YOUTH
ENDOWMENT
FUND

Security ratings for Secondary Data Analysis (SDA) studies

Guidance for research teams and peer reviewers

August 2025

Contents

Acknowledgements

We'd like to thank everyone who has contributed to the development of this document, in particular:

- the Education Endowment Foundation (EEF) for their work on security ratings that provided the framework for our approach
- Jennifer Stevenson, who has led the work to adapt the EEF approach to work in the context of YEF Secondary Data Analysis impact studies; and,
- a range of contributors who have provided helpful comments and input, including Iain Brennan, Celeste Cheung, Jonathan Kay, Camilla Nevill, Alex Sutherland and Umar Toseeb.

All views expressed in this report are those of the Youth Endowment Fund.

Purpose

This document describes the process to arrive at a security rating for Youth Endowment Fund (YEF) Secondary Data Analysis (SDA) impact studies. SDA studies can help to address research questions where Randomised Controlled Trials (RCTs), and other research designs are infeasible, unethical, or inefficient. They typically use retrospective, quasi-experimental approaches that allow for causal inference on the effects of drivers, policies, practices and interventions on crime and violence outcomes[1]. The document is written **for SDA research teams and the peer reviewers** who conduct the assessment for evidence security of SDA studies.

We also fund secondary data analysis that involve purely descriptive analysis, for example on the scale or nature of an issue, or quantitative exploration of the drivers of youth violence. These studies or study components will not have a security rating attached to them.

The YEF's Magnifying Glass (MG) evidence security rating assessment system for YEF impact evaluations is **based on the padlock system developed by the Education Endowment Foundation (EEF) but is adapted to the youth justice sector** and associated outcomes. This security rating system for SDA studies is very closely aligned with the MG security rating system for impact evaluations[2], but with some key changes to account for the distinct characteristics of SDA studies. These differences are presented below. The system makes the interpretation of SDA study evidence strength consistent with the rating of evidence strength of evaluations, that is, RCTs and prospective quasi-experimental evaluations.

Like the EEF's system, the MG rating primarily represents **the extent to which the result of a study can be attributed to an intervention, policy or practice rather than other factors.** It does not include an assessment of the size or direction of effect. The rating does not represent the overall quality of a study. For example, it does not consider the appropriateness and relevance of the research questions, or the extent to which the study builds on existing research or contributes to evidence gaps.

[1] For ease, "Intervention" is used in the rest of the document as an overarching term for any programme, policy, practice, or exposure.

[2] YEF Magnifying Glass Guidance for impact evaluations: https://youthendowmentfund.org.uk/wp-content/uploads/2025/04/YEF-Magnifying-Glass-Guidance.pdf

While information reduction is always controversial in scientific contexts, to achieve our mission of preventing young people becoming involved in crime **it is crucial that we can communicate to practitioners, funders and policy-makers to what extent they can trust our published findings.**

Key differences between SDA studies and evaluations

SDA impact studies have the following characteristics that make them distinct from YEF evaluations (specifically RCTs and prospective quasi-experimental evaluations). These differences necessitated a separate version of the security rating system:

- There is rarely primary data collection, and they largely rely on existing data from administrative datasets, annual national surveys and longitudinal studies.
- They are retrospective, and so the comparison and / or intervention groups may be identified after the fact.[3] Participants in the comparison group are typically chosen based on secondary data (e.g., administrative or survey data). The intervention and comparison participants are not recruited to be part of the study.
- There is no programme team involved in the study as delivery or exposure has already been completed. SDA impact studies do not typically evaluate the impact of a manualised program. They frequently assess a broader national or local policy change where there may be significant variation in how that policy was enacted. There may be little or no information on how it was delivered in practice.
- Sample size calculations will be more complex for some study approaches than they are for an RCT, and may involve the use of simulation approaches.
- They may explore a greater range of research questions than YEF impact evaluations, for example exploration of multiple interventions. They may report a greater number of estimates and more sensitivity analyses than other evaluations.

[3] We occasionally fund the re-analysis of external RCT data through their Secondary Data Analysis funding stream, for example to explore additional sub-groups or longitudinal outcomes. These studies should be appraised using the YEF evaluation evidence assessment system instead of the SDA assessment system.

Overview

The YEF assigns the final security rating, considering assessments by two peer reviewers and the author's opinion.

The process for determining the appropriate security rating is the following:

1. **Two peer reviewers** will use this guidance to provide a security rating,
2. The **YEF arbitrates** between peer reviewer ratings if they differ and presents this to the author,
3. The **author** has an opportunity to respond,
4. The **YEF assigns** the final security rating.

The security rating is determined by three criteria:

- **Design:** The type of design used to create a comparison group with which to determine an unbiased measure of the impact on the primary outcome(s). Higher MGs are given for designs better suited to deal with confounding.
- **The minimum detectable effect size** (MDES): The MDES that the study was powered to achieve, which is heavily influenced by sample size. We expect research teams to undertake MDES calculations at the beginning of their study and again at the interim and final reporting stages.
- **Threats to internal validity:** A series of markers that explain whether the results could be explained by anything other than the intervention.

These are not the only things that are important in determining the security of the results. They are, however, the key factors that differentiate the security of findings for impact-focused SDA studies. The security rating system is only applied to the primary outcome(s). Subgroup analyses are not included in the security ratings unless otherwise stated.

