see Appendix 1.

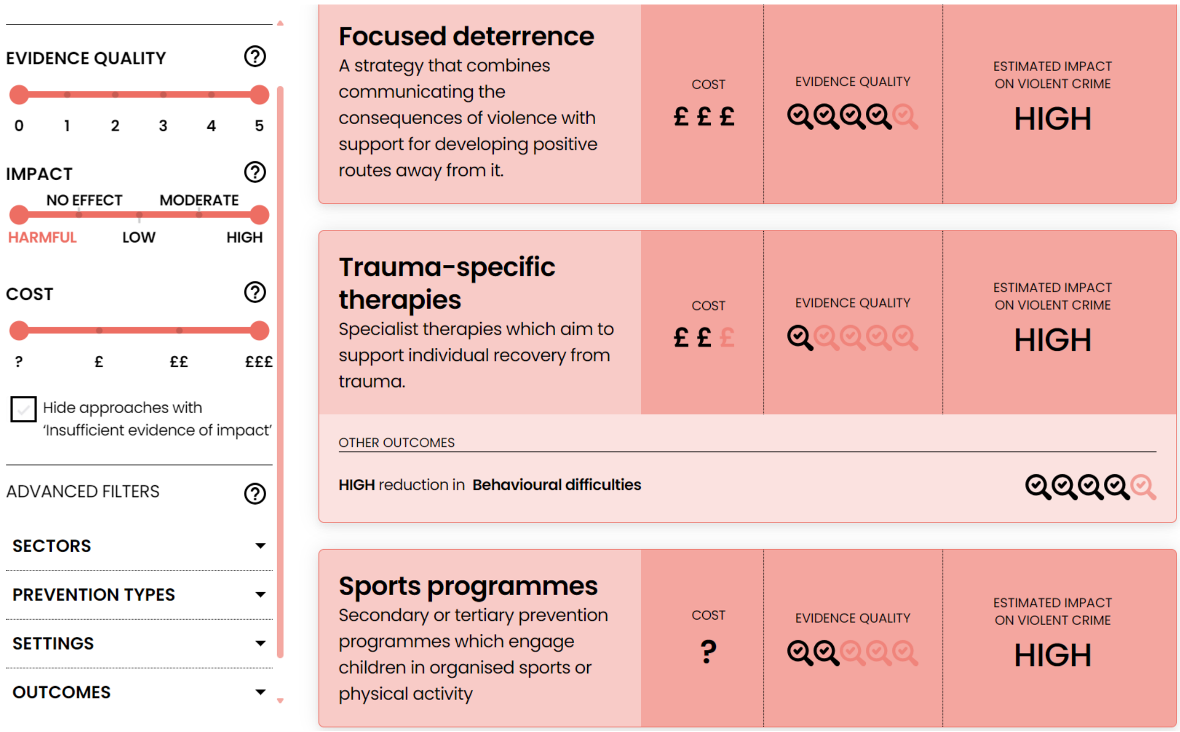
2. Toolkit Structure

There are three core levels to the YEF Toolkit, which will be summarised clearly here. These include the (1) Toolkit front page, (2) Summary page, and (3) Technical Report.

Toolkit Front Page

The Toolkit Front Page provides a quick overview of a variety of approaches and their relative success at reducing violence (see Figure 1). The Toolkit Front Page provides a brief description of the given approach alongside three key ratings (cost, evidence quality, estimated impact on violence). To make these easily accessible, icons such as magnifying glasses to indicate the strength of evidence quality are used. Where available, other outcomes that are known to be highly relevant to engagement in violence are indicated, alongside their evidence quality rating.

Figure 1. Toolkit Front Page showing brief summaries and key ratings



Focused deterrence
A strategy that combines communicating the consequences of violence with support for developing positive routes away from it.
COST £££ | EVIDENCE QUALITY 🔍🔍🔍🔍 | ESTIMATED IMPACT ON VIOLENT CRIME HIGH

Trauma-specific therapies
Specialist therapies which aim to support individual recovery from trauma.
COST ££ | EVIDENCE QUALITY 🔍 | ESTIMATED IMPACT ON VIOLENT CRIME HIGH
OTHER OUTCOMES
HIGH reduction in **Behavioural difficulties** 🔍🔍🔍🔍

Sports programmes
Secondary or tertiary prevention programmes which engage children in organised sports or physical activity
COST ? | EVIDENCE QUALITY 🔍🔍 | ESTIMATED IMPACT ON VIOLENT CRIME HIGH

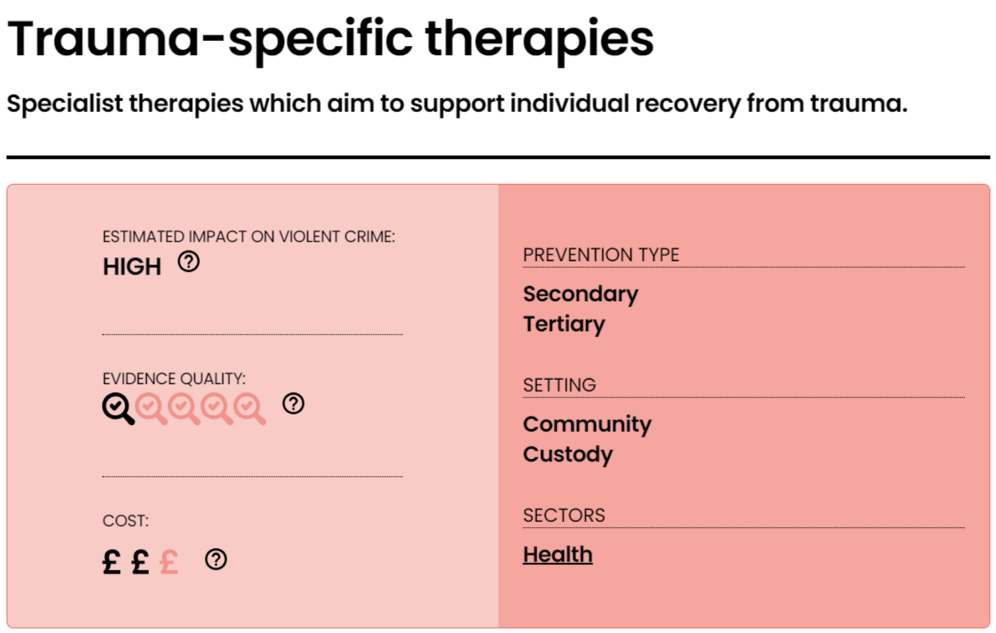
Summary Page

The Summary Page provides an in-depth narrative on the specific Toolkit approach selected. At the top is an expanded box (see Figure 2), which provides additional information on the prevention type, setting, and sector. The narrative for every summary covers the following key topics:

1. **What is it?** This provides a detailed description of the approach, including core components or activities required.
2. **Is it effective?** This section outlines the estimated impact on violence and on any other relevant outcomes.
3. **Who does it work for?** This section provides a narrative overview of the best available evidence that examines which CYP have been involved in evaluations. This section considers ethnicity, gender, experience of deprivation, Special Educational Needs and Disabilities (SEND) or care experience, and whether any adaptations have been tested to improve acceptability, attrition, and effectiveness.

4. **How secure is the evidence?** This section indicates how confident we are in the impact rating. Summaries of the evidence security for any other outcomes, along with degree of confidence, are also provided.
5. **How can you implement it well?** This section utilises the best available evidence to provide a narrative synthesis of how the approach can be used in practice.
6. **How much does it cost?** This section provides a narrative summary of the costs incurred when delivering the approach, providing UK evidence where available.
7. **What programmes are available?** This provides links to programs in the Foundations Guidebook, where relevant.
8. **Topic summary.** This provides a bullet-pointed summary of the approach.
9. **Take away messages** This provides a short list of YEF recommendations for actions to be taken by commissioners, decision-makers, and policymakers.
10. **YEF Projects and Evaluations.** This provides a brief overview and links to completed and/or ongoing projects funded by YEF on the given approach.
11. **Downloads.** This provides a link to a PDF download of the technical report, and the implementation guidance (where available).

Figure 2. Toolkit Summary Page showing in-depth top box



Trauma-specific therapies

Specialist therapies which aim to support individual recovery from trauma.

ESTIMATED IMPACT ON VIOLENT CRIME:
HIGH ?

EVIDENCE QUALITY:
●○○○○ ?

COST:
£ £ £ ?

PREVENTION TYPE
Secondary
Tertiary

SETTING
Community
Custody

SECTORS
Health

Technical Report

The Technical Report is the most detailed of all three layers of the Toolkit. Here you will find a step-by-step report of the methodology used to create the Front Page headline ratings, as well as the narratives needed for the Summary Page. There is an individual technical report written, based on this guide, for each approach on the Toolkit. By scrolling to the bottom of any of the Summary Pages, there will be a downloadable technical report, detailing the methodology and findings for the given approach.

3. Selecting evidence for the Toolkit

Approaches to be covered by Toolkit strands are agreed by YEF through an internal process of scoping and prioritisation which considers the evidence that is available, the relevance of the approach to YEF's work and mission, and the salience to the Toolkit's audience. This section outlines what information is needed to create the three levels of a Toolkit strand, and how to locate this information. The following sections will describe in detail how to calculate the Toolkit-specific measures including the impact rating, evidence security rating, and cost rating.

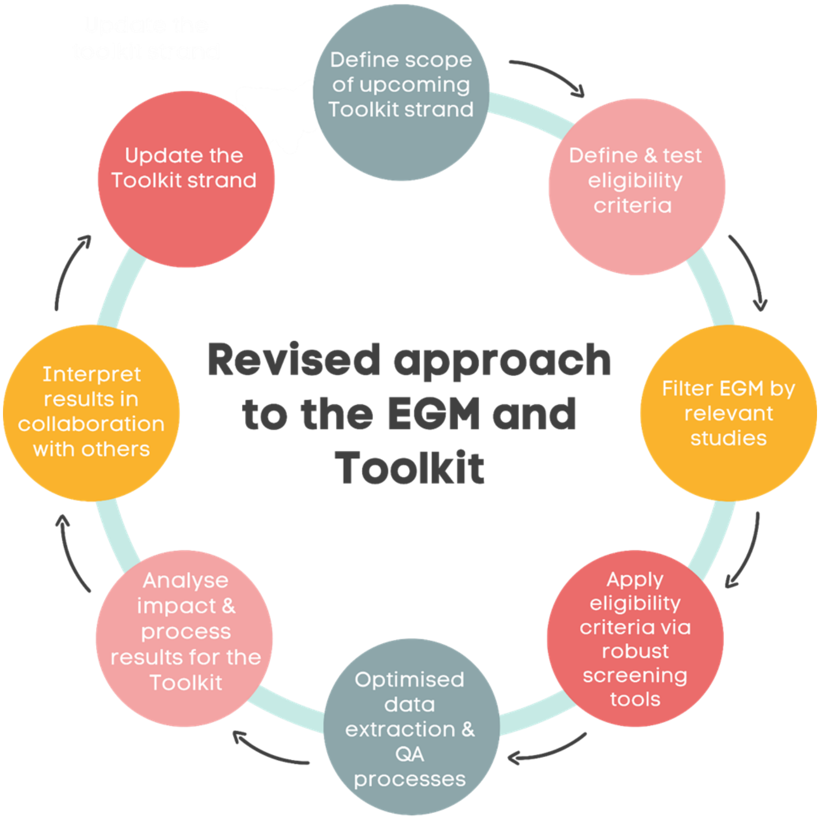
Locating evidence within the Evidence and Gap Map

The updated Evidence and Gap Map (EGM) is a comprehensive and up-to-date source of studies that are relevant to YEF's work. The EGM is housed within EPPI-Reviewer, an application for conducting systematic reviews (EPPI-Reviewer, 2024). For more information on how the EGM searches for, screens and extracts data from studies, see the EGM protocol available on Open Science Framework (OSF).[1] The EGM extracts the following data for all studies:

- Year and Type of publication
- Region and Country
- Toolkit strand (or unclassified)
- Study design
- Focus of intervention: Person or Place-based
- Location/setting of intervention
- Type of targeting
- Target group of intervention
- EDIE characteristics (according to PROGRESS-Plus)
- YEF outcomes framework
- Implementation outcomes (including cost)
- Quality of systematic review (if applicable)

[1] Protocol is available to access here: https://osf.io/vamxy

Figure 3. The workflow for new Toolkit strands


Revised approach to the EGM and Toolkit
Define scope of upcoming Toolkit strand
Define & test eligibility criteria
Filter EGM by relevant studies
Apply eligibility criteria via robust screening tools
Optimised data extraction & QA processes
Analyse impact & process results for the Toolkit
Interpret results in collaboration with others
Update the Toolkit strand

Error! Reference source not found. presents the workflow for new Toolkit strands. Each T oolkit strand begins with a scoping note which sets out the PICOS criteria for the strand: the population, intervention, comparison group, outcomes, and study designs to be included. This should specify any ways in which the strand will diverge from or restrict the PICOS criteria used for the EGM (see EGM protocol for more information).

To find relevant studies within the EGM, the reviewer can filter according to several criteria, including the relevant Toolkit strand, the outcomes targeted by the intervention, the study design, and whether the study includes information on process or cost. Whilst strands prioritise studies conducted in the UK and/or Ireland, research from other countries are also included. For new Toolkit strands where studies have not yet been categorised within the EGM, the reviewer can filter by ‘uncategorised’ and use additional filters to identify studies that fulfil the PICOS criteria set out in the scoping note.

Once the searches have been carried out, studies can be sorted into the appropriate sections (NB, studies may feature in more than one of the sections listed below):

- Effectiveness: Studies to be used in the meta-analysis to produce the headline effectiveness rating. These will be quasi-experimental studies, or Randomized Controlled Trials that provide an estimate of the impact of the intervention on the outcomes of interest.
- EDIE: Studies with sub-group analysis exploring how outcomes vary by gender, age, ethnicity, SEND, and care experience, or with qualitative consideration of how different groups of children and young people may experience the intervention.
- Implementation: Studies with process insights into how the approach can be applied with CYP.
- Cost: Studies with information on cost, particularly economic or cost-benefit analysis.

Understanding whether the evidence has come from the UK (and/or Ireland) or internationally is critical for finalising many sections in the Toolkit. For instance, high or moderate quality UK and/or Ireland data is prioritised in narrative summaries in the EDIE and implementation sections. In addition, the number of studies informing each of the approaches in the Toolkit, evidence quality rating, estimated impact on violence rating, EDIE and implementation, from the UK and/or Ireland vs. internationally must be calculated. Details regarding studies locations will be extracted from the Description of Intervention codeset and recorded initially in the Location Template (see Appendix 2), before being reported on the Summary Page.

Extracting data for the Toolkit

Data is extracted within EPPI-Reviewer to ensure that all data pertaining to a study is housed in a central location. Within EPPI-Reviewer, there is a codeset named 'Toolkit – additional data extraction' which contains codes for extracting the following data:

- Quality appraisal rating – the overall rating from the YEF-Evidence Quality Assessment Tool (YEF-EQA)

- Scope of data extraction – whether the data extraction focuses on just one component of a study with a broad focus
- Study aims and purpose
- Conflict of interest – whether the authors declare their source of funding and any conflict of interest
- Intervention details – description of the aims, duration and intensity of the intervention
- Population and EDIE characteristics
- Study design / method – study timing and timing of outcome measurement
- Comparison details
- Outcomes data
- Study results and conclusions
- Implementation details
- Outcomes
- Cost details

To ensure consistency in data extraction, reviewers receive detailed training on what is and is not within the scope of data extraction and to work through examples together.