These three criteria are combined to generate an overall padlock rating in four steps:

- **Step 1:** The first two criteria – Design and MDES – are awarded a rating on a scale from zero to four. It is not possible for SDA studies to achieve more than four magnifying glasses because we preserve the highest MG rating (five) for RCTs only.
- **Step 2:** An interim magnifying glasses rating is determined by the lowest of these two ratings.

- **Step 3:** The interim magnifying glasses rating can be adjusted downwards by assessing threats to internal validity.
- **Step 4:** The final magnifying glass rating is determined.

In the following, we first describe all criteria and how they influence the security rating. We expect peer reviewers to read this at least once. While applying the guidelines, you'll be asked to complete an assessment form. Appendix 2 contains a worked example. Separate ratings and assessments should be undertaken for each research question (see box below).

Once the security rating has been agreed, the assessment will added to the final report, with a summary on the reasons for the decision provided in the executive summary of the report.

HOW MANY QUESTIONS SHOULD BE TESTED?

Most YEF evaluations have one intervention of interest, one primary outcome, and therefore one primary research question. In contrast, single SDA impact studies may aim to explore more than one research questions, for example exploration of multiple interventions.

There may also be multiple outcomes of interest, for example impacts on educational attainment, exclusion and absences, besides impacts on crime and violence; or when all-crime is the primary outcome and impact on violent crime (a subset) is further examined. Study teams may also be interested in determining for which subgroups an intervention has an effect, for example, by ethnicity[4].

In cases where research teams explore the impact of different interventions, or the impact of an intervention on multiple distinct outcomes, each will be considered a separate research question and will be assigned a security rating with its own assessment form. This should be agreed with the YEF during the set-up process and described in the study plan.

We expect research teams to:

- Pre-specify and justify the main research question(s), discussing the relevant evidence gaps.

[4] At the YEF we have a particular focus on race equity and encourage research that sheds light on the experiences of children and young people for Black, Asian and other minority backgrounds, including via subgroup analysis.

- Minimise the total number of research questions to avoid fishing / data mining. **Teams are strongly encouraged to keep to one or two main research questions and not exceed three in total.** This is largely because multiple inferences are more prone to false-positive errors.

For each research question, teams should:

- Pre-specify and justify their outcome(s) of interest, including why these are expected to change as a result of the intervention, and the measure(s) used. Where there is more than one outcome of interest, teams may categorise these into primary and secondary outcomes[5] and apply a hierarchical structure, where appropriate. Teams should be clear on which is their preferred outcome on which they're basing power calculations. **The security rating will be applied to the preferred / primary outcome.**
- Pre-specify the preferred approach (e.g. a difference-in-differences) and model specification, or pre-specify decision rules or criteria on how they will decide which approach and model will be their preferred. **The security rating will be applied to the preferred approach.**
- Pre-specify and justify any additional analysis (e.g. subgroup analysis, discussing why the intervention effect is expected to be heterogeneous). Where appropriate, teams may categorise these as secondary or exploratory research questions. **These will not have a security rating applied to them.**
- Apply appropriate multiple hypothesis testing, such as the Benjamini-Hochberg's step-down procedure[6] or Romano-Wolf correction[7].

[5] Pocock, S. J., Rosello, X., Owen, R., Collier, T. J., Stone, G. W., & Rockhold, F. W. (2021). Primary and secondary outcome reporting in randomized trials: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*, *78*(8), 827-839. https://doi.org/10.1016/j.jacc.2021.06.024

[6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B (Methodological)*, *57*(1), 289-300. https://www.jstor.org/stable/2346101

[7] List, J. A., Shaikh, A. M., & Yang, X. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*, 773-793. https://doi.org/10.1007/s10683-018-09597-5

Individual assessment criteria

Criterion 1: Design

This criterion relates to the validity of the comparison group used as an estimate of the counterfactual. Table 1 summarises the scale for rating quality of design. YEF SDA impact studies are expected to be designed to attain at least 3 magnifying glasses (MG) except in rare circumstances or in the case of studies funded in early funding rounds. It is not possible for SDA studies to achieve more than four MGs because we preserve the highest MG rating for RCTs only. We occasionally fund re-analysis of RCT data through our SDA funding stream, for example to explore additional sub-groups or longitudinal data. These studies should be appraised using the YEF Magnifying Glass Guidance for impact evaluations instead of this assessment system.

Table 1 does not include all possible approaches or techniques that we would fund. If a research team's preferred approach is not listed, we will expect them to justify whether their study should be considered 3 or 4 MGs in terms of the ability of the approach to control for unobservable or observable confounders, for the YEF's and peer reviewers' consideration when assigning the MG rating.

The security of the design should be ascertained from (1) the description of the design in the report and protocol, (2) evidence that valid methods were used to identify the comparison group (for example, appropriate methods to reduce imbalance, appropriate and successful matching, support of identification assumptions).

TABLE 1. SECURITY OF THE DESIGN

Rating	Design
5🔍	Randomised design.
4🔍	Design for comparison that considers some type of selection on unobservable characteristics (e.g. Regression Discontinuity Designs, Difference-in-Differences, Matched Difference-in-Differences).
3🔍	Design for comparison selection on all relevant observable confounders (e.g. Matching/Weighting or Regression Analysis with variables descriptive of the selection mechanism).
2🔍	Design for comparison that considers selection only on some relevant confounders

1🔍	Design for comparison that does not consider selection on any relevant confounders.
0🔍	No comparator.

Regression Discontinuity Designs (RDDs), and Matched Difference-in-Differences (MDD) can achieve 4 MG because they attempt to control for some unobservable characteristics. In the case of RDDs it can be considered “as randomised” around the assignment cut-off, while MDD attempts to control for time-invariant heterogeneity. This is also the case for DD, but the assumption of parallel trends necessary for the validity of the estimate is made more tenable using matching. Methods that only attempt to control for observable characteristics (for example, matching/weighting), can only achieve 3 MGs or less.

Criterion 2: Minimum Detectable Effect Size (MDES)

This is the ability of the study to detect a given scale of impact. MDES is highly dependent upon the sample size but is also influenced by other factors, including the outcome variance and the intra-cluster correlation (ICC) (in clustered or multi-level designs). Although in the case of YEF-funded retrospective QEDs, data has already been collected and the study sample size is largely out of the research team’s control, sample size analyses can still help to determine the feasibility of a rigorous analysis of aims given the study population before data analysis is completed. We expect research teams to undertake MDES calculations at the beginning of their study and again at the interim or final reporting stage. This is to ensure that we fund studies designed to detect meaningful effects of interventions of interest and to reduce the risk of type II errors (the study incorrectly concludes there is no statistically significant effect owing to the true effect being too small to detect given the study’s sample size). This issue is particularly relevant for teams using smaller longitudinal or cohort datasets. Our aim is to reduce youth violence and its two most common outcomes are offending via administrative or self-report data (e.g. the SRDS), and the strengths and difficulties questionnaire (SDQ), although it does also commission studies with other primary outcomes.

The MDES criteria provides a broad rule of thumb on the likely power of the study, at the beginning of the study, and provides a useful guide to evaluators on our expectations of study size and power. But it cannot replace detailed sample size calculations using assumptions based on evidence.

We encourage evaluators to use the DELTA[8] guidance in determining the target difference for sample size calculations, including searching the relevant literature and working with stakeholders to identify a difference that is meaningful and important enough to change practice. These sample size calculations should account for the distribution of the outcome measure. Justification can be made to adjust MGs up or down by one, up to a maximum of 4 MGs, where a strong rationale using the DELTA guidance can be provided.

The MDES thresholds indicated in the table below are applicable to all YEF studies and is inclusive of all outcome distributions[9] (continuous, dichotomous, count, etc.), unless in the protocol the evaluators have provided a justified exception for a higher MDES i.e. when detecting small effects is not feasible, meaningful, or practical given the study's constraints.

The thresholds below have been adapted from EEF's padlock system, following a review[10] of effect sizes across all studies included in the YEF Evidence and Gap Map and in consultation with our Technical Advisory Group. These thresholds are consistent with the thresholds in the YEF Magnifying Glass Guidance for impact evaluations.

TABLE 2. MDES AND ASSOCIATED MAGNIFYING GLASSES RATING

Magnifying glasses (MGs)	**Offending** *(measured through admin data or SRDS)*	**SDQ Total difficulties**	**Other outcomes**
5			
4	<= 0.19	<= 0.39	<= 0.29
3	0.2- 0.29	0.4- 0.49	0.3- 0.39
2	0.3-0.39	0.5-0.59	0.4-0.49
1	0.4-0.49	0.6-0.69	0.5-0.59

[8] Cook, J. A., Julious, S. A., Sones, W., Hampson, L. V., Hewitt, C., Berlin, J. A., Ashby, D., Emsley, R., Fergusson, D. A., Walters, S. J., Wilson, E. C. F., MacLennan, G., Stallard, N., Rothwell, J. C., Bland, M., Brown, L., Ramsay, C. R., Cook, A., Armstrong, D., Altman, D., & Vale, L. D. (2018). DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ*, *363*, k3750. https://doi.org/10.1136/bmj.k3750

[9] Applying different thresholds for different outcomes, for different subsets of outcomes (e.g. violent crime and non-violent crime as distinct subsets of all crime) *and* for different distributions would mean an impractical number of sets of thresholds.

[10] YEF Effect Size Database: https://youthendowmentfund.org.uk/reports/effect-size-database/

0	>=0.5	>=0.7	>=0.6

Criterion 3: Threats to internal validity

The magnifying glass ratings for SDA impact studies can be adjusted downward in response to potential threats to internal validity. Potential threats to internal validity are the following:

1. Confounding
2. Concurrent interventions
3. Contamination, spillover effects and misclassification of interventions
4. Implementation fidelity and compliance with the intervention
5. Attrition and missing data
6. Measurement of outcomes
7. Selective reporting

To determine whether an adjustment to the magnifying glasses rating needs to be made, the reviewer will have to determine a) which threats are present, b) the severity, and c) likely direction of bias (i.e., towards or away from zero).

Please use your expert judgement and the signalling questions for each criterion to estimate whether these threats are unknown, low, moderate or high, and in which direction they likely bias results.