We leverage machine learning tools within EPPI-Reviewer to extract all outcome data in duplicate, including key details such as intervention types and outcomes of interest. For components where suitable machine learning tools are not available, a random 10% sample of the total studies are checked in duplicate by two reviewers using EPPI-Reviewer's 'comparison' mode. This step ensures consistency, identifies potential areas of low inter-rater reliability, and allows for clarification or additional reviewer training if necessary. Once the review team achieves a satisfactory level of consistency in data extraction, the remaining studies are divided among reviewers for individual extraction using EPPI-Reviewer's 'Normal' data entry mode. This process is designed to maintain a balance between thoroughness and efficiency while ensuring the highest standards of reliability

Quality appraising evidence for the Toolkit

Primary studies selected for inclusion in the Toolkit should be appraised using the YEF Evidence Quality Assessment (YEF-EQA) tool which covers qualitative or mixed methods process evaluations, single group pre / post designs with only one post-intervention timepoint, quasi-experimental designs, and Randomized Controlled Trials

(see Appendix 2. Location details template

Study reference (Author, year, EPPI study ID):

---

---------------------

	Number of UK Studies	Number (and Location) of International Studies
Overall, for Strand		
Contributing to Evidence Quality Rating		
Contributing to Estimated Impact on Violence		
Contributing to EDIE Information		
Contributing to Implementation		
Contributing to Cost Data		

Appendix 3 for the full version of the tool). The YEF-EQA appraises several elements of study quality, including:

- Study design
- Recruitment and sampling
- Positionality, assumptions and biases
- Outcomes
- Data analysis
- Implications and recommendations

Each item is rated as High (3 points), Medium (2 points) or Low (1) point. Some items are conditional and only apply to specific study types. Since not all domains apply to every study type, the final score is calculated as a percentage of the applicable maximum points, using the following process:

1. Calculate the total possible points based on the applicable items (excluding N/A).
2. Sum the total points awarded across applicable items.
3. Compute a percentage score:

$$\text{Final Score} = (\text{Total Awarded Points}/\text{Total Possible Points}) \times 100$$

The final score is then in the format of a percentage. According to the scoring bands, this is then used to determine the overall quality of the study as High (90-100%), Moderate (70-89%), Low (50-69%), or Very Low (below 50%).

The YEF-EQA is available as a codeset in EPPI-Reviewer, enabling reviewers to work through the codes and enter the final rating into the relevant study record within the platform. To ensure consistency and reliability, quality appraisal is checked by a senior team member, with discrepancies discussed and resolved collaboratively via team training sessions.

As quality appraisal is carried out within EPPI-Reviewer, the platform has the potential to train and evaluate a machine learning model to assist in this process. While such a model could eventually act as a 'second coder,' replacing the need for two human reviewers, this functionality remains a prospective option for the future and is not yet fully operational.

The quality appraisal ratings are recorded in EPPI-Reviewer and also reported in the Toolkit technical report. Each section that has primary studies (effectiveness, EDIE, and implementation) has a template to record the included studies together with the quality appraisal rating. In the effectiveness section, the YEF-EQA ratings provide the foundation for the evidence security rating, which indicates how much confidence we can have in the overall impact rating. Each section also includes a narrative discussion of the quality of evidence underlying it, particularly noting any issues that affect the interpretation of study results such as bias.

4. Estimating the effectiveness of Toolkit interventions

The Toolkit brings together the best available evidence to show what the most effective approaches are in preventing involvement in violence for CYP. For each approach, a meta-analysis of the results of relevant primary studies provides an overall estimate of the effectiveness of the intervention. These estimates can be compared across approaches to find the most and least effective interventions. The Toolkit presents additional measures to clearly communicate effectiveness to end users.

This chapter:

- Describes the different estimates that can be used to communicate the effectiveness of an intervention.
- Provides an overview of the meta-analysis process, including the investigation of heterogeneity.
- Outlines the various metrics used to communicate effectiveness.
- Provides the template for the reporting of estimates of effectiveness for the three different levels of the Toolkit.

Person-based versus place-based interventions

Throughout this chapter we distinguish between person-based interventions and place-based interventions. In person-based interventions, activities focus on changing the behaviour, attitudes, knowledge and skills of an individual person. The outcome measured is the likelihood of the individual person carrying out the behaviour of interest (e.g., being violent). A person-based intervention is effective if it reduces the likelihood of an individual perpetrating violence. Effectiveness is attributed to the intervention by comparing a group of individuals who receive the intervention (intervention or treatment group) with a group of individuals who do not receive the intervention (control or comparison group).

By contrast, a place-based intervention focuses on reducing the incidence of violence in a particular geographic area. For example, CCTV surveillance or street lighting change the environment in a specific area with the aim of preventing incidents of crime and violence. The outcome measured would be the number of crime or violent incidents occurring in that area within a given timeframe. A place-based intervention is effective if it reduces the number of violent incidents that occur. Effectiveness can be attributed to the intervention in one of two ways:

1. Comparing the intervention area with a similar area that does not receive the intervention (control or comparison area) over the same time period.
2. Comparing the intervention area before and after the intervention is implemented, which captures changes over time.

In place-based interventions, research findings on crime rates are used to estimate the impact on violence perpetration.

Place-based and person-based interventions are fundamentally different and therefore require different approaches to measure their effectiveness. When drafting a Toolkit strand it is critical to state whether the intervention is place-based or person-based and to follow the relevant process for calculating and reporting effectiveness, as specified below.

Some interventions may blend both person- and place-based strategies. For example, a policing strategy may target geographic hot spots (place-based) while offering individualised services (person-based) such as mentoring or referrals. In such cases, reviewers should:

- Classify the dominant modality (based on main outcome and design).
- Disclose mixed features in the narrative summary.
- Justify the analytic approach selected (e.g., use of RIRR vs. SMD).

Appropriate estimates of effectiveness

Person-based interventions

The effectiveness of person-based interventions hinges on using context-sensitive methodologies that account for individual behavioural changes. These include Randomized Control Trials (RCTs), quasi-experimental designs (QEDs), and longitudinal approaches. RCTs remain the gold standard for establishing causality, while QEDs provide viable alternatives when RCTs are unfeasible or raise ethical concerns (Cham et al., 2024). Comparatively, longitudinal approaches allow for the assessment of sustained changes over time, offering deeper insights into the long-term impact of interventions (Hill et al., 2016).

There are several effect sizes that can be used to communicate the effectiveness of a person-based intervention:

- **Cohen's d** is used to indicate the standardized difference between two means (i.e., the mean of the intervention group and the mean of the comparison group). A larger value of d indicates a greater difference between the two groups. Cohen's d is one of the most widely used effect sizes and is either reported directly by studies or can be converted from the Odds Ratio (see below). Cohen's d is useful in comparing effects regardless of the underlying measure used but is not intuitive to communicate to a non-technical audience.
- **Hedges' g** is also used to indicate the standardized difference between two means but is used for small sample sizes, which can often be the case in research focusing on youth violence. Hedges' g is interpreted in the same way as Cohen's d and for sample sizes greater than 100, the two metrics are interchangeable.
- The **Odds Ratio (OR)** is the ratio of the odds of the outcome in the intervention group to the odds of the outcome in the comparison group. It represents the odds that the outcome will occur for an individual in the intervention group relative to the odds of the outcome occurring for an individual in the intervention group. An OR of less than 1 indicates that the odds of the outcome are lower in the intervention group compared to the comparison group, while an OR of more than 1 indicates that the odds of the outcome are higher in the intervention group than in the comparison group. The OR is easy to calculate and is useful for studies with binary outcomes, which feature heavily in the Toolkit. However, comparing the odds ratios for different interventions does not give an intuitive indication of the magnitude of the difference for a non-technical audience.
- The **Relative Risk or Risk Ratio (RR)** is the ratio between the risk of the outcome

occurring in the intervention group and the risk of the outcome occurring in the comparison group. As with the OR, an RR of less than 1 indicates a lower risk of the outcome for the intervention group while an RR of more than 1 indicates a higher risk of the outcome for the intervention group. However, the RR is generally more easily understood as it comes closer to people's intuitive understanding of probability (Cummings, 2009).

- The **Relative Risk Reduction (RRR)** is the difference between 1 and the relative risk. The RRR illustrates how much the intervention has reduced the risk of the outcome in the intervention group compared to the comparison group and helps to interpret the relative risk by indicating the magnitude of the risk reduction. The RRR is very intuitive to understand. However, the RRR puts the change in risk in the context of the baseline risk and so can over-promise the effectiveness of an intervention when the baseline prevalence of the outcome is low.
- The **Absolute Risk Reduction** (**ARR**; also known as the **risk difference**) is the risk in the comparison group minus the risk in the intervention group. It gives the percentage point difference in risk between the intervention and comparison group. The ARR provides a consistent measure of effectiveness, particularly when baseline prevalence is low (George et al., 2020). Presenting the ARR is the most useful measure of effectiveness to aid decision-making, a key consideration for the Toolkit (Irwig et al., 2008).

Place-based interventions

The evaluation of place-based interventions requires metrics specifically tailored to the unique characteristics of geographic and temporal data. Studies often assess the effectiveness of a place-based intervention by examining changes in the intervention area as compared to another area. This comparison frequently involves event counts or rates (e.g., crime incidents) across specific geographic areas and/or time periods. Traditional effect size indices, such as Cohen's d or odds ratios, often fail to capture the complexity of these count-based outcomes commonly observed in crime prevention strategies (e.g., CCTV; Wilson, 2022). To accurately estimate the effectiveness of such interventions, it is essential to use effect size measures that reflect the nature of count-based data:

- **Incident Rate Ratio (IRR):** Calculated as the ratio of post- to pre-intervention rates within a single group, IRR measures changes over time but does not account for relative changes between treatment and control groups. It is useful for within-group comparisons but lacks the broader context provided by Relative Incidence Rate Ratio.
- The **Relative Incidence Rate Ratio (RIRR)** captures the proportional change in the rate of incidents (e.g., crime) in the intervention area compared to the control area, accounting for pre-intervention baseline differences. This approach ensures that variations due to differences in time periods, population sizes, or geographic scales are appropriately addressed (Wilson, 2022). The RIRR is therefore the most appropriate measure of effectiveness for place-based approaches and is a robust and meaningful measure for evaluating the relative change in event rates.

Calculating estimates of effectiveness for meta-analyses

Person-based interventions

For person-based interventions, the meta-analysis of primary studies uses the Standardised Mean Difference (SMD). The most common SMD is Cohen's d which is simply the difference between the intervention group mean ($\overline{x_1}$) and the comparison group mean ($\overline{x_2}$), divided by the pooled standard deviation (s_{pooled}):

$$d = \frac{\overline{x_1} - \overline{x_2}}{s_{pooled}}$$

Many papers have been published to assist the calculation of the SMD from primary research (Rosnow et al., 1996; Rosnow et al., 2000). Calculating the SMD enables the transformation of many statistical tests of significance such as t-tests, F tests, and chi square values to a common metric which communicates the magnitude of the intervention effect.

SMD values should always be presented with Confidence Intervals (CI). CIs represent the range of values that encompass the true mean and can also act as a test of statistical significance. When researchers choose a 95% CI, this means that there is a 5% chance that the interval does not contain the true mean. CI is presented with two values, the first is a Lower Confidence Interval (LCI), and the second is an Upper Confidence Interval (UCI). If we think of a value of 0 as representing no effect, then if either interval includes the value of 0, then we also know the effect size is not statistically significant at the 0.05 level (assuming the CI is 95%). The p-value also communicates the statistical significance of the effect size, indicating the probability of observing that SMD value purely by chance. The traditional cut-off of $p<0.05$ indicates that the SMD value would only be observed by chance in 5% of hypothetical repetitions.

Within EPPI-Reviewer, the reviewer uses the outcomes data codeset to record the data available from the primary study. Cohen's d is then calculated alongside its variance within EPPI-Reviewer.

Place-based interventions

RIRR can be extracted directly from primary studies where available or calculated using the following formula[2]:

$$RIRR = \frac{(x_{11}/\ t_{11})}{(x_{01}/t_{01})} / \frac{(x_{00}/\ t_{00})}{(x_{10}/t_{10})}$$

Where x is the number of crimes and t is the sampling frame (generally either a time period or the population size), and the first subscript indicates the treatment condition (1 = treatment and 0 = control) and the second subscript indicates time (1 = post-test and 0 = pre-test). When t is equal across the four counts[3], the equation can be simplified to use only the counts rather than the rates, as follows:

$$RIRR = \frac{x_{11}x_{00}}{x_{01}x_{10}}$$

[2] NCB are working alongside the team at EPPI to include the RIRR calculation within EPPI-Reviewer software

[3] The simplified equation can also be used when t differs between conditions but remains the same pre and post-test; when t differs from pre to post-test but remains the same between conditions; or when the change in t is consistent in both conditions from pre to post-test.

When using count data, overdispersion is an issue that can lead to underestimation of the standard error and CIs of the RIRR. Overdispersion in crime count data arises from two sources. Firstly, crimes 'clump' together ("crime begets crime") and so are not truly independent events. For example, the same offender may commit multiple offences over a short time period, or multiple offenders may be arrested for the same crime, leading to 'clumps' of crimes rather than a random distribution of crime over time. Secondly, using a Poisson[4] regression model assumes that the probability of an event occurring is constant across a given unit of time or space. Again, this assumption does not hold for crime count data. For a meta-analysis of count data, the recommended approach to adjust for overdispersion is to use a quasi-Poisson model with an overdispersion parameter (Wilson, 2022). The overdispersion parameter can be computed from the mean rates and standard deviations provided in the primary study. This adjustment ensures accurate estimation of standard errors and CIs, reducing the risk of biased conclusions.

RIRR can also be calculated from dichotomous data if there are primary studies within the meta-analysis which have dichotomized the outcome rather than providing a count. This may be the case when most observations are either 0 or 1, with very few observations of 2 or more. For example, in a study of body-worn cameras, the outcome was complaints per police officer. As very few police officers received two or more complaints, the outcome was dichotomized into 'no complaints' and 'complaints'. For dichotomized data, the logged IRR and logged RIRR can be estimated for the meta-analysis from the difference between the two logged risk ratios (Wilson, 2022).

Finally, the IRR/RIRR can be converted to a percentage change metric using the following equation:

$$\% \, change = 100 \ \times (IRR - 1)$$

Meta-analysis of person-based and place-based interventions

Within each Toolkit strand, reviewers conduct a meta-analysis of the available evidence

[4] The Poisson distribution assumes that events happen randomly over a specified period of time and do not clump together. Applying the Poisson distribution to crime counts may violate a key assumption of the Poisson distribution as crime counts may clump together in particular time periods.

to provide an overall estimate of effectiveness for each approach. The meta-analysis uses SMD for population-based interventions and RIRR for place-based interventions to give an accurate effect for each Toolkit approach. Once data has been extracted for all primary studies, it is exported to a dataframe to facilitate advanced analyses with R[5]. The following sections outline the methodology for carrying out a rigorous meta-analysis, including accounting for dependence between effect sizes and investigating heterogeneity.

Robust Variance Estimation

The meta-analysis should account for the likely dependence between effect sizes by using **Robust Variance Estimation (RVE).** RVE is a statistical technique used in meta-analysis to address the issue of dependent effect sizes, which often arise when multiple effect sizes are extracted from the same study or related studies (Pustejovsky & Tipton, 2022). Traditional meta-analytic methods assume that effect sizes are independent; however, this assumption is frequently violated in practice, leading to underestimated standard errors and potentially misleading conclusions.

RVE provides a solution by allowing for the inclusion of all effect sizes in a single meta-regression model without requiring precise knowledge of the dependence structure among them. It employs a working model to approximate the dependence, ensuring that standard errors are adjusted appropriately to account for the correlation between effect sizes. This adjustment leads to more accurate statistical inferences, even when the exact form of dependence is unknown.

Where Robust Variance Estimation (RVE) alters the standard error or p-value relative to a naïve model, a narrative explanation should be provided. This should explain:

- Whether statistical significance changed as a result of RVE.
- Why this change matters (e.g., previously underestimated uncertainty).
- Implications for interpreting the strength of evidence.

Example: "While initial analysis suggested a highly significant reduction in violence, the

[5] R is a free, open-source programming language and software environment widely used for statistical computing and graphics.

application of RVE revealed increased uncertainty, and the revised p-value (p = 0.063) weakens confidence in this effect."

The RVE method extends earlier work on heteroskedasticity-robust and clustered standard errors in general linear models[6], adapting these concepts to the context of meta-analysis. By doing so, RVE enhances the robustness of meta-analytic findings, particularly in fields where studies often report multiple, related effect sizes. For the Toolkit, RVE is applied using *clubsandwich* and *robumeta* in R (Fisher & Topton, 2015; Pustejovsky, 2024).

Investigating Heterogeneity

Statistical heterogeneity is the variability in outcomes between primary studies that is not due to chance. Heterogeneity will always be present, but it is important to understand the amount that exists. Statistical heterogeneity can be checked in a number of ways (Higgins et al., 2003). First, visually using forest plots and checking for overlapping of CIs. Second, using tests such as the Cochran Q test (Chi-Square or χ^2), percentage of total variation across studies (I^2) and the Tau squared statistic (τ^2 or Tau^2). When using the Cochran Q test, authors often agree with the presence of heterogeneity when $p<0.1$. This figure may be chosen as it counterbalances the relatively low power of the test (in cases where there are a large number of included studies, Q is expected to be highly significant). The I^2 test represents the total variation across studies and is unlike the Q test in that it is independent from the number of studies; instead, I^2 is based on treatment effect and outcomes. The following equation from the Cochrane handbook shows how I^2 and Q are interrelated:

$$I^2 = 100\% \times \frac{(Q - df)}{Q}$$

I^2 ranges from 0-100% with 0% representing total absence of observed heterogeneity. The impact of heterogeneity was determined as low (25%), medium (50%) or high (75%) (Higgins et al., 2003). For the purposes of the Toolkit, we use the threshold of 60% to inform

[6] In general linear models, heteroskedasticity-robust standard errors adjust for non-constant variance in error terms across observations, ensuring valid statistical inferences even when this assumption is violated. Clustered standard errors account for correlations within grouped data, such as students within the same school, providing accurate standard errors by considering intra-cluster similarities.

the evidence security rating. For outcomes with an I^2 value above 60%, this is considered 'high heterogeneity' and knocks off 1 evidence security rating. For I^2 values below 60%, the evidence security rating is not penalised. Finally, τ^2 observes statistically significant heterogeneity when >1. Tau is the difference between total observed variance (Q) and within-studies variance.

If substantial heterogeneity is detected, many reviewers decide not to combine the effect sizes or present a synthesis of the findings. Others, however, investigate which study characteristics might be influencing the level of heterogeneity through techniques known as moderator analysis.

Moderator analysis

Moderator analysis is where explanations for heterogeneity are explored through analysis of certain characteristics of the study. The effect size for outcomes like offending behaviour can be influenced by factors such as intermediate outcomes (mediators) and contextual, design, or implementation aspects (moderators). Understanding these moderators helps users decide if an intervention is suitable for their context and highlights critical considerations for its design and implementation.

For the Toolkit, information is extracted from studies on the moderators tested and their results, focusing on:

- **Study -level moderators:** type of publication, quality appraisal as assessed by the YEF-EQA tool, conflict of interest
- **Quality moderators:** Study design, comparison type, quality appraisal as assessed by the YEF-EQA tool
- **Intervention-level moderators:** Whom the intervention was targeted at, implementation features (e.g., duration, intensity, location).
- **Outcome moderators:** Analysis type, data type, type of outcome or domain (e.g., crime vs. wellbeing).
- **Higher-level outcome groupings** (where appropriate): mapping outcome types to the higher level YEF Outcomes Framework categories.

These analyses help contextualise effect variation and strengthen narrative conclusions about effectiveness. Moderator analysis can be handled in a way that is analogous to the one-way analysis of variance (ANOVA)[7] ('subgroup analysis'); or analogous to linear regression[8] in primary research ('meta-regression'). The decision of which type of moderator analysis to use will often depend on the type of characteristics available.

It is important to understand that both types of moderator analyses are exploratory and should never be implemented to test hypotheses. Even if the meta-analysis contains only random and quasi-random trials, the studies involved in these moderator analyses have not been randomised, they are observational in nature and at a higher risk of bias. Additionally, as these types of analyses generally have lower power due to missing data in the primary research, there is an increased risk of presenting incorrect results which appear simply through chance (false positive conclusion), and potential for various biases.

Subgroup analysis

Subgroup analysis calculates the SMD or RIRR within each subgroup and then compares effectiveness and heterogeneity with the other subgroups in the category. Subgroup analysis presents details about the variance within the subgroups (Qw) which is unexplained, and the variance between the subgroups (Qb), and whether those differences are statistically significant.

For subgroup analysis for the Toolkit, the following ten factors are explored:

1. **Conflict of interest.** Classify studies into two categories, the first for studies where either the author or the reviewer identified a potential conflict of interest was likely, and the second for studies where a conflict of interest was deemed unlikely.
2. **Publication type.** Categorize studies as either published or non-published. 'Non-published' studies include those which are not available commercially and/or have not been through the peer review process.
3. **Risk of Bias.** Categorize studies based on their overall risk of bias as assessed

[7] One-way analysis of variance is a statistical method used to compare the means of two or more groups to determine if there is a statistical difference between them.
[8] Linear regression is a statistical technique that uses a linear equation to predict the value of a variable based on the value of another variable.

using the YEF-EQA framework. Assign each study a rating of high, medium, low, or very low, reflecting the level of confidence in the reliability and validity of the study's findings.

4. **Training for Intervention Providers.** Categorize studies according to whether individuals delivering the intervention received specialized training or not.
5. **Geographic location of study.** Categorize studies according to their geographic location. The precise categorization depends on the geographic spread of studies within the Toolkit strand, for example, studies may be classified by continent where there is a large geographic spread, or by country or region.
6. **Gender.** Classify studies as male-targeted if the majority of participants (more than 50%) are male, female-targeted if the majority of participants (more than 50%) are female, or gender-balanced if there is an approximately equal representation of male and female participants.
7. **Ethnicity.** Classify studies as majority group-targeted if the majority of participants (more than 50%) are from the majority ethnic group, minority group-targeted if the majority of participants (more than 50%) are from a minority ethnic group, or ethnically diverse if there is a balanced representation of ethnic groups.
8. **Experience of deprivation.** If data allows, categorize studies as either low, middle, high, or equal, depending on the predominant socioeconomic background of the sample participants.
9. **SEND status.** Classify studies as SEND-targeted if the majority of participants were identified as having SEND and as SEND-inclusive if SEND participants were included but did not constitute the majority.
10. **Care-experienced.** Classify studies as care-targeted if the majority of participants had spent time in formal care (e.g., foster care or residential care), and as care-inclusive if participants with care experience were included but did not constitute the majority.

Meta-regression

Meta-regression extends subgroup analysis by allowing the simultaneous examination of multiple variables, including continuous ones such as mean age. The categorical variables listed above can also be entered into the model as a series of dummy variables. In a meta-regression, the outcome variable is the SMD or RIRR and the characteristics extracted are the predictors or criterion variables. A meta-regression analysis can be represented by a simple scatter plot, with the variable of interest presented along the x-axis, and the SMD along the y-axis. The regplot function in metafor (Viechtbauer, 2010) also allows the precision of each primary research to be proportional to the size of the plotting symbols provided. In addition to testing the statistical significance of the potential moderators on the SMD, it is also important to test the fit of the model using the coefficient of determination, also known as the R^2 index. This index calculates the proportion of the variance of the SMD or RIRR that is explained by the meta-regression model and covariates chosen to test.

Sensitivity analysis

Sensitivity analysis is a repetition of the primary meta-analysis using alternative decisions or assumptions to determine if the results are consistent. The goal is to ensure that the conclusions are not unduly influenced by specific studies or methodological choices. This ensures that the review is robust and prevents individual studies from exerting an undue influence on findings. Process sensitivity analysis is carried out using either the SMD (for population-based approaches) or the RIRR (for place-based approaches) alongside their heterogeneity statistics. These analyses are recalculated using the metafor package in R to systematically explore how different analytical choices impact the overall findings, ensuring that the results are robust and not overly dependent on specific assumptions or individual studies. The exact sensitivity analyses will vary according to the underlying data within each Toolkit strand, but some examples include removing studies with outliers of extreme effects, and removing studies rated as ‘low or very low quality’ based on the YEF-EQA.

Publication bias

Publication bias most simply refers to the tendency for studies with negative effects or non-statistically significant findings to remain unpublished (Rosenthal, 1979). To mitigate the effects of publication bias, various grey literature sources are included in searches for the EGM to ensure a fully comprehensive Toolkit strand.

In meta-analyses, publication bias describes the phenomenon whereby studies with stronger positive effects are more likely to be published in peer-reviewed journals and, consequently, included in meta-analyses (Quintana, 2015). This phenomenon, known as the file-drawer problem, is described as: "journals are filled with the 5% of the studies that show Type 1 errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > 0.05$) results" (Rosenthal, 1979, p. 638). When studies are collated, meta-analysis techniques enable the detection of publication bias using several methods.

One way to visually assess publication bias is through a funnel plot (Egger et al., 1997). A funnel plot is a scatterplot that plots each included study according to the effect size[9] and standard error. If a meta-analysis lacks a representative number of studies with small sample sizes and statistically insignificant results (and thus less likely to be published), the funnel plot appears asymmetric, indicating publication bias, as shown in Figure 4. In contrast, when a meta-analysis includes a representative number of studies with sufficient sample sizes and statistically significant findings, the funnel plot appears symmetrical, indicating no publication bias, as shown in Figure 5.

[9] We will use SMD for illustrative purposes but other effect sizes like RIRR can be used interchangeably with these methods.

Figure 4. Funnel plot from a meta-analysis with publication bias

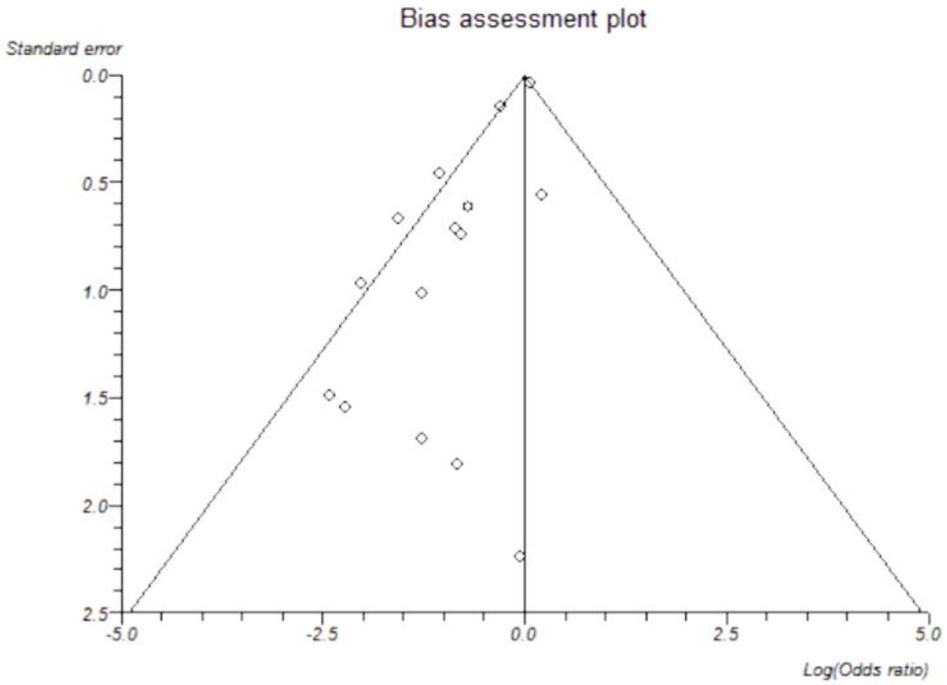
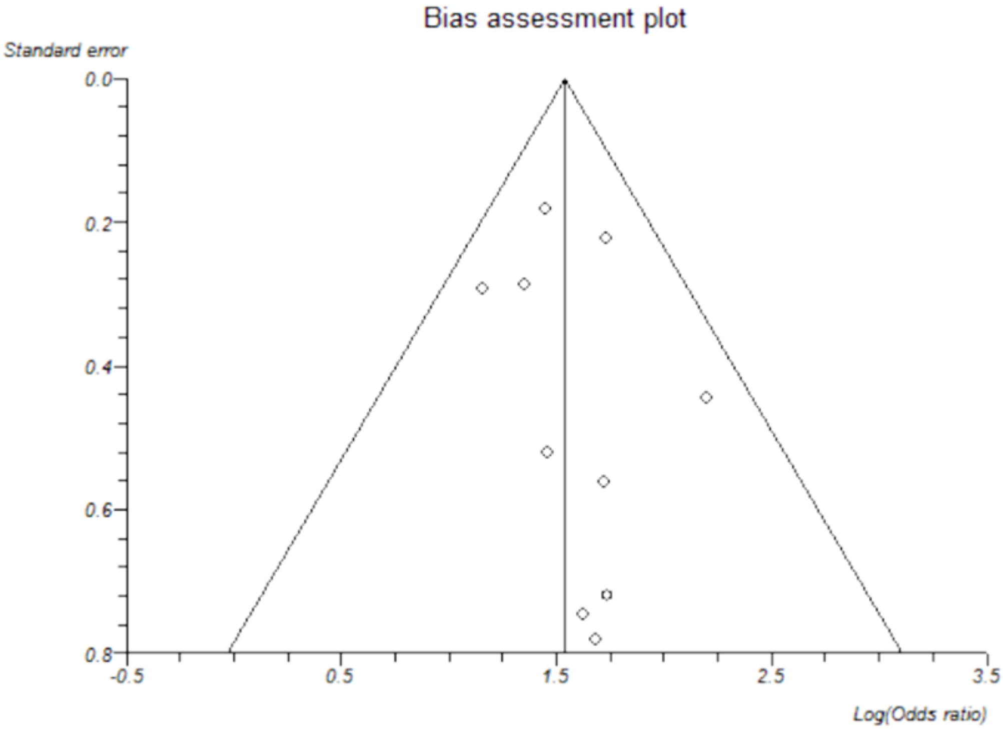


Figure 5. Funnel plot from a meta-analysis without publication bias



As funnel plots are visual and therefore subjective, several regression tests are typically reported alongside them. One test for funnel plot asymmetry is the rank correlation between the standard error and the SMD (Begg & Mazumdar, 1994). If the rank correlation test reports statistical significance ($p < 0.05$), it indicates asymmetry in the funnel plot. Another test, Egger's regression method, investigates a linear relationship between the standard error and the SMD (Egger et al., 1997). Statistical significance ($p < 0.05$) from Egger's regression also suggests asymmetry. Egger's regression is considered more powerful than the rank correlation test for detecting publication bias in meta-analyses with fewer than 30 studies (Sterne et al., 2000).

If the analyses described above indicate publication bias within a Toolkit strand, the trim and fill method (Duval & Tweedie, 2000) is applied. This nonparametric technique removes smaller studies, re-estimates the combined SMD (or RIRR), and imputes missing studies to improve funnel plot symmetry (Higgins & Green, 2011). However, the trim and fill method has limitations, including underestimating the combined SMD (or RIRR) when study heterogeneity is substantial (Peters et al., 2007) and correcting for publication bias when it does not exist (Terrin et al., 2003).