TABLE 3. ADJUSTMENTS TO MAGNIFYING GLASSES BASED ON THREATS TO INTERNAL VALIDITY

Weighting of threats by level of risk and direction of bias	**Adjustment to magnifying glasses**
Up to two threats classified as 'moderate risk' or 'no information available'; **AND** the direction of any likely biases is unknown or operates in opposite directions; **AND** all other threats deemed as 'low risk'	No adjustment made
• Up to four threats classified as 'moderate risk' or 'no information available' and the directions of any likely biases are unknown; **OR** • Up to two threats classified as 'moderate risk' or 'no information available' and the direction of any likely biases operates in the same direction; **OR** • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1

● One threat classified as 'high risk' and two threats are classified as 'moderate risk' or 'no information available'; **OR** ● Two or more threats are classified as 'high risk'	-2

Threats to internal validity (1): Confounding

A confounder is a variable that is correlated with receiving an intervention and has an independent impact on outcomes. Confounding can be time-invariant when it is based on characteristics that do not change over time, e.g. gender; or time-variant, when it is related to characteristics that change over time, e.g. a pupil's attitude towards school. Furthermore, confounding can be based on variables that are observable and measurable, or on variables that are unobservable and unmeasurable.

Guidance questions (all designs)

1. What are potential confounders for the intervention and their likely effects on outcomes?
   - Are they measured with errors in a way that is correlated with the intervention and outcomes?
   - Have the authors discussed or illustrated the relevant confounders (e.g., using a Directed Acyclic Graph)?
2. What type of confoundedness is controlled by the chosen design?
   - Which are the identification assumptions?
   - What evidence do the authors present for the assumptions of **exchangeability/ignorability**[11] and **positivity**[12] in causal inference?
3. Variables that are measured after the treatment, including those that might be affected by the treatment (mediating variables) should not be controlled for in the statistical model. This would produce biased estimates of impact, unless the model can account for this bias, for example in the case of Marginal Structural Models.
4. Is there balance in covariates between the treatment and control groups?

[11] Exchangeability means that the counterfactual outcome and the treatment are independent. In other words, the treated and untreated are exchangeable when the treated would have experienced the same average outcome as the untreated, had they remained untreated; and vice versa. Conditional exchangeability is also known as 'weak ignorability' or 'ignorable treatment assignment' in statistics, 'selection on observables' in social sciences, and 'no omitted variable bias' or 'exogeneity' in econometrics. Hernán, M. A., & Robins, J. M. (2025). *Causal inference: What if.* Chapman & Hall/CRC. https://miguelhernan.org/whatifbook

[12] Positivity means that all individuals in the population of interest have a non-zero probability of being assigned to each of the treatment levels. Hernán, M. A., & Robins, J. M. (2025). *Causal inference: What if.* Chapman & Hall/CRC. https://miguelhernan.org/whatifbook

5. Are sensitivity analyses run where important confounders are controlled for, especially those for which imbalances are found[13]?
6. Consider sample size when assessing balance[14]. Small studies are more likely to have imbalance due to chance.

The guidance questions above are relevant to all SDA impact designs. However, studies that use one of the more commonly funded YEF designs, specifically Regression Discontinuity Designs, Difference-in-Differences (including Synthetic Controls) or designs relying on Matching / Weighting will be assessed according to the design-specific confounding criteria in the following sections. For teams that use other approaches, studies will be assessed using the confounding criteria in the table in this section. This includes studies using Interrupted Time Series analysis and Marginal Structural Models.

In the study plan and final report, research teams should describe to what extent confounding is likely to be a threat to the internal validity, and the relevant and important confounders (grounded in theory or existing evidence) that they are able and not able to convincingly control for.

They should present results of tests and robustness checks that have been established for their method. Teams and peer reviewers may refer to recommendations within other design-specific sections in this guidance, for example, some of the Difference-in-Differences guidance below will be relevant for Synthetic Control studies.

TABLE 4. RISK LEVEL BASED ON CONFOUNDING

Description of variables predictive of the intervention and the outcome and justification of approach to control for these variables	Balance in observable characteristics between groups	Multiple specifications	Robustness checks	Risk level
Good	Good	Explored and find similar results	Considered	Low *Risk level is low only if all of these*

[13] It is likely that not all required sensitivity analyses for confounders can be pre-specified in the analysis plan. We encourage teams to list any uncertainties, along with suitable sensitivity analyses and decision rules, based on their understanding of the context, theory and existing evidence. Teams should also be able to justify any tests in the interim/final report that were not pre-specified.

[14] Since quasi-experimental designs generally require larger sample sizes than randomised controlled trials, teams may consider benchmarking against a randomised controlled trial with treatment and control groups the same size as their treatment group.

				*conditions are met (**AND** logic).*
Satisfactory	Small differences that are controlled for analytically with alternative methods	Explored but results depend on the method chosen	n/a	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
Unsatisfactory – failing to consider some relevant confounders	Large imbalances that are not accounted for	n/a	n/a	High *Risk level is high as soon as one of these conditions is met (**OR** logic).*

Confounding – Considerations for specific research designs

Regression discontinuity design

RDD.1. Describe the process which determines the cut-off and how it defines treatment allocation.

RDD.2. For (i), present graphical evidence of the discontinuity in treatment assignment around the threshold.

RDD.3. For (ii), the assumption would be violated if individuals have control over the value of the assignment variable around the threshold, meaning that they can (at least imperfectly) *choose* whether they receive the intervention or not.

RDD.3.1. Run balance tests on observable pre-intervention characteristics. These tests are expected to be met in the area surrounding the arbitrary cut-off. Balance tests could be included for several widths of the inclusion window. As with other balance tests, this can't rule out imbalance in unobservable characteristics.

RDD.3.2. Run density checks of the running variables at either side of the cut-off, for example McCrary Manipulation Test.

RDD.4. Run additional robustness checks including:

RDD.4.1. Different functional forms of the assignment variable. Note that in an infinitesimally narrow window, any functional form of the assignment variable could be approximated with a linear function.

RDD.4.2. Different widths of the assignment window.

RDD.4.3. A broad range of relevant control variables.

Considerations depending on the design: Regression discontinuity designs

- Is there evidence of a discontinuity in the probability to be assigned to treatment around the cut-off? Is the discontinuity sharp? If teams are using a fuzzy regression discontinuity design, they should provide a strong justification and evidence for its validity.
- Is there evidence of manipulation of the running variable or any other variable around the cut-off?
- Are the results robust to sensitivity analyses, including covariates, testing different inclusion windows and functional forms of the running variable?

TABLE 5. RISK LEVEL BASED ON CONFOUNDING FOR REGRESSION DISCONTINUITY DESIGNS

Discontinuity in treatment allocation around cut off	**Discontinuity in the assignment variable and other covariates**	**Appropriate robustness checks show…**	**Risk level**
Sharp	No evidence of discontinuity	Similar results	Low *Risk level is low only if all of these conditions are met (**AND** logic).*
Fuzzy	Limited evidence of discontinuity (manipulation in assignment variable or other covariates around the cut-off)	Some differences in the impact estimates	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
No evidence of discontinuity	Evidence suggestive of discontinuity in assignment variable and other covariates around the cut-off	Large differences in impact estimates	High *Risk level is high as soon as one of these conditions is met (**OR** logic).*

<u>Difference-in-Differences</u>

DD.1. Provide contextual information describing the quasi-experimental variation that creates a feasible comparison group, including definition of groups, the precise timing of the intervention period and whether the timing of the intervention varies by participant / unit. Provide evidence suggesting whether shocks after intervention delivery started can be expected to differentially affect any of the groups (and thus be conflated with the intervention effects).

DD.2. Compare pre-intervention trends in outcomes between both groups. This can include in-time placebos where a "placebo treatment period" is identified before the actual intervention occurred. The expected treatment effect for the placebo treatment period should be indistinguishable from zero.

DD.3. Run additional robustness checks which may include:

DD.3.1. Tests of balance in pre-intervention characteristics. Even if balance is not required to assess the validity of the approach, it is likely to make the "parallel trend assumption" more tenable. Using Matched Diff-in-Diffs minimises the imbalance in observable characteristics.

DD.3.2. Analytical models including other control variables

DD.3.3. Estimation of treatment effects for each period of the intervention when the intervention collects outcome data for several periods. This could provide information on how treatment effects vary over time.

DD3.4. Use of a triple difference-in-difference approach, when there is access to an appropriate additional comparison group and a simple Diff-in-Diffs approach might still be affected by unobserved, time-varying confounders.

Considerations depending on the design: Difference-in-Differences (DD)

- Is there evidence of parallel trends before the intervention starts?
- Is there evidence that any other shocks were common to both treatment and comparison group?

TABLE 6. RISK LEVEL BASED ON CONFOUNDING FOR DIFFERENCE-IN-DIFFERENCES DIESIGNS

Parallel trends assumption	**Risk level**
Evidence suggests assumption is met (including in-time and/or in-space placebo tests) **AND** matched Difference-in-Differences is used	Low
Evidence suggests assumption is met (including in-time and/or in-space placebo tests)	Moderate
Weak or no evidence of parallel trends is presented	High

Matching/Weighting

MAT.1. Explain how different variables are expected/hypothesised to be correlated with the treatment status and outcomes (i.e. confounders that will be considered). A key component of these evaluations requires exploring the validity of these hypothesised relationships.

MAT.2. Explore the sensitivity of results including appropriate sensitivity analyses which may include alternative specifications of the Matching/Weighting, additional variables and, interaction effects. As there is no consensus on the primacy of one approach or a specific matching algorithm irrespective of the characteristics of the sample, it is necessary to discuss why the chosen approach is suitable to analyse the sample under study.

MAT.3. Assess the balance in the distribution of relevant covariates included in the matching/weighting between treatment and comparison groups, before and after the matching is done.

MAT.3.1. Express differences in terms of standardised differences, as those are not dependant on sample sizes. These could be accompanied by significance tests and measures of closeness-of- fit.

MAT.3.2. Assess differences in mean values and higher order moments between the groups (See Austin 2011).

MAT.3.3. When some differences remain even after matching/weighting, consider the use of alternative methods that attempt to control for some of the residual variance by including additional variables as covariates.

MAT.4. Explore the area of common support and the characteristics of those included.

MAT.4.1. Compare the characteristics of those included in the common support and those for whom no match was found. Explain whether common support is imposed, why, as well as its implications.

MAT.4.2. Consider using methods that employ information from all individuals (for example, inverse probability weighting on the propensity score). When using Inverse Probability Weighting, consider exploring the distribution of weights and including robustness excluding large weights.

MAT.5. As Matching/Weighting cannot account for unobservable heterogeneity, include additional robustness checks of the sensitivity to hidden / omitted variable bias, e.g. using Rosenbaum Bounds.

MAT.6. Select the approach to use based on its ability to reduce imbalance. It is strongly preferred that this choice is made before outcomes are observable to the research team or made independently of the outcome values.

Considerations depending on the design: Matching/Weighting

- Is the choice of variables included in the Matching/Weighting well explained? Are those predictive of the intervention take up and outcomes? Is there any meaningful variable not included?