For meta-analyses within Toolkit strands, a funnel plot is included to visually assess publication bias across included studies (Sterne et al., 2000). The rank correlation test (Begg & Mazumdar, 1994) and Egger's regression test (Egger et al., 1997) are also employed. Where asymmetry in the funnel plot and statistical significance in these tests are observed, this indicates either publication bias or a tendency for smaller studies to show larger treatment effects. Where publication bias is detected, Rosenthal's fail-safe N (1979) estimates the number of additional studies required to affect the significance of the intervention effects, and the trim and fill method provides a graphical imputation of potential missing studies. This information is all communicated to the reader through the Technical report.

Communicating effectiveness

Meta-analysis results are presented in the Technical Report for each Toolkit strand. The Toolkit also presents several other effectiveness estimates to help communicate impact to a non-technical audience and to help end users compare between approaches.

Person-based interventions

YEF impact rating

The YEF impact rating provides a simple, clear indicator of the relative impact of the approach on violent crime, allowing users to quickly compare different approaches. The YEF impact rating is calculated from the Cohen's d statistic derived from the meta-analysis. Where the meta-analysis includes multiple different outcome types, the impact rating is based on the outcomes considered to be a 'primary outcome' by YEF, favoring outcomes that directly measure involvement in violence or crime.[10] Once the headline outcome is chosen, the Cohen's d value can be converted into the YEF impact rating using Table 1. The intervals for the YEF impact rating are based on the distribution of effect sizes for reviews included in the YEF Toolkit and studies included in the YEF Effect Size Database. Studies were grouped according to effect size into three roughly equal groups, which correspond to the small, moderate and high effect categories (only one review reported a 'harmful' effect; YEF, 2021). The 'no effects' category includes very small negative and positive effects. The rationale for classifying these effect sizes as having 'no effect' is that the number of people who would need to receive the intervention to achieve a positive outcome or to avoid a negative outcome is very large. For example, a program might need to work with hundreds of young people to prevent one young person committing a crime.

The YEF impact rating calculated here is the 'raw' impact rating and may be downgraded during the evidence security assessment process, outlined in the next chapter.

[10] See the YEF Outcomes Framework for definitions and examples of primary outcomes, secondary outcomes, and contextual factors: https://youthendowmentfund.org.uk/outcomes/]

Table 1. Conversion of Cohen's d to YEF impact rating

Cohen's *d* from meta-analysis	YEF impact rating
d ≤ -0.02	Harmful
-0.02 > d < 0.02	No effects
0.02 ≥ d < 0.10	Low impact
0.10 ≥ d < 0.25	Moderate impact
d ≥ 0.25	High impact

NB In this table, the start point of each given range is part of the relevant effect size, while the end point marks the start of the next interval

Relative and absolute risk reduction

The Toolkit uses 'risk' as the basic concept to communicate effectiveness. 'Risk' is the likelihood of an outcome occurring for a given group. It is calculated as the proportion of the group who have the outcome, for example, the proportion of young people who commit a crime, where a is the number of young people who commit a crime and b is the number who do not commit a crime:

$$Risk = \frac{a}{a+b}$$

To calculate risk reduction based on the results of the meta-analysis, first convert Cohen's *d* into an Odds Ratio and subsequently derive a 2x2 table to obtain the number of young people with and without the outcome in the intervention group and in the comparison group. Cohen's d can be converted into an Odds Ratio using the formula provided in Lipsey & Wilson, 2001:

$$\ln(OR) = \frac{d}{0.5513}$$

Alternatively, the online calculator provided by the Campbell Collaboration can be used

(Wilson, 2023).[11] Table 2 illustrates the information required to calculate an odds ratio.

Table 2. 2x2 table for calculating an odds ratio

	No outcome	**Outcome**	**Total**
Intervention	a	b	$a + b$
Comparison	c	d	$c + d$

Using the odds ratio to calculate the numbers in each group who do or do not have the outcome requires two assumptions:

1. Equal numbers (n=100) in intervention and comparison groups
2. The prevalence of different outcomes in the comparison group

The baseline estimates for (2) used by YEF are listed below and justified in Box 1:

- Involvement in crime: 25% prevalence
- Reoffending: 50% prevalence
- Involvement in violence: 15% prevalence
- Violence recidivism: 29% prevalence

[11] https://www.campbellcollaboration.org/calculator/

Box 1: Justification of baseline prevalence estimates

Offending

Still providing the best available evidence on offending rates, data from the longitudinal Cambridge Study in Delinquent Development (CSDD; Farrington, 2012; Farrington & Malvaso, 2023) was used to justify prevalence estimates. According to the CSDD, 25% of 411 London boys aged 10-17 years received a conviction for offending.

Reoffending

According to the Youth Justice Statistics 2022-2023 for England and Wales (Youth Justice Board, 2024), the proven reoffending rate for CYP after one year is 32.2%. As such, this figure can be expected to increase to approximately 50% within three years. The reoffending baseline estimate of 50% is consistent with past research (see Wilson et al., 2018).

Violence

YEF (2023) surveyed over 7500 13 to 17-year-olds in England and Wales, finding that 15% had been involved in violence over the past 12 months.

Violent Recidivism

According to the Youth Justice Statistics 2022-2023 for England and Wales, the proven reoffending rate for CYP with an index offence relate to violence was 29%.

Tables 3 to 6 illustrate the 2x2 tables that can be derived from these assumptions.

Table 3. 2x2 table for calculating odds ratios for intervention targeting involvement in crime

	Not involved in crime	Involved in crime	Total
Intervention	100 - x	x	100
Comparison	75	25	100

Table 4. 2x2 table for calculating odds ratios for intervention targeting reoffending

	Don’t reoffend	Reoffend	Total
Intervention	100 - x	x	100
Comparison	50	50	100

Table 5. 2x2 table for calculating odds ratios for intervention targeting involvement in violence

	Not involved in violence	Involved in violence	Total
Intervention	100 - x	x	100
Comparison	85	15	100

Table 6. 2x2 table for calculating odds ratios for intervention targeting violence recidivism

	Don’t commit violence again	Commit violence again	Total
Intervention	100 - x	x	100
Comparison	71	29	100

By re-arranging the formula for the odds ratio, x can be calculated as follows:

$$x = \frac{a + b}{\frac{d}{c} \times OR + 1}$$

The relative risk reduction (expressed as a percentage) is then:

$$Relative\ Risk\ Reduction = \frac{(d - x)}{d} \times 100$$

Sensitivity testing should be carried out to test a range of prevalence estimates for the relevant outcomes to understand how different assumed prevalence rates affect the relative risk reduction.

The absolute risk reduction can also be calculated from the derived 2x2 table as follows:

$$Absolute\ Risk\ Reduction = (d - x)\ \times 100$$

The absolute risk reduction provides a clear measure of how much an intervention reduces risk in absolute terms (rather than as a proportion of the underlying risk) which can provide a more realistic picture of an intervention's impact for decision-makers.

Place-based interventions

YEF impact rating

Applying Cohen's d to count-based data in place-based interventions is methodologically unsound, as d is designed for continuous data and measures the standardized difference between two means. This approach can lead to inaccuracies when applied to aggregated rates or counts. To address this, the Toolkit uses Wilson's RIRR, which is tailored to count data and ensures accurate analysis of an intervention's impact. An RIRR of less than one indicates that the intervention is beneficial while an RIRR greater than one indicates that the intervention is harmful.

To ensure impact ratings can be presented consistently across both place-based and person-based approaches in the Toolkit, the log-transformed RIRR is aligned with Cohen's d to give a YEF impact rating.[12] This is based on Wilson (2022) which describes the proportional interpretive power of log-transformed RIRR.

Step 1: Define thresholds

RIRR > 1.05: Harmful effects

RIRR < 1.05 and > 0.95: No effect

RIRR < 0.95 and > 0.78: Small effects

RIRR < 0.78 and > 0.35: Moderate effects

[12] Thresholds for interpreting raw and log-transformed RIRR values were informed by consultation with Professor Dave Wilson, whose expertise contributed to refining the presentation of the methodology.

RIRR < 0.35: Large effects

Step 2: Align with Standardised Mean Differences (SMD)

Table 7. RIRR thresholds aligned with Cohen's d

Raw RIRR values	Log RIRR Values[13]	Cohen's d*	YEF impact rating
>= 1.05	>= 0.049	≤-0.02	Harmful: implies that the rate of the event of interest (e.g., crime) is 5% higher in the treatment group relative to the control group, (or the intervention led to a 5% increase in incident rates).
< 1.05 and >= 0.95	< 0.049 and >= -0.05	-0.02 < d < 0.02	No effects: indicates a neutral zone with changes in incident rates not exceeding ±5%.
< 0.95 and >= 0.78	< -0.05 and >= -0.24	0.2 ≤ d < 0.1	Small effects: indicates a 10% decrease in incident rates
<0.78 and >= 0.35	< -0.24 and >= -1.05	0.1 ≤ d < 0.25	Moderate effects: reflecting a 10% to 25% decrease in incident rates
< 0.35	< -1.05	d 0.1 ≤ 0.25	Large effects: indicating decrease exceeding 26% in incident rates

*Approximate alignment

[13] Log RIRR values are natural logarithms of the corresponding RIRR values.

Step 3: Calculation of the percentage change

The formula for the percentage change is:

$$\%\ change = 100\ \times (IRR - 1)$$

An example RIRR of 0.87 corresponds to a log-transformed RIRR of approximately -0.139, indicating a 13%[14] reduction in incident rates relative to the control group. Using the proposed thresholds, this would fall into the "Small" impact category.[15]

Step 4: Report that it is an approximate measure.

To ensure transparency, in the Technical Report, all results should explicitly indicate that they are based on log-transformed RIRR values, alongside the percentage change and category of impact.

Reporting on effectiveness

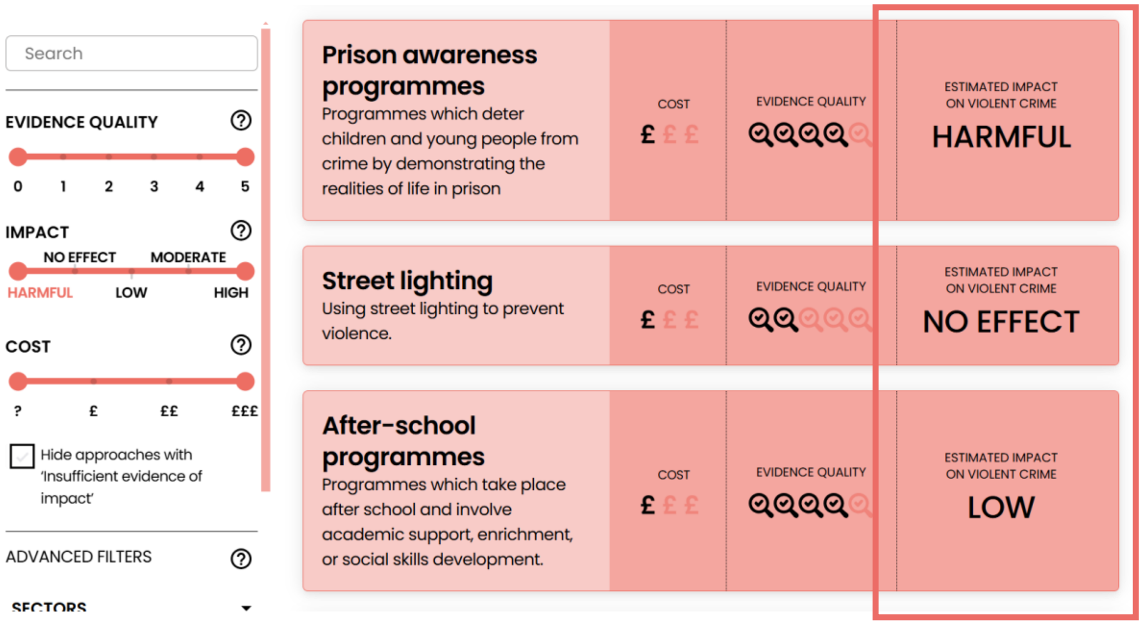
Toolkit front page

The Toolkit front page clearly displays the YEF impact rating on the right-hand side (see Figure 6). Users can also sort and filter according to the impact rating.

Figure 6. Snapshot of the YEF Toolkit front page showing the impact rating on the right-hand side

[14] Based on based on Wilson's equation: % change = 100 × (IRR - 1)

[15] Transforming the RIRR into Cohen's d without the steps above would correspond to an approximate Cohen's d of -0.077 and therefore categorised as 'harmful' under current thresholds. This would be methodologically inappropriate as it does not account for the fact that these measures are derived from different statistical models and data types



□ Hide approaches with 'Insufficient evidence of impact'

	COST	EVIDENCE QUALITY	ESTIMATED IMPACT ON VIOLENT CRIME
Prison awareness programmes Programmes which deter children and young people from crime by demonstrating the realities of life in prison	£	4/5	HARMFUL
Street lighting Using street lighting to prevent violence.	£	2/5	NO EFFECT
After-school programmes Programmes which take place after school and involve academic support, enrichment, or social skills development.	£	4/5	LOW

Summary page

The summary page should clearly state the number of studies used to calculate the impact rating that were conducted in the UK or Ireland.

Under the heading 'Is it effective?', the Summary Page provides an overview of the evidence on the intervention's effectiveness. A statement summarises the relative risk reduction, as outlined below:

> *"The evidence suggests that [intervention name] may reduce the risk of [outcome] among participants by [% reduction], compared to baseline violent offending in England and Wales."*
>
> For example, *"the evidence suggests that A&E navigator programmes may reduce the risk of reoffending among participants by 38.5%, compared to baseline violent offending in England and Wales."*

The Summary Page also outlines any results of note from the moderator or sub-group analysis carried out as part of the meta-analysis, to highlight how the effectiveness of the intervention may vary according to contextual factors or participant characteristics.

Technical report

The 'impact' section of the Technical Report details the evidence and methods used to produce the YEF impact rating. The report should include a list of the primary studies used (summarised in the table template in **Error! Reference source not found.**, and a Table l isting all outcomes from the meta-analysis with effect sizes with standard error, 95% CIs, and *p*-values (Table 8 and Table 9).

Table 8. Template table for reporting effect sizes for person-based interventions

Outcome name	Cohen's d (SE)	95% confidence intervals	p-value	Odds ratio	'Raw' YEF impact rating

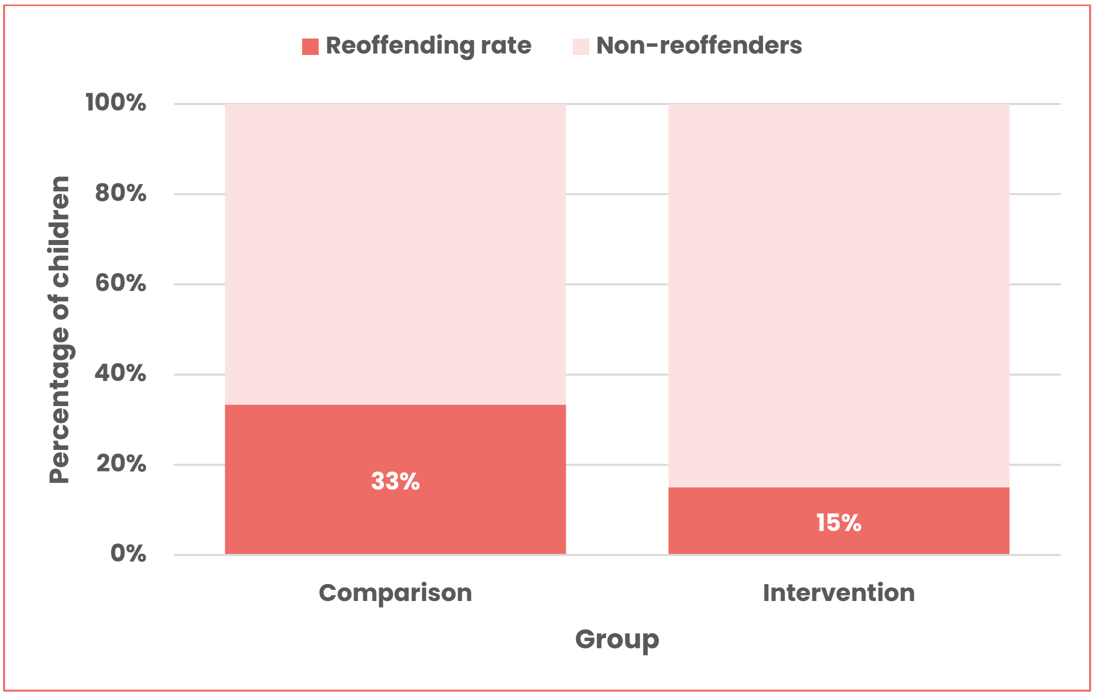
Table 9. Template table for reporting effect sizes for place-based interventions

Outcome name	RIRR	95% confidence intervals	Log RIRR	Cohen's d	'Raw' YEF impact rating

For each outcome, the Technical Report should also show the 2x2 table and calculation used to calculate the Relative Risk Reduction and the Absolute Risk Reduction. The heterogeneity section presents the I^2 value and discusses the implications for the meta-analysis. Where I^2 is greater than 60%, tau^2 is also presented. The section 'moderators and mediators' detail the moderator analyses conducted (sub-group analysis and/or meta-regression and sensitivity analyses) and summarises the results. The Technical Report should include an appendix presenting a detailed breakdown of the meta-analysis results, including forest plots and moderator analyses.