- ➢ Is the choice of Matching/Weighting method explained and argued appropriately?
- ➢ Was the Matching/Weighting successful to balance the baseline characteristics of the groups?
- ➢ How sensitive are the results to the use of different specifications?

TABLE 7. RISK LEVEL BASED ON CONFOUNDING FOR MATCHING/WEIGHTING DESIGNS

Description of variables to be included in the matching/weighting which are predictive of the intervention and outcomes	**Balance in observable characteristics between groups (after matching/ weighting)**	**Multiple specifications**	**Robustness checks**	**Risk level**
Good	Good	Explored and find similar results	Considered	Low *Risk level is low only if all of these conditions are met (**AND** logic).*
Satisfactory	Small differences that are controlled for analytically with alternative methods	Explored but results depend on the method chosen	n/a	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
Unsatisfactory – failing to consider some relevant confounders	Large imbalances that are not accounted for	n/a	n/a	High *Risk level is high as soon as one of these conditions is met (**OR** logic).*

*For example, if multiple specifications are explored and results depend on the method chosen, this is always a moderate risk, independent of findings in the other categories.

Threats to internal validity (2): Concurrent interventions

For this criterion, we are concerned about the systematic participation of treatment units in another intervention alongside the intervention or exposure of interest, that prevents research teams from isolating the effect of interest. For example:

- A difference in differences study interested in estimating the impact of Child and Adolescent Mental Health Services (CAMHS) alone could face a bias due to concurrent interventions if participation in CAMHS led to the

automatic offer of an additional support that wasn't available more broadly.
- An Interrupted Time Series analysis interested in exploring whether youth offending changed before and after the introduction of a new Stop and Search policy could face a bias due to concurrent interventions if a youth offending-focused intervention was introduced at a similar time.

At the very minimum, SDA study teams should be describing the 'Business as Usual' provision and comparing this with the intervention or exposure of interest. If there are concurrent interventions that are common across both study groups as part of 'Business as Usual' provision (e.g., following national rollout that is simultaneously implemented), this does not introduce biases nor reduces the security of findings of the study. However, it may affect the interpretation of the results, and it would be useful for research teams wherever possible to describe the wider context of current interventions that the new intervention was introduced into. When intensive concurrent interventions are expected as part of 'Business as Usual' or comparison provision, teams could consider powering studies to detect smaller MDES compared to without these concurrent interventions.

SDA study teams may attempt to explore concurrent interventions or exposures using desk-based approaches such as reviewing existing research, using administrative or survey datasets or speaking to the relevant intervention delivery teams. If not possible, we encourage teams to make this clear in their study plan and report that they were unable to explore this issue. As above, teams should in any case be able to describe what the comparison condition entails.

TABLE 8. RISK LEVEL BASED ON CONCURRENT INTERVENTIONS

Criteria	**Risk level**
• Concurrent interventions or exposures are explored and there is no evidence suggesting differential uptake of those interventions, or no other interventions or exposures are identified or expected; **OR** • Evidence of concurrent interventions is found but controlled for analytically.	Low
• Concurrent interventions are explored **AND** there is evidence of minor differential uptake between groups which is not controlled for analytically; **OR** • No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	Moderate
Concurrent interventions are explored **AND** there is evidence of large differential uptake between groups.	High

Threats to internal validity (3): Contamination, spillover effects and misclassification of interventions

For this criterion, we are concerned with contamination and spillover effects, as well as misclassification of intervention status. Contamination is where individuals in the comparison group directly receive the intervention or exposure of interest as well as those in the treatment group. Spillovers are cases where the outcomes of individuals not in the intervention group, including those in the comparison group and those not in the study are indirectly affected by the intervention, for instance due to proximity or social networks.

Misclassification of intervention status is a systematic error where participants are incorrectly assigned to either an intervention group or a comparison group. This can occur because of:

- Assignment of participants to the intervention/exposure group or comparison group rely on events or measurements occurring *after* the start of follow up. This can create a period where the intervention group cannot experience the outcome as a result of the study design ('immortal time'). It can result in intervention participants being considered as exposed to an intervention during their follow-up time even during periods when they are not, or individuals who partially experience an intervention and experience the outcome during the immortal time getting reclassified as the comparison group.
- Poorly collected or recorded information on group status or because of how the groups are defined.

When participants are misclassified, observed differences between groups may not accurately reflect the impact of the intervention. There can be:

- Differential misclassification (e.g. more likely to misclassify exposed individuals as the comparison group if they have an "undesirable" outcome, leading to an overestimation of the intervention's effect)
- Non-differential misclassification (meaning the misclassification occurs equally in both directions across groups and independently of the outcome, which typically biases the result towards the null).

Considerations

➢ Are the criteria for what constitutes intervention and comparison status precise and unambiguous?

- Is the comparison group likely to have been affected by the intervention, either directly by receiving the intervention or indirectly (e.g, if they behaved differently as a result of the intervention, which may affect their outcomes positively or negatively)?
- Does group status rely on reliable data sources (e.g., administrative records, program attendance logs, rather than just self-report) to determine group status? If group status is based on self-report data, is there a risk of differential or non-differential recall bias by group?
- For longitudinal studies where the identification strategy relies on measurement at multiple time points, is there potential for immortal time bias?
- Are sensitivity analyses used to account for how robust results are to different assumptions about the extent and direction of group status misclassification, contamination, or spillover effects?

TABLE 9. RISK LEVEL BASED ON CONTAMINATION, SPILLOVER EFFECTS AND MISCLASSIFICATION OF INTERVENTIONS

Evidence of misclassification	**Contamination or spillover effects**	**Sensitivity analyses**	**Risk level**
Explored – no evidence	Explored – no evidence	n/a	Low *Risk level is low only if all of these conditions are met (**AND** logic).*
Explored – evidence of minor misclassification issues **OR** No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	Explored – evidence of minor contamination or spillover effects (e.g., 20% of the control units implement something similar)[15] **OR** No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	Similar findings as main analysis	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
Explored – meaningful misclassification issues	Explored – meaningful evidence of spillover effects (e.g., 50% of the control units implement something similar)	Different results than the main analysis	High *Risk level is high as soon as one of these conditions is met (**OR** logic).*

[15] Please note that this is only indicative. The decision of the relevance of the threat would depend on the judgement of the peer reviewer depending on the intensity of the misclassification issues.

Threats to internal validity (4): Implementation fidelity and compliance with the intervention

Compliance refers to the extent to which participants adhered to the assigned treatment status. Implementation fidelity is the extent to which the intervention, programme or policy was implemented as intended (as per the delivery model), in terms of content and process.

This criterion is concerned with:

- whether the intervention, compliance and implementation fidelity are well-defined, unambiguous and aligned with the identified causal mechanism and outcomes.
- whether the intervention was implemented with fidelity and compliance during the study period.

All SDA studies will be assessed against the first of these two points, i.e. whether the intervention, compliance and fidelity are well defined and aligned with the outcomes of interest. However, only some SDA impact studies will be assessed against the second point, according to the guidance below. This topic should be discussed during the set-up of the SDA study and made clear in the study plan.

Some SDA impact studies are primarily interested in the impact of participating in an intervention, *as it was delivered for those who took part*. The estimand of interest for these impact studies would be the Local Average Treatment Effect (LATE) or the Complier Average Causal Effect (CACE). For example, a study interested in the effects of *regularly attending* a youth club or *participating* in community activities on the likelihood of a young person becoming involved in violent crime. In these cases, the Not Relevant category below should be used for this part of the criterion. It is still important that the study plan and report appropriately describes the intervention of interest, including references to its critical components and methods of delivery as relevant.

This contrasts with studies that explore *the impact of an intervention offer* and consider whether the intervention was delivered with fidelity. The estimand of interest for these impact studies would be the Intention to Treat (ITT). For example, a study interested in the impact of young people being assigned to receive one-to-one mentoring on their behavioural and emotional problems, which also explored whether they participated and complied with the mentoring model and whether the impact varied by this compliance. If this is the case, the implementation fidelity and compliance part of the assessment criteria below should be applied as usual.

SDA study teams may attempt to assess fidelity and compliance using desk-based approaches such as reviewing existing research, using administrative or survey datasets or speaking to the relevant intervention developers or implementors where possible. If no information is available, teams may also consider similar policies delivered to similar populations, to allow some bounds on compliance to be estimated.

Considerations

- Was the intervention appropriately described including references to its critical components and methods of delivery? Is the intervention well-defined and unambiguous? I.e. what is the evidence for the assumption of consistency[16] in causal inference?
- Was the 'implementation logic' adequately specified to assess the fidelity with the intervention and potential effects on outcomes?
- Are deviations from ideal implementation reasonably considered "usual practice"?
- Are the requirements for participants to be considered as having adhered to their assigned treatment status, clear and unambiguous (e.g. number of sessions attended)?
- Are the levels of compliance (e.g. young person, family, school etc.) clearly specified?
- Were the intervention content and process delivered as intended?

TABLE 10. RISK LEVEL BASED ON IMPLEMENTATION FIDELITY AND COMPLIANCE WITH THE INTERVENTION

Intervention and implementation fidelity and/or compliance are well defined and aligned with the implementation logic and the causal mechanism identified in the logic model	**Implementation fidelity and/or compliance with the intervention**	**Risk level**
Yes	High **OR** Not relevant	Low *Risk level is low only if all of these conditions are met (**AND** logic).*

[16] Consistency means that the exposure or treatment is sufficiently well-defined, such that the observed outcome for every treated individual equals their outcome if they had received treatment, and the observed outcome for every untreated individual equals their outcome if they had remained untreated. Hernán, M. A., & Robins, J. M. (2025). *Causal inference: What if.* Chapman & Hall/CRC. https://miguelhernan.org/whatifbook

Yes	Moderate **OR** No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
Not well defined or poorly aligned with the logic model	Very low	High *Risk level is high as soon as one of these conditions is met (**OR** logic).*

Threats to internal validity (5): Attrition and missing data

This criterion explores the potential for bias and loss of statistical sensitivity introduced by missing data (on the outcome, intervention group status, and/or control variables), which can be a common issue in SDA impact studies. It allows for adjustments based on the total amount of missing data, differential missingness between intervention and comparison group, reason for missingness, and analyses to account for missing data.

We expect that SDA teams discuss in both their study plan and final report whether missing data are likely to have led to a threat to internal validity. This should include discussion of the extent of missingness, any known patterns of missingness (missing completely at random, missing at random, missing not at random) and reasons for missingness (e.g. information or variables predictive of missingness); whether this might vary by intervention or comparison group status; the direction of potential bias (including any selection bias; and implications on the findings and conclusions.