The Technical Report features a chart to help end users visualise the intervention's impact in terms of the Absolute Risk Reduction. The chart is a stacked bar chart with two bars comparing the outcome prevalence in the comparison group with the prevalence in the intervention group (Figure 7[16]). The report should state the caveat that these charts are not directly comparable between approaches as the outcomes and axes may differ.

Figure 7. Example of a chart for the YEF Toolkit summary page showing the absolute risk reduction



5. Evidence security rating

Assessing the security of Toolkit evidence

The Toolkit intends to show ‘best bets’ for reducing violence. It gives an overall picture of the approaches that are most likely to succeed, given the available research. The impact rating outlined in the last section provides one piece of this puzzle, but the security of the evidence must also be considered when communicating how promising an approach may be. An intervention that has a high impact rating that is based on several high-quality studies will be a much ‘better bet’ than an intervention with the same impact rating, but which is based on only a few studies, or on lower-quality studies. The evidence

[16] Data shown in Figure 7 is only used as an example, it is likely that these will change in practice.

security rating, therefore, communicates the strength of the evidence that has contributed to the impact rating and indicates how confident the end user can be that the intervention will have the estimated impact.

Calculating the impact rating for each area of the Toolkit is based on a meta-analysis using a robust evidence synthesis methodology that ensures methodological quality. AMSTAR 2 (A MeaSurement Tool to Assess Systematic Reviews 2; Shea et al., 2017) is a critical appraisal tool designed to evaluate these methodological standards. The EGM review process addresses each of the AMSTAR 2 criteria as follows:

1. **Protocol Registration.** A detailed protocol outlines the pre-determined objectives, eligibility criteria, and methods for the EGM (the foundation for all meta-analyses carried out for the Toolkit). To promote transparency this protocol is registered on Open Science Framework (OSF).[17]

2. Comprehensive searching. Utilizing OpenAlex[18] allows for the implementation of a thorough search strategy across multiple sources, ensuring the identification of all relevant studies. The search process is meticulously documented, including search terms and search dates, to ensure reproducibility.

3. Study selection and data extraction by multiple reviewers. By incorporating machine learning algorithms, these tasks can be semi-automated, with the machine acting as a second reviewer. Machine learning models can be trained to screen studies for inclusion and extract important data, thereby expediting the process. Discrepancies between human and machine assessments are resolved through discussion or consultation with a third reviewer. This method maintains the integrity of the review while enhancing efficiency.

4. Noting exclusion decisions. Reviewers provide clear reasons for the exclusion of any studies at the full-text screening stage via the EPPI-Reviewer coding tools. Due to the sheer number of records, this is effectively communicated using a PRISMA flow diagram,

[17] [Placeholder link to protocol on OSF]

[18] OpenAlex aims to be a comprehensive repository of the World's research and currently contains more than 200 million records. The OpenAlex database updates every few weeks and a copy of the dataset is made available within EPPI-Reviewer. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3754; https://openalex.org/

detailing the number of studies excluded and the reasons for exclusion via categories.

5. Risk of Bias assessment. All systematic reviews in the EGM receive an AMSTAR 2 rating, while all primary studies used to calculate the impact rating are appraised using the YEF-EQA which includes an item on risk of bias in individual studies.

6. Reporting funding sources. The team extracts information on funding sources for each included study to assess potential conflicts of interest that may influence study outcomes.

7. Assessing Risk of Bias during analysis. During the meta-analysis, moderator analysis is used to assess the influence of a study's quality on the outcome. Sensitivity analyses determine how the exclusion of studies at high risk of bias affects the overall results (for further detail see 'Investigating heterogeneity' section above).

8. Assessing heterogeneity. Reviewers investigate and report on the heterogeneity among study results. The Toolkit utilizes statistical methods, such as the I^2 statistic, and explores potential sources of heterogeneity through subgroup analyses or meta-regression, if applicable.

9. Appropriate meta-analysis. The Toolkit Technical Guide has been produced through consultation with experts, ensuring the statistical methods used are appropriate. Robust variance estimation, one of the most advanced meta-analytical techniques available to date, is used to ensure the analysis accounts for dependence between effect sizes. Limitations of the data (for example, due to missingness or small samples), are clearly communicated within the Technical Report.

10. Assessing publication bias. Reviewers evaluate the potential for publication bias using methods like funnel plots or statistical tests when sufficient studies are available and discuss the implications of any detected bias on the review's conclusions (for further detail see 'Publication bias' section above).

11. Declaration of Conflict of Interest. There are no potential COIs among the review authors and the team reports the funding sources in each Toolkit Report to maintain transparency.

By systematically addressing each of these criteria, the review process for the EGM and each subsequent Toolkit strand will align with the standards set by AMSTAR 2, thereby enhancing the credibility and reliability of the findings.

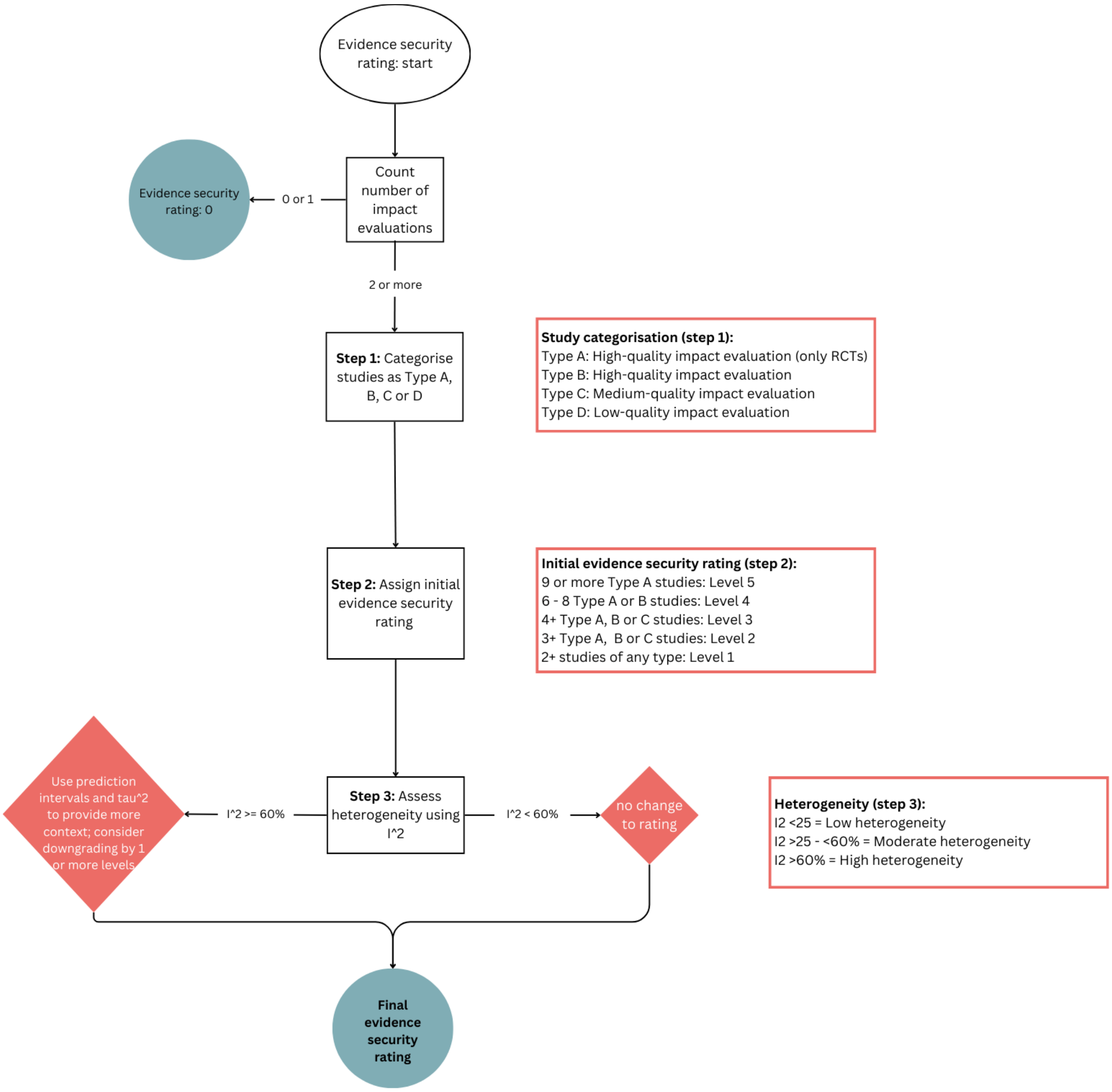
Evidence security rating process

How We Rate Evidence Security

This process begins by assessing the type and quality of impact evaluations available. The more high-quality studies, the higher the initial rating. Evidence security may be adjusted based on how consistent the results are across studies. If results vary widely and the differences are unexplained, this may lead to a lower rating. This process helps users see how much trust they can place in the findings.

The evidence security rating shows how confident we can be in the impact rating for each Toolkit strand. It is based on the number and quality of impact evaluations, as well as how consistent the findings are across the included studies (measured via heterogeneity). See **Error! Reference source not found.** for a summary of the evidence security rating process.

Figure 8. Flowchart summarising YEF evidence security rating process


Evidence security rating: start
Count number of impact evaluations
0 or 1
Evidence security rating: 0
2 or more
Step 1: Categorise studies as Type A, B, C or D
Study categorisation (step 1):
Type A: High-quality impact evaluation (only RCTs)
Type B: High-quality impact evaluation
Type C: Medium-quality impact evaluation
Type D: Low-quality impact evaluation
Step 2: Assign initial evidence security rating
Initial evidence security rating (step 2):
9 or more Type A studies: Level 5
6 - 8 Type A or B studies: Level 4
4+ Type A, B or C studies: Level 3
3+ Type A, B or C studies: Level 2
2+ studies of any type: Level 1
Step 3: Assess heterogeneity using I^2
I^2 >= 60%
Use prediction intervals and tau^2 to provide more context; consider downgrading by 1 or more levels
I^2 < 60%
no change to rating
Heterogeneity (step 3):
I2 <25 = Low heterogeneity
I2 >25 - <60% = Moderate heterogeneity
I2 >60% = High heterogeneity
Final evidence security rating

The first step categorizes studies based on their study design and their quality. Impact evaluations refer to studies that assess the impact of an intervention by comparing the outcomes for an intervention group with those for an appropriate comparison group. When participants are allocated randomly to either an intervention or comparison group, the study is an RCT. When participants are not allocated randomly but other methods are used to ensure the comparison group is similar to the intervention group, the study is 'quasi-experimental'. Due to the potential bias introduced by non-random allocation, quasi-experimental studies are not considered to have the same strength of evidence as high-quality RCT's, which are the 'gold standard' for measuring impact. The quality of primary studies used by the meta-analysis to calculate the effect size is assessed using the YEF Evidence Quality Assessment Tool (YEF-EQA). Each primary study is rated as either:

- **Type A:** High-quality randomised controlled trial (RCT)
- **Type B:** High-quality quasi-experimental study
- **Type C:** Moderate-quality RCT or quasi-experimental study with minor limitations
- **Type D:** Low-quality RCT, quasi-experimental study or PPD study with major limitations

The second step in the Evidence Security Rating process involves assigning an initial Evidence Security level from 5 to 1. The highest level that can be achieved for Type D is Level 1, the highest for Type C is Level 3, the highest for Type B is Level 4, and the highest for Type A is Level 5 (Table 10). However, it is important to note that there are Toolkit strands where RCTs (Type A) evidence is not possible due to practical or resource constraints. For example, it would be unethical to randomly assign young people to pre-court diversion.[19]

[19] This is sometimes referred to as the "equipoise" dilemma. In this example, randomly assigning young people to pre-court diversion versus traditional punitive measures would mean some individuals are denied access to an intervention that could positively affect their future outcomes. This could be perceived as unfair or harmful.

Table 10. Initial Evidence Security Rating (Steps 1 and 2)

Level	Number of studies	Types of studies	
Level 5	Nine+	high-quality	Impact evaluations (type A)
Level 4	Six to eight	high-quality	Impact evaluations (type A or B)
Level 3	Four or five	high-quality	Impact evaluations (type A or B)
	8+[20]	moderate quality	Impact evaluations (type C)
Level 2	Three+	any quality	Impact evaluations (type A, B or C)
Level 1	Two or more	any quality	Impact evaluation (type A, B, C or D)

The third step of the Evidence Security Rating assesses the level of heterogeneity in meta-analyses. When heterogeneity is high, the effect sizes from the different primary studies vary widely, giving us less confidence in the combination of results. When heterogeneity is low, the effect sizes from the different primary studies are similar and there is more confidence in the synthesis of results.

Quantifying Heterogeneity

During the meta-analysis, heterogeneity is measured in the following ways:

- **I^2 Statistic:** Represents the percentage of total variation across studies due to heterogeneity rather than chance.

[20] In earlier versions, any number of Type C studies were suggested as eligible for Level 3. Based on further review, this has been revised to reflect that a higher threshold (e.g., 8+) of moderate-quality studies is required to achieve the same confidence level as fewer high-quality studies.

- **Tau² (τ^2):** Estimates the between-study variance in a random-effects meta-analysis.
- **Prediction Intervals:** Provide a range in which the true effect size of a new study is expected to fall, considering existing heterogeneity.

Table 11. Guidelines for Downgrading Based on Heterogeneity

I^2 Value	τ^2 Value	Interpretation	Action
< 25%	≤ 0.01	Low heterogeneity	No downgrade
26%–59%	0.01–0.05	Moderate heterogeneity	Consider downgrading by one level if prediction intervals are wide or cross thresholds of clinical importance.
>60%	>0.05	High heterogeneity	Consider downgrading by one or two levels, especially if heterogeneity affects the robustness of conclusions.

While I^2 can be calculated from only two studies, the power to detect and accurately quantify heterogeneity with a small number of studies is very limited. As the number of studies included increases, the reliability of heterogeneity estimates improves (von Hippel, 2015). Therefore, the following guidance should apply:

Two Studies: Avoid downgrading solely based on I^2 or τ^2; instead, discuss the limitations of heterogeneity assessment in the report.

Three to Four Studies: Report I^2 and τ^2 with confidence intervals; consider downgrading if prediction intervals indicate substantial inconsistency.

Five or More Studies: Apply standard downgrading criteria as outlined in Table 11 above.

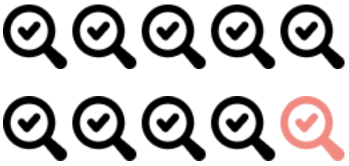
Where sufficient studies are available, **meta-regression** may also be used to explore the causes of heterogeneity by examining how study-level characteristics, such as study quality, intervention duration, or setting, influence effect sizes. This analysis can help identify whether specific factors are systematically associated with larger or smaller effects, supporting more nuanced interpretation of the findings. However, meta-regression should be used cautiously, particularly when the number of studies is limited. It is generally recommended that at least ten studies contribute to each covariate tested, to reduce the risk of spurious associations. Findings from meta-regression should be considered exploratory and interpreted as part of a broader narrative about the robustness and consistency of the evidence base.

Reporting on evidence security

The Toolkit includes an indicator of how much confidence the user can have in the impact rating based on the strength of the evidence base available. The strength of evidence considers the amount of evidence in terms of the number of studies, the quality of the underlying studies, and the quality and heterogeneity of the evidence base.

- **Top level (Toolkit Front Page):** The evidence security rating (scored out of five) is communicated visually through the number of magnifying glasses (see Figure 9).
- **Second level (Summary Page):** Under the heading, "How secure is the evidence?", a summary sentence indicates how confident we can be in the impact rating (see examples in Figure 10 and 10). This section then summarises the evidence security for any other outcomes and the reasons for the evidence security rating.
- **Third level (Technical Report):** The report should detail the calculation of the evidence security rating for all effect sizes with a clear rationale for the final rating. The report should also present the quality appraisal rating for all included studies, considering how the quality of the studies may impact on our interpretation of the effectiveness and implementation of the approach.

Figure 9. YEF evidence security rating system



= We have very high confidence in this impact rating

= We have high confidence in this impact estimate

●●●○○ = We have moderate confidence in this impact estimate

●●○○○ = We have low confidence in this impact estimate

●○○○○ = We have very low confidence in this impact estimate

○○○○○ = There is not enough research to create an impact estimate*

** 0 magnifying glasses indicates that the toolkit approach consists only of qualitative evidence or systematic reviews of implementation evidence*

Figure 10. Examples of evidence quality ratings

Focused deterrence is rated 4 out of 5 for evidence quality. We can be confident that a focused deterrence approach will have a high impact on violent crime.

Trauma-specific therapies is rated 1 out of 5 for evidence quality. We cannot be confident that trauma-specific therapies will have a high impact on violent crime without further high-quality evidence.

After-school programmes are rated 4 out of 5 for evidence quality. We can be confident that after-school programmes will have a low impact on violent crime.

CCTV is rated 2 out of 5 for evidence quality. We have low confidence that CCTV will have a low impact on violent crime without further high-quality evidence.

Cognitive Behavioural Therapy (CBT) is rated 3 out of 5 for evidence quality. We have some confidence that CBT will have a high impact on violent crime but require further high-quality evidence.

6. EDIE Section

Previous work has shown that a number of personal characteristics are associated with greater risk of involvement in violent crime, including being male, being from a minority ethnic group, having a special educational need or disability or a social-emotional or mental health need, being care experienced, being from a deprived neighbourhood, and having lower educational attendance (e.g., McAra & McVie, 2016). It is therefore important to highlight any available information on the effectiveness of interventions specifically for these groups. Additionally, the Toolkit should consider how generalizable the evidence is to the UK context, specifying the proportion of studies conducted in the UK and/or Ireland and highlighting quality evidence from UK studies. To specify, the concepts associated with EDIE are clearly defined in Table 12.

Table 12. EDIE key definitions

Concept	Definition
Equality	Same resources and opportunities are available to individuals or groups
Diversity	Recognizes that differences between people (including, but not exclusively, protected characteristics) should be valued, respected and promoted
Inclusion	Positive action is taken to ensure practices for including people are fair for all, with people feeling empowered and enabled to be themselves
Equity	Resources and opportunities are allocated to individuals based on their specific circumstances, enabling equal outcomes for all

In this report, the following terminology is used:

- Care-experienced young people
- Young people with special educational needs and disabilities (SEND)
- Minority ethnic young people, or where possible the name of specific ethnic

groups, (e.g., Black Caribbean young people)

Where citing authors of other studies, they may have used different terminology. As such, this will need to be cited as reported by other authors[21]. Throughout the report, where available, we are as detailed as possible, and report on specific special educational needs (e.g., neurodiversity), or specific ethnic groups (e.g., Black Caribbean young people).

Some personal characteristics are likely to be present in the literature more than others and have been defined below (see Table 13). However, this is not a recommendation to limit EDIE factors to the below categories, should other personal characteristics be present in the literature then these must also be appraised and recorded. For example, other EDIE factors that may be present in the literature include mental health needs, neurodiversity, religion, and sexual orientation.

Table 13. Core definitions of personal characteristics

Concept	Definition
Gender	An individual's deeply felt internal perception of oneself, based on socially constructed roles and behaviours. This may or may not differ from their designated sex at birth (WHO, 2024)
Ethnicity	Social groups that share a distinctive, yet common culture, background, religion and language. Minority ethnic young people refer to young people who belong to UK minority ethnic groups, including Asian, Black, and mixed ethnic, other ethnic groups, and White minority ethnic groups, including Gypsy, Roma, and Irish Traveler groups (Gov.uk, 2024).
SEND	Learning difficulties and/or disabilities which make it challenging for CYP to learn at a similar rate to their peers.
Care Experienced	Children and young people who have been in care (whether foster care, a residential placement, or via family/kinship care).

[21] Throughout the EDIE section, in all documents and on the Toolkit, a reflective approach to language is used, trying to be sensitive and respectful. Where possible language that is used comes directly from people with relevant characteristics (i.e., based on self-description found in included research studies), however, it is important to acknowledge that people have different preferences.

Experience of Deprivation	As specified in the Index of Multiple Deprivation, deprivation refers to a lack of resources across income, employment, health, disability, education, skills training, crime, and housing. This could also be measured through factors such as access to free school meals.
Educational Attendance and Attainment	Extent to which a CYP attends school and their level of achievement (e.g., qualifications attained/learning level).

EDIE evidence in the Toolkit

EDIE details

The Toolkit provides details on who is in the sample. In particular, this section focuses on how representative the evidence-base is of the general population (for primary interventions) or young people in the Criminal Justice System (CJS) (for tertiary interventions). Those involved in secondary interventions would reflect a transitional population, falling between the general and CJS populations. The implications of who the sample is for our understanding of how effective the intervention is for different people is considered. This includes reflecting on any moderator analyses as outlined in Section 5 above.

Extracting EDIE Data

The process for finding and reporting EDIE data in the Toolkit is as follows:

1. Identify relevant studies within the Evidence and Gap Map by filtering by the strand.
2. Check that identified studies fit the PICOS and inclusion/exclusion criteria outlined in the scoping note.
3. Appraise the identified studies using the YEF-EQA tool for mixed methods or qualitative process evaluations. For studies that incorporate process insights as part of a wider evaluation, appraise the process aspect separately.
4. Extract data into EPPI-Reviewer for each study using the EDIE codeset.

5. Write a narrative summary of the EDIE details for the technical report, discussing the quality of the evidence base and noting any areas that are not covered by the evidence base.
6. Write the content for the second level of the Toolkit (the summary page), summarising the available evidence under the heading “Who does it work for?”, prioritising high and medium quality evidence from the UK and/or Ireland where this is available.

Reporting on EDIE

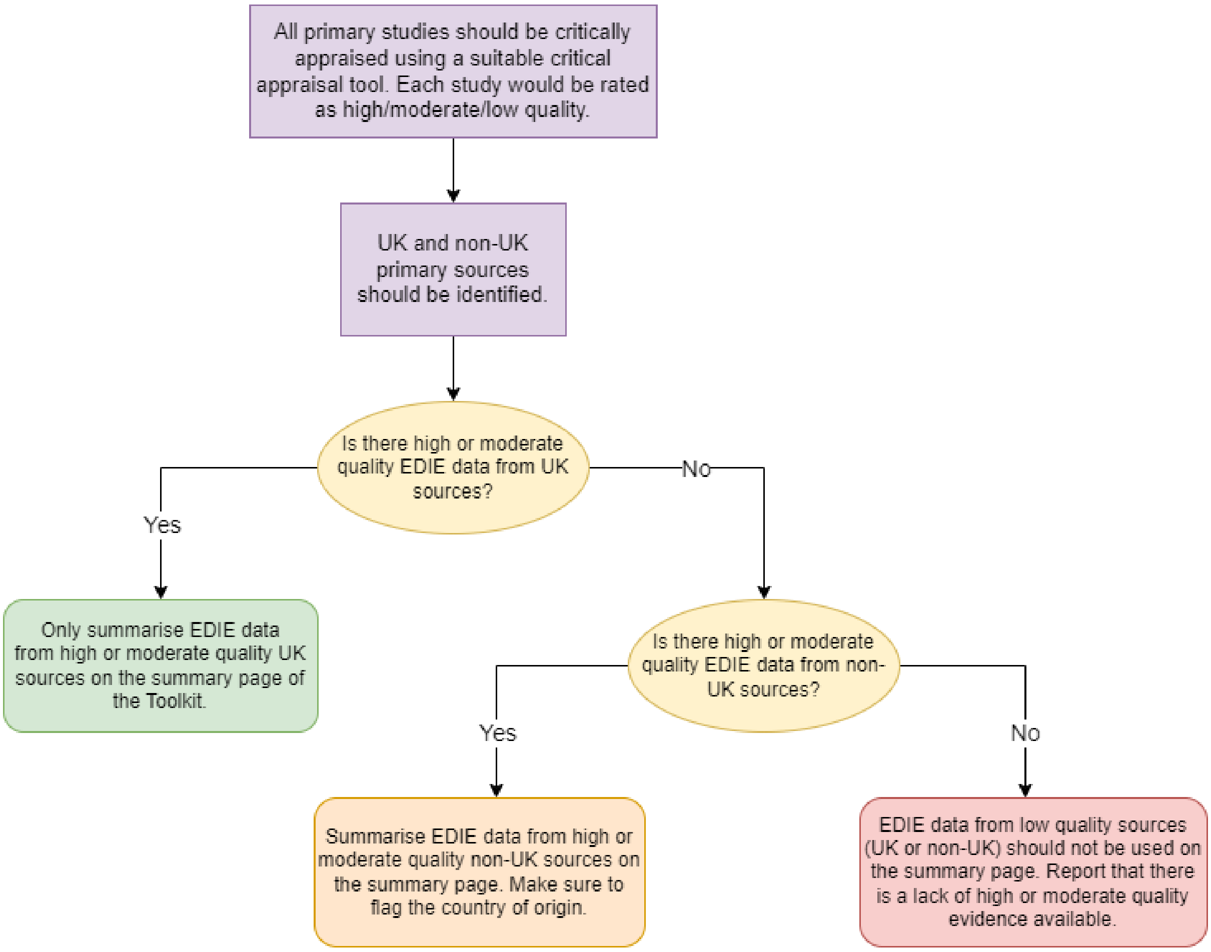
Toolkit front page

EDIE information does not feature on the front page of the Toolkit.

Summary page

To summarise EDIE information on a Toolkit approach, the prioritisation framework shown in Figure 11 should be followed. This highlights that a quality appraisal must first be completed on the primary studies underpinning the approach, with the locations where the studies were undertaken noted. Only high or moderate quality studies should be included in the narrative on the Summary Page of the Toolkit. Where this is available, UK studies should be prioritised and summarised in the narrative. Where this is unavailable, high or moderate quality studies from international sources can be included in the narrative, but the country of origin must be flagged. Importantly, on the Summary Page of the Toolkit, low quality sources should not be reported.

Figure 11. Prioritization framework for writing narratives on summary page


All primary studies should be critically appraised using a suitable critical appraisal tool. Each study would be rated as high/moderate/low quality.
UK and non-UK primary sources should be identified.
Is there high or moderate quality EDIE data from UK sources?
Yes
No
Only summarise EDIE data from high or moderate quality UK sources on the summary page of the Toolkit.
Is there high or moderate quality EDIE data from non-UK sources?
Yes
No
Summarise EDIE data from high or moderate quality non-UK sources on the summary page. Make sure to flag the country of origin.
EDIE data from low quality sources (UK or non-UK) should not be used on the summary page. Report that there is a lack of high or moderate quality evidence available.

A summary paragraph should be written based on the selected primary studies from the prioritisation framework. This should describe the study population, specifying whether the approach has been tested with a general population or with specific groups of young people, and whether studies have excluded or failed to represent certain groups. If effect sizes have been calculated for sub-groups, these should be described here (see moderator analyses above). Statement examples are provided in Table 14 to ensure consistency in Summary Page EDIE narratives across approaches.

Table 14. Statement examples for EDIE Summary Page narratives

EDIE/UK	Statement Example
High/ moderate findings available from UK on EDIE.	Findings from high and moderate quality UK studies suggest that this approach is more effective for male young people than for female young people. There is some evidence that it is more effective for White young people than for Black or Asian young people. There also is some evidence that pairing young people with facilitators of the same ethnic background improved effectiveness for Black, Asian, and mixed ethnic young people. There has been no research exploring the effectiveness of this approach according to SEND.
High/ moderate findings available from UK on EDIE, which are contradictory.	There were contradictory findings from high quality UK studies. One study found this approach to be more effective for male young people, whilst another study found it to be more effective for females.
No high/ moderate findings available from UK on EDIE, but high/ moderate findings from other countries available.	There were no high or moderate quality UK studies which explored the impact of [enter strand name] and EDIE. However, a high-quality study from the US found this approach works well for those without SEND, but not for those with SEND. No research exploring the effectiveness of this approach according to ethnicity or gender has been conducted.
Only low-quality findings from UK or other countries available.	There has been a lack of high or moderate quality studies from both the UK and other countries which have explored this approach in relation to personal characteristics or outcomes, such as gender, ethnicity and SEND.
No insights regarding EDIE available.	To date, there has been no research exploring the effectiveness of this approach according to gender, ethnicity, SEND.

Technical report

The Technical Report should describe the evidence base for EDIE information, discussing the quality of the underlying studies and noting any areas that are not covered. Unlike the Summary Page, it includes all sources (regardless of quality or location). The report should have a detailed narrative of all EDIE information available, including study populations for all included studies. This will continue with a detailed narrative of any groups that have been excluded or under-represented in the research and what the implications of this might be for the effectiveness of the intervention in different contexts. The report should detail any sub-group analysis that has been carried out on outcomes for different groups. The report should also note how groups have been defined or categorized and whether there are any issues, for example whether differences in definitions between studies make it difficult to combine effect sizes or whether the categories used are too broad or too narrow to be useful.

7. Implementation evidence

The Toolkit describes the key features of an intervention and provides information on how to implement the intervention well, including any contextual factors that might affect the intended outcomes. The implementation section draws on evidence from process evaluations of the intervention, which may be published separately or may be included as part of a wider evaluation.

Intervention details

The Toolkit makes use of the Template for Intervention Description and Replication Checklist (TIDieR; Hoffman, 2014[22]) to capture information reported on:

- The *name* of the intervention including the long form of any abbreviations or if unnamed a very brief description of the intervention to identify the type of intervention.
- *Why* the intervention is expected to work, in other words, the theory of change and presumed causal mechanisms. The Toolkit summarises the theory of change where this is available, to explain how and why the activities lead to the desired impact on violent crime. The Toolkit includes any intermediate outcomes and how these relate causally to the activities and to the long-term outcome of reduced violence and crime, as well as any assumptions. If a Theory of Change diagram or logic model is available, then this is included as an appendix to the Toolkit technical report. If not, a narrative description of the aims, rationale or essential components of an intervention will be captured.
- *What* is delivered, in terms of the physical or information materials or resources used and the process or procedures for using them.
- *Who* delivers the intervention: the key professionals or other personnel involved in the intervention and any specific training or accreditation they require to deliver the intervention.

[22] Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman D. G., Barbour V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McColloch, P., Wyatt, J. C., Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, *348*, 1-12. https://doi.org/10.1136/bmj.g1687

- *How* an intervention is delivered including the medium (e.g., face to face, online) and format (e.g., group, individual, mass media) and any other relevant delivery features. For example, was it interactive or not? Are people obliged to participate for example court mandated or not?
- *Where* it is delivered: the location or setting of the intervention.
- *When and how much*, for example the duration of the intervention and its intensity (e.g., how many sessions over how many weeks/months), how long was each session and any information on flexibility in the dosage (e.g. attending 70% of sessions is considered sufficient).

Implementation outcomes

Effective implementation is essential to realising outcomes of effective interventions. The implementation section draws on evidence from process evaluations of the intervention, which may be published separately or may be included as part of a wider evaluation.

Implementation outcomes capture the effect of deliberate and planned strategies and efforts to implement new treatments, services, practice or interventions. Studying implementation outcomes answers questions about how well an intervention was implemented and what impacted on how well or poorly an intervention was able to be implemented. To capture implementation outcomes the toolkit data extraction made use of Procter et al.'s (2011) Implementation Outcomes Framework[23] to capture and categorise the barriers and facilitators to achieving good implementation.

The data extraction for the toolkit is an extension of what is already captured in the EGM. For the EGM the focus was on whether or not implementation outcomes were measured. In other words, does a study report on indicators of how well the programme/intervention was implemented or not? For toolkit data extraction we capture why implementation did or did not go well and factors that influenced implementation. This is typically thought of as barriers and facilitators to implementation. Information on barriers and facilitators will

[23] Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2010). Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

be presented using Procter et al.'s (2011) Implementation Outcomes as headings so that the reader can understand the evidence, and gaps in the evidence, on the following implementation outcomes:

- **Acceptability**

  Stakeholders' perceptions that the intervention or change is agreeable, palatable, or satisfactory. Example indicators: participant's views on the intervention, participant engagement, satisfaction with content or delivery.

- **Adoption**

  The decision or action to employ an intervention or implementation target. Example indicators: Uptake of the intervention by services, schools, or communities.

- **Appropriateness**

  The perceived fit or relevance of the intervention to the given context or problem. Example indicators: Adaptations made to improve the intervention's fit with the context, perceived usefulness.

- **Feasibility**

  The extent to which the intervention can be successfully implemented in a specific setting. Example indicators: Evidence of practicality or utility, ability to deliver the intervention in the target environment.

- **Fidelity**

  The degree to which the intervention was delivered as intended. Example indicators: Training quality, dosage and intensity of the intervention, adherence to the prescribed approach.

- ***Reach/Penetration***

  The extent to which the intervention has been integrated into a service setting or reached eligible recipients. Example indicators: Ratio of recipients served to the target population, evidence of saturation or integration.

- ***Sustainability***
  The ability to maintain or institutionalise the intervention over time. Example indicators: Evidence of routinisation, integration into policies or practices, durability of implementation efforts.

Where implementation barriers/facilitators or influences on an implementation outcome were not measured and/or reported this is stated.

The Toolkit specifically notes any factors that disproportionately affect CYP based on their gender, ethnicity, experience of deprivation, experience of care, SEND, or any other protected characteristic. If the available studies do not consider how the experiences of different groups of young people may vary, the Technical Report states this as a weakness in the current evidence base.

Views of Children and Young People

The Toolkit includes the views of CYP who have participated in the intervention, ideally in the form of direct quotes if these are available. If no studies report on the views of CYP with regards to the intervention, then this is stated.

Extracting implementation data

The process for finding and reporting implementation data in the Toolkit is as follows:

1. Identify relevant studies within the EGM by filtering by strand and then either by study design to identify process evaluations or by whether the study contains process insights.
2. Check that identified studies fit the PICOS and inclusion/exclusion criteria outlined in the scoping note.
3. Appraise the identified studies using the YEF-EQA tool for mixed methods or qualitative process evaluations. For studies that incorporate process insights as part of a wider evaluation, appraise the process aspect separately.
4. Extract data into EPPI-Reviewer for each study using the implementation code set.

5. Record study details in the 'implementation study details' template (see Appendix 5) to be included in the technical report appendix.

Reporting on Intervention Details and Implementation

Toolkit front page

The Toolkit front page describes the intervention in a single sentence and clearly indicates whether the intervention is place-based or person-based. Each approach summary includes a section dedicated to implementation considerations, providing insights into factors that influence the success of interventions.

Summary page

Information for the summary page should be prioritised in line with the prioritisation framework shown in the EDI-E section above (see Figure 11 in Summary page). Moderate or high-quality studies from the UK or Ireland should be prioritised first. If no studies that meet this criterion are available, then moderate or high-quality studies from other countries can be used instead, but this should be made clear in the narrative. If no moderate or high-quality studies are available, then the Summary Page should state this clearly. Unlike the EDIE section, low-quality studies can be used to give factual information about the intervention and how it is implemented, but reviewers should report which aspects of the studies are of low quality and how this might affect the conclusions drawn. The Summary Page should clearly state what proportion of studies used within the section are from the UK or Ireland.

Under the heading "What is it?" the Summary Page provides a detailed description of the approach based on the intervention details extracted from primary studies.

Under the heading, "How can you implement it well?", the Summary Page describes the factors that either hinder (barriers) or improve delivery (facilitators). The views of CYP who have participated in the intervention are reported if available.

Where strands rely heavily on international evidence regarding implementation, it is important that Toolkit Users understand how transferable the information is. As such, the following statement should be added to any strand relying on international evidence alone:

"XXX [enter strand name] has mainly been used outside of the UK, when implementing adapt the principles and materials to your local context."

Technical report

The Technical Report should describe the evidence base for implementation information, discussing the quality of the underlying studies and noting any areas that are not covered. Then the report should have a detailed narrative of the intervention itself, covering a description of the key components of the intervention, including the Theory of Change if available.

Implementation evidence will be summarised narratively, starting with bullet point summaries of themes identified. In-depth implementation evidence will then be provided, organised according to the Procter et al.'s (2011) implementation outcomes.

The views of CYP who have participated in the intervention will be summarised along with information provided on how well CYP were able to give their views, using Lundy's model of children's participation (Lundy, 2007[24]).

The studies used to inform this section should be described in the implementation study details template (see Appendix 5) and included as an appendix to the report, together with the Theory of Change diagram if one is available. The presence or absence of each of Proctor et al.'s (2011) implementation factors should be recorded on Appendix 6 for all studies.

8. Cost data in the Toolkit

The Toolkit provides a summary of cost information for interventions, focusing on the average cost per participant. Costs are banded into three ratings based on monetary ranges:

[24] Lundy, L. (2007). 'Voice' is not enough: conceptualising Article 12 of the United Nations Convention on the Rights of the Child. *British Educational Research Journal, 33*(6), 927–942. Portico. https://doi.org/10.1080/01411920701657033

- Low: Denoted by one ‘£’ sign.
- Medium: Denoted by two ‘££’ signs.
- High: Denoted by three ‘£££’ signs.

These bands provide a headline summary on the first level of the Toolkit.

The Summary Page of the Toolkit offers a more detailed narrative about cost. This includes a breakdown into three categories where information is available (Table 15).

Table 15. Description of types of cost incurred by Toolkit intervention with examples

Type of cost	Example
Frontline delivery costs	These are costs that are completely attributable to the delivery of the specific intervention. The key costs for this cost category are staff costs associated with the preparation and delivery of the intervention / approach. Also included in this cost category are the equipment, materials and supplies.
Overhead costs	These are typically fixed in nature and are usually thought of as indirect costs as they are usually spread across all the activities being delivered by the respective organisation. A proportion of these overhead costs need to be attributed to delivery of the intervention. This cost category includes rent, utilities and administrative costs.
Other costs	This includes costs not directly tied to delivering an intervention but essential for maintaining its quality, such as staff training and development and ongoing supervision. Often overlooked, these "hidden costs" should still be accounted for in overall project budgets.

Extracting cost data

Cost data is primarily extracted from the EGM. A review of 101 papers coded as having cost data revealed that only 16 provided sufficiently disaggregated information across all three cost categories. Updates to the Toolkit are unlikely to substantially change this picture of data availability. To address this gap, NCB developed a pro-forma template for collecting and collating cost information. This template (available via the project’s OSF page: https://osf.io/a9bhd/)

1) Allows organizations delivering a particular intervention to enter data for the three broad cost categories (outlined in Table 15 above) with individual cost items included under each broad cost category.
2) Automatically calculates total cost, average costs (per participant starting / completing / successfully completing the intervention), and presents aggregated cost data for the three broad cost categories.

The average cost data line is used to feed into the creation of the overall cost rating (see below for more details of the cost rating bands). In situations where cost data is not available, the template will be used by a panel of experts[25] to ensure that there is a standardized and consistent approach to costing up interventions.

Challenges in Estimating Costs for Place-Based Approaches

For place-based interventions such as Problem-Oriented Policing, CCTV, and street lighting, estimating a standardised "average cost per participant" is particularly challenging. These approaches often do not have clearly defined participants, and the scale and design of delivery can vary widely depending on location, local resources, and implementation partners. As a result, it is difficult to assign a precise cost rating based on participant numbers alone.

Where appropriate, the Toolkit supplements the cost band with a narrative explanation that outlines the types of resources typically involved (e.g., police time, infrastructure, coordination with local authorities), provides indicative cost examples where available, and explains the reasoning behind the cost band assigned. In some cases, a “?” symbol is used to indicate that cost estimates are uncertain or highly variable.

Reporting on Cost

Toolkit front page

The Toolkit summarises intervention costs with an average cost per participant, providing a snapshot of costs relative to other approaches. These costs include setup, delivery, and

[25] YEF are developing a protocol for how these experts would be engaged to cost up interventions, where this is needed.

ongoing costs (e.g., training), but exclude counterfactual costs (i.e., cost of not implementing an intervention, such as cost of incarceration) unless explicitly noted. Cost data from the UK are prioritised, but where only international data is available, costs are adjusted for inflation and converted to GBP using current exchange rates. Costs associated with non-completers are excluded unless otherwise noted in the technical report. The cost bands are outlined in **Table** 16.[26]

Table 16. Cost bands for YEF Toolkit interventions

Band	**Average cost per participant (£ Stg) as estimated from available cost data**
Low (£)	£0 - £499
Medium (££)	£500 - £1,499
High (£££)	£1,500 +

Cost information is presented as a summary with banded ratings (Low/Medium/High) based on the average cost per participant.

Summary page

The Summary Page provides a narrative summary of costs, including:

- Average cost per participant.
- Frontline delivery, overhead, and other costs (where available).
- Any limitations in data availability.

UK data are prioritised in this summary, and international data is flagged with country-specific considerations.

[26] The bands have been recalibrated slightly from previous versions of the Toolkit so that each of them do not overlap. E.g., in previous versions of the Toolkit, the medium band was £500 - £1,500.

Technical report

The Technical Report includes a detailed cost analysis, if the underlying data allows. This will:

- Explain cost-effectiveness calculations and the creation of the cost ratings.
- Discuss the quality of cost data, including its sources and limitations.
- Note any missing or incomplete cost categories.
- Incorporate counterfactual costs when available.

This report will also document how costs were calculated and includes a comprehensive narrative of the evidence base for cost data. The Technical Report will reference the pro-forma template in an appendix, alongside any relevant methodology for expert panel engagement.

Future updates

Cost bands and data will be periodically reviewed (approximately every five years but depending on how specific factors evolve such as the rate of inflation) to align with real-world changes. To address the current limitations in cost data availability across the evidence base, the pro-forma template developed by NCB for person-based interventions will play a crucial role in standardising cost data collection and improving the overall quality and comprehensiveness of cost information included in the Toolkit. However, the team also acknowledges that cost assessment for place-based interventions remains an area for future development. Pragmatic assessments and expert judgement will continue to inform cost ratings for place-based approaches, supported where possible by use of the NCB cost template.

9. References

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088-1101.

Cham, H., Lee, H. & Migunov, I. (2024). Quasi-experimental designs for causal inference: an overview. *Asia Pacific Educ. Rev.* 25, 611–627. https://doi.org/10.1007/s12564-024-09981-2

Cochrane Handbook, 9.5.2 Identifying and measuring heterogeneity. https://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm [Accessed 02/12/2024]

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. *Archives of pediatrics & adolescent medicine*, *163*(5), 438-445.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56(2),* 455-463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315(7109),* 629-634. https://doi.org/10.1136/bmj.315.7109.629

EPPI Reviewer. https://eppi.ioe.ac.uk/cms/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4& [Accessed 03/12/2024]

Farrington, D. P. (2012). Childhood risk factors for young adult offending: Onset and persistence. In F. Lösel, A. Bottoms, & D. P. Farrington (Eds*.), Young adult offenders: Lost in transition?* (pp. 48–64). Routledge.

Farrington, D. P., & Malvaso, C. G. (2023). Interactions between child-rearing and other risk factors in predicting delinquency, and implications for prevention. *International journal of offender therapy and comparative criminology*, 1-17. https://doi.org/10.1177/0306624X231188231

Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. *arXiv preprint* arXiv:1503.02220. https://doi.org/10.48550/arXiv.1503.02220

George, A., Stead, T. S., & Ganti, L. (2020). What's the risk: differentiating risk ratios, odds ratios, and hazard ratios?. *Cureus, 12(8).* https://doi.org/10.7759/cureus.10047

Gov.uk. (2024). *Writing about ethnicity.* https://www.ethnicity-facts-figures.service.gov.uk/style-guide/writing-about-ethnicity/

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions (Vol. 4).* John Wiley & Sons

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327(7414),* 557.

https://doi.org/10.1136/bmj.327.7414.557

Hill, K.G., Woodward, D., Woelfel, T. *et al.* Planning for Long-Term Follow-Up: Strategies Learned from Longitudinal Studies. *Prev Sci, 17,* 806–818 (2016). https://doi.org/10.1007/s11121-015-0610-7

Irwig L, Irwig J, Trevena L, et al. (2008). *Smart Health Choices: Making Sense of Health Advice.* London: Hammersmith Press. Chapter 18, Relative risk, relative and absolute risk reduction, number needed to treat and confidence intervals. Available from: https://www.ncbi.nlm.nih.gov/books/NBK63647/

Krug, E. G., Mercy, J. A., Dahlberg, L. L., & Zwi, A. B. (2002). *The world report on violence and health.* World report on violence and health

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Sage Publications, Inc.

McAra, L., & McVie, S. (2016). Understanding youth violence: The mediating effects of gender, poverty and vulnerability. *Journal of criminal justice*, *45*, 71-77. https://doi.org/10.1016/j.jcrimjus.2016.02.011

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Statistics in medicine, 26(25), 4544-4562. https://doi.org/10.1002/sim.2889

Pustejovsky, J.E., Tipton, E. Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prev Sci, 23,* 425–438 (2022). https://doi.org/10.1007/s11121-021-01246-3

Pustejovsky, James E. (2024). Meta-analysis with cluster-robust variance estimation. https://cran.r-project.org/web/packages/clubSandwich/vignettes/meta-analysis-with-CRVE.htmlQuintana, D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in psychology, 6.* https://doi.org/10.3389/fpsyg.2015.01549

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86(3),* 638. https://doi.org/10.1037/0033-2909.86.3.638

Rosnow, Ralph L., Rosenthal, Robert, and Rubin, Donald B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*(6): 446-453. https://doi.org/10.1111/1467-9280.00287 Available from: https://dash.harvard.edu/handle/1/3199067 [Accessed 27/11/2024].

Rosnow, Ralph. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*(4), 331–340. https://doi.org/10.1037/1082-989X.1.4.331

Shea, B. J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E, H. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ. 21,* 358.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53(11),* 1119-1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22(13),* 2113-2126. https://doi.org/10.1002/sim.1461

Viechtbauer W (2010). "Conducting meta-analyses in R with the metafor package." *Journal of Statistical Software, 36(3),* 1–48. https://doi.org/10.18637/jss.v036.i03

Von Hippel, P. T. (2015). The hetereogeneity statistic *I2* can be biased in small meta-analyses. *BMC Medical Research Methodology,* 15: 35. https://doi.org/10.1186/s12874-015-0024-z

Wilson, D.B., Brennan, I. and Olaghere, A. (2018). Police-initiated diversion for youth to prevent future delinquent behavior: A systematic review. *Campbell Systematic Reviews, 14,* 1-88. https://doi.org/10.4073/csr.2018.5

Wilson, D. B. (2022). The relative incident rate ratio effect size for count-based impact evaluations: When an odds ratio is not an odds ratio. *Journal of Quantitative Criminology, 38(3),* 323–341. https://doi.org/10.1007/s10940-021-09494-w

Wilson, D. B. (2023). *Practical meta-analysis effect size calculator* (Version 2023.11.27). https://www.campbellcollaboration.org/calculator/

World Health Organization. (2024). *Gender and health.* https://www.who.int/health-topics/gender#tab=tab_1

World Health Organization (2002). *Word report on violence and health.* https://www.who.int/publications/i/item/9241545615

YEF (2021). *Technical Guide Version 4-1. December* 2021. Available from: https://youthendowmentfund.org.uk/wp-content/uploads/2021/06/YEF-Toolkit-technical-guide-December-21.pdf [Accessed 29/10/2024]

YEF (2023). *Children, violence and vulnerability.* Available from: YEF_Children_violence_and_vulnerability_2023_FINAL.pdf [Accessed 10/12/2024]

Youth Justice Board (2024). *Youth Justice Statistics 2022 to 203 England and Wales.* Available from: https://assets.publishing.service.gov.uk/media/65b391a60c75e30012d800fa/Youth_Justice_Statistics_2022-23.pdf [Accessed 10/12/2024]

10.Appendices

Appendix 1. Detailed updates to technical guide

Selecting evidence for the Toolkit

Under the previous approach, evidence for the Toolkit was identified from the EGM but also via supplementary searches in Google Scholar and suggestions from Professor David Farrington and Dr Hannah Gaffney (YEF, 2021). In the current approach, the requirement for additional hand searching and consultation with experts has been removed to move towards a more automated model of the Toolkit. Recent updates to the EGM ensure that it is a more comprehensive repository of studies that regularly refreshes with the most up-to-date evidence. There is also the option for researchers to upload their own research for the review team to check and incorporate into the EGM. In this way, the Toolkit continues to include comprehensive, up-to-date evidence with input from experts, but the time and resource required for manual updates is reduced.

Extracting data for the Toolkit

In the previous version of the Toolkit, data was extracted from studies but was not stored centrally. Instead, the PDF Technical Report gave a narrative summary of all the data extracted. Under the new approach, all data extracted is entered into the EGM within EPPI-Reviewer so that all data pertaining to one study is in a single location, enhancing replicability and transparency.

Quality appraisal of studies for the Toolkit

Previously the Toolkit did not report on the quality appraisal of primary studies. Under the new approach, primary studies are appraised using the YEF-EQA, with the Toolkit reporting on the quality rating and any implications so that Toolkit users can better understand the strengths and weaknesses of the whole evidence base.

Estimating the effectiveness of Toolkit interventions

Calculating the headline effect size

Previously, the headline impact estimate for each Toolkit strand was obtained from a high-quality published meta-analysis[27]. Under the new approach, each strand conducts a new meta-analysis based on the available primary studies within the EGM to ensure the effectiveness estimate is based on the most up-to-date evidence available.

The previous version of the Toolkit used Cohen's d as the effect size for place-based interventions as well as person-based interventions. Methodological developments in the field have highlighted that Cohen's d is not a suitable effect size for place-based interventions. The new Toolkit approach follows the recommended best practice for place-based interventions which is to use RIRR as the effect size. The Front Page of the Toolkit will now highlight the difference between place-based and person-based interventions so that end users are aware that the YEF impact rating for place-based interventions is the 'Estimated impact on violence crime rates in targeted areas' rather than the 'Estimated impact on violent crime' reported for person-based interventions.

Communicating effectiveness

Previously the Toolkit presented the Relative Risk Reduction as a percentage. The current version retains the Relative Risk Reduction but also adds a visual that represents the Absolute Risk Reduction. The visual clearly demonstrates the difference between the intervention and comparison groups and helps end users to understand the magnitude of the difference, particularly where the underlying prevalence of the outcome is lower.

Evidence security rating

The previous evidence security rating was based on four criteria:

1. The number of primary studies used in the meta-analysis used to calculate the headline impact rating

[27] In some cases where a published meta-analysis was not available, the Toolkit team commissioned a meta-analysis to inform the Toolkit strand

2. The AMSTAR 2 rating of the systematic review which provides the headline impact rating
3. The heterogeneity of studies informing the headline impact rating
4. Whether the headline impact estimate is a direct measure of crime or violence or not

Under the new approach, the rating now considers the quality of the underlying primary studies as well as the quantity. As Toolkit strands are now based on a bespoke meta-analysis rather than a published systematic review, the AMSTAR 2 rating is no longer used. However, the new meta-analysis process meets the criteria laid out in AMSTAR 2 ensuring that the synthesis of evidence for the Toolkit is consistently high-quality. Heterogeneity is still considered but prediction intervals and tau^2 are now used in addition to inform the decision of whether to downgrade the rating or not.

EDI-E

Previously, the Toolkit did not regularly collect or report on data regarding EDIE. The updated version of the Summary Page includes a new section termed 'Who does it work for?'. Here a narrative summary is provided on whether personal characteristics have been considered in high/moderate quality UK (as a preference) or international research. Within the Technical Report, all the evidence-base (regardless of quality or location) is summarized, any sub-group analyses are reported, and definitional issues considered.

Implementation evidence

The updated version of the Toolkit broadly includes the same detail about the implementation of each approach. The main change is that more detail is now provided on how to find evidence for the implementation section, and more detail is reported on the studies included, for example listing the studies that provide implementation evidence together with their methods and quality rating. This provides more transparency about the evidence underlying all aspects of the Toolkit and enables end users to more easily find the sources of evidence and to understand any biases or gaps in the evidence base.

Cost data

The updated version of the toolkit maintains the current cost rating bands (low / medium / high) with minor adjustments to the monetary range of the bands to ensure they are mutually exclusive. For those interventions where we are unable to extract sufficient cost data from the EGM, we have developed a cost template that can be completed by those who commission or deliver the intervention. The cost template is structured to capture cost data under three broad cost categories (frontline delivery costs, overhead costs and other costs) and enables automated collation of cost data metrics such as average cost per participant starting and completing the intervention as well as an overall intervention cost. The template can also be used by YEF at the point of commissioning to help ensure consistent cost data is collected. We have also suggested as part of this update that the cost bands are adjusted periodically (c. every five years) to ensure they reflect the actual cost of intervention delivery.

Appendix 2. Location details template

Study reference (Author, year, EPPI study ID):

____

	Number of UK Studies	Number (and Location) of International Studies
Overall, for Strand		
Contributing to Evidence Quality Rating		
Contributing to Estimated Impact on Violence		
Contributing to EDIE Information		
Contributing to Implementation		
Contributing to Cost Data		

Appendix 3. YEF Evidence Quality Assessment (YEF-EQA) tool

			Implementation / process evaluation (qualitative or mixed methods study)	Feasibility study (pre/post)	Impact evaluation (RCT or QED)	Rating
1	**Study design**	a	Is the intervention or approach clearly named and described, including all relevant components?	Is the intervention or approach clearly named and described, including all relevant components?	Is the intervention or approach clearly named and described, including all relevant components?	**High:** full and clear description, so that the main components and how they are delivered are clear **Medium:** Partial description **Low:** Little or no description
		b	Are the evaluation questions clearly stated?	Are the evaluation questions clearly stated?	Are the evaluation questions clearly stated?	**High:** Specific, clearly stated evaluation questions are presented. **Medium:** Hypotheses or research aims are clear, but no explicit evaluation questions are stated. **Low:** No clear evaluation questions, only vague references to aims or

						hypotheses
		c	Is the qualitative methodology described?			**High:** full and clear description **Medium:** partial description **Low:** little or no description ***(if low – skip next item)***
		d	***(If 1c is medium or high)*** Does the qualitative methodology align with the research objectives, specifically addressing the evaluation questions?			**High:** The chosen qualitative methodology is well-suited to the study's objectives, clearly addressing the evaluation questions and providing a comprehensive understanding of the context, processes, and outcomes. The methodology is explicitly linked to the research goals and the specific aspects of the process being evaluated. **Medium:** The qualitative methodology is generally appropriate, but there are some limitations in how well it aligns with the study objectives.

						The methodology addresses some key questions but may overlook others or have limitations in depth or scope. **Low:** The qualitative methodology does not adequately align with the study's objectives. It does not effectively address the evaluation questions or is not appropriate for understanding the processes or context being examined.
2	**Recruitment & sampling**	**a**	Is the recruitment or sampling strategy described?	Is the recruitment or sampling strategy described?	Is the recruitment or sampling strategy described?	**High:** recruitment and sampling strategy fully described, including considerations for place-based interventions where entire locations are the study subjects. **Medium:** recruitment and/or sampling partially described **Low:** recruitment or sampling strategy not described /

						insufficient detail ***(if low – skip next item)***
		b	***(if 2a is medium or high)*** Is the recruitment or sampling strategy appropriate to address the evaluation questions?	Is the recruitment or sampling strategy appropriate to address the evaluation questions?	Is the recruitment or sampling strategy appropriate to address the evaluation questions?	**High:** appropriate recruitment or sampling, accounting for challenges in defining study populations at a geographic level. **Medium:** somewhat limited sampling & recruitment strategy **Low:** inappropriate recruitment or sampling or unclear from description
		c	Has there been an assessment of recruitment bias and reporting on diversity of sample?	Has there been an assessment of recruitment bias and reporting on diversity of sample?	Has there been an assessment of recruitment bias and reporting on diversity of sample?	**High:** sample characteristics are reported, and the sample represents the diversity of the target population well **Medium:** clear assessment and reporting of recruitment bias or diversity of sample. For place-

						based studies, if the sample is contextually valid but not diverse, it can still be rated as medium while acknowledging limitations **Low:** no clear assessment of recruitment bias or diversity of sample
		d	Where relevant, are administrative data sources used appropriately?	Where relevant, are administrative data sources used appropriately?	Are administrative data sources used appropriately?	**High:** Administrative data is used with clear justification, comprehensive description of sources, and acknowledgment of biases or limitations. **Medium:** Some discussion of administrative data but lacks full justification or detail on potential biases. **Low:** Administrative data is used inappropriately, or no discussion on quality and limitations.

3	**Positionality, assumptions and biases**	a	Are the researcher's own position, assumptions and possible biases outlined?			**High:** The researcher provides a detailed and thoughtful reflection on their positionality, assumptions, and biases, explaining how these may influence the study design, interpretation of findings, and conclusions. **Medium:** The researcher offers some discussion of their positionality, assumptions, and biases, but the reflection is brief or somewhat general. Acknowledgment of potential researcher biases is considered sufficient if study limitations are discussed. **Low:** The researcher does not explicitly address their positionality, assumptions, or biases, or provides only a vague or superficial mention with no critical reflection on their potential impact on the

						research.
4	**Outcomes**	**a**		Are the outcomes clearly defined? Where appropriate, use of an existing, validated measurement tool?	Are the outcomes clearly defined? Where appropriate, use of an existing, validated measurement tool?	**High:** Both the outcomes are clearly defined and validated measurement tools are used where applicable, with clear citations for those tools. **Medium:** Outcomes are clearly defined, but the measurement tools are not validated or not adequately referenced. Studies using administrative data as outcome measures should be rated Medium, provided they offer transparency and justification **Low:** Outcomes are not clearly defined, even if validated tools are used. If no outcome definition exists (or is very

						vague), the study deserves a Low rating.
		b			Has confounding been adequately controlled?	**High**: Adequate controlling for confounding **Medium:** controlling for some confounding variables **Low:** Inadequate controlling for confounding
		c			Has a trial been stopped early? Selective outcome reporting: Have endline and longest follow up been reported?	**High:** Trial run to completion and all outcomes reported **Medium:** Trial completed, but some anticipated outcomes mentioned in intro section or at baseline (e.g., author mentions likely success at long-term follow-up, or specific subgroups included) were then not fully reported or not all reported with equal emphasis

						Low: Trial stopped early and/or it is clear that there is selective outcomes reporting (e.g., only statistical significant results)
5	**Data analysis**	**a**	Is the data analysis approach adequately described?	Is the data analysis approach adequately described?	Is the data analysis approach adequately described?	**High:** full and clear description **Medium:** some description, but lacks clarity **Low:** little or no description **(if low – skip next item)**
		b	**(If 5a is medium or high)** Is the data analysis sufficiently rigorous?	(If 5a is medium or high) Is the data analysis sufficiently rigorous?	(If 5a is medium or high) Is the data analysis sufficiently rigorous?	**High:** data analysis sufficiently rigorous **Medium**: data analysis approach is pragmatic, but well-justified **Low:** data analysis not sufficiently rigorous or insufficiently described

6	**Implications & recommendations**	**a**	Are the implications or recommendations clearly based in the evidence from the study?	Are the implications or recommendations clearly based in the evidence from the study?	Are the implications or recommendations clearly based in the evidence from the study?	**High:** implications or recommendations clearly based in study evidence **Medium:** some implications/recommendations are firmly rooted in the evidence, and some less so. **Low:** implications or recommendations not clearly based in study evidence
7	**Assignment to treatment and comparison groups**	**a**			Is assignment to treatment and comparison groups done at the appropriate level (e.g. individual, community)?	**High:** assignment to treatment and comparison group at the appropriate level **Medium:** Assignment not randomised but applied at the appropriate level with strong justification for group comparisons, such as a well-matched quasi-experimental design. **Low:** assignment to treatment and comparison group at inappropriate level

		b			Are the methods used to assign participants to treatment and comparison groups sufficiently rigorous?	**High:** sufficiently rigorous methods **Medium:** Assignment methods are not fully rigorous (e.g., no randomisation or advanced statistical controls), but some effort is made to ensure comparability between groups-such as matching on key variables or clear justification for the comparison group. **Low**: insufficiently rigorous methods

Appendix 4. Characteristics of included studies for effectiveness

Authors (Year)	Country	Study Design	Intervention	Population/ Place	Comparison	Outcomes Measured	Quality Level	Findings

Appendix 5. Implementation study details template

Authors (Year)	Country	Study Design	Intervention	Quality Level	Implementation Outcomes	Experiences of CYP

Appendix 6. Availability of evidence according to each of Proctor et al.'s (2011) implementation outcomes

Authors (Year)	Acceptability	Adoption	Appropriate-ness	Feasibility	Fidelity	Reach/ penetration	Sustainability	Cost
Example A	Yes	No	Yes	No	No	No	No	No
Example B	Yes	Yes	Yes	No	No	No	No	No

YOUTH ENDOWMENT FUND

Impetus

Home Office

The Youth Endowment Fund Charitable Trust Registered Charity Number: 1185413