Missing data in SDA studies may result from participants:

- Participants being pre-excluded from a particular dataset. For example, pupils with frequently changing addresses are not always tracked by the National Pupil Database (NPD), which is more likely for low-income young people or young people with complex or difficult home situations.
    - We encourage teams to explore differences in key characteristics of participants who are and who are not included in a given dataset with national census data or refer to existing research on this issue.
    - Missingness in this form is not considered in the risk table below, but we will expect an appropriate acknowledgement and discussion of any resulting selection bias in the study plan and final report.
- Participants dropping out from a longitudinal study. For example, higher rates of school absence and special education needs status are

associated with higher non-response or dropout from the Avon Longitudinal Study of Parents and Children (ALSPAC).[17]

- Participants for whom data collection is incomplete, including missingness on the outcome, intervention group status, and/or control variables.

The YEF Analysis Guidance[18] (pp.12-15) provides guidelines on the appropriate analytical approaches (e.g. multiple imputation) depending on extent and pattern of missingness, and whether missingness is on control variables or outcomes. Teams may also refer to the Office for National Statistics' Review of Methods for Missing Data[19]. Where appropriate, teams may consider engaging with data managers for administrative and survey datasets to explore whether additional variables or proxy variables could address gaps for important missing variables.

Considerations

- ➢ What was the total amount of missing data at the unit of the intervention and the unit of analysis (if different)? For instance, if a school study with pupil-level outcomes we might ask, were all schools originally included in a school-level policy present in the end-point dataset (unit of intervention); if all schools were present, were the data available for all pupils within those schools (unit of analysis)?
- ➢ Where participants are completely missing from the dataset, is there an appropriate discussion of the extent of missingness, potential patterns and reasons for missingness, direction of any bias and implications?
- ➢ Was linkage bias likely to be an issue? For instance, if specific groups of young people are less likely to be successfully linked across datasets because of missing information.
- ➢ Was there differential missingness between intervention and control groups?
- ➢ Were observable variables predictive of missingness? Specifically, was the treatment indicator predictive of missingness, net of other covariates included in the final analysis model?

[17] Cornish, R. P., Macleod, J., Boyd, A., & Tilling, K. (2021). Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: A study using linked education and primary care data. *International Journal of Epidemiology*, *50*(1), 293-302. https://doi.org/10.1093/ije/dyaa192

[18] YEF Analysis Guidance: https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf

[19] Technical review of methods for missing data (prepared for the Office for National Statistics by Alma Economics): https://static1.squarespace.com/static/56963a52c647ad2ec2573846/t/65d87094c9708f6196b8cfbd/1708683413694/Review+of+missing+data+methods.pdf

- Are the results of the analyses accounting for missing data reasonably similar to the main analysis (e.g. point estimates are meaningful and consistent in size and direction)
- Are results robust to further sensitivity analyses to account for missing data?

TABLE 11. RISK LEVEL BASED ON ATTRITION AND MISSING DATA

Total amount of missing data	Logical connection	Differential missing data	Logical connection	Analyses accounting for missing data	Risk level
Low (<10%)	AND	No	AND	Similar to complete-cases analyses	Low
-	-	Yes	AND	Minor deviations from complete-cases analyses	Moderate
Moderate (10-20%)	AND	No	AND	Similar to complete-cases analyses	Moderate
-	-	Yes	AND	Similar to complete-case analyses	Moderate
-	-	-	-	Minor deviations from the complete-case analyses	Moderate
-	-	-	-	Differ from complete-case analyses	High
High (>20%)					High
No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.					Moderate

Threats to internal validity (6): Measurement of outcomes

This criterion is concerned with the use of reliable, valid and acceptable outcome measures that are free from ceiling/floor effects and where scorers are blind to intervention group allocation (where relevant). This criterion also explores the potential for bias from the time-ordering of events and measurement of outcomes.

Bias may be introduced if outcomes are misclassified or measured with error. Non-differential measurement error is error that is unrelated to the intervention received. This will not cause bias but can affect precision. Differential measurement error is error that is related to the intervention received and can bias the intervention-outcome relationship. This is often referred to as detection bias, which can arise when (i) if scorers are aware of intervention received (particularly when the outcome is subjective); (ii) different methods (or intensities of observation) are used to assess outcomes of participants receiving different interventions; and (iii) measurement errors are related to intervention received (or to a confounder of the intervention-outcome relationship). Blinding of scorers aims to prevent systematic differences in measurements according to intervention received. However, blinding is frequently not possible or not performed for practical reasons.

In SDA impact studies where study teams do not control the timing of interventions and measurement of outcomes and where the event timing isn't clearly established, ambiguous temporal ordering may result (where it is unclear which occurred first). This includes reverse causality bias, when a study assumes a cause-and-effect relationship in one direction when the opposite is true but could also include issues of bidirectional causality. For example, there may be a risk of reverse or bidirectional causality in a study exploring the effects of living in a neighbourhood experiencing high rates of stop and search on young people’s mental health and offending rates, if there is not a clear temporal ordering of the policy introduction or change in intensity and data collection on the outcomes of interest over time. Teams could consider exploiting exogenous shocks to the exposure of interest (e.g., intensity of stop and searches) to overcome these issues.

Considerations

- Are the outcome measures a valid and reliable measure of the relevant construct for the population of interest? If a proxy is used, is there evidence that the proxy is highly correlated with and accurately captures the outcome/concept of interest?

- Is there evidence of measurement error that may have biased the results?
- Are the outcome tests administered and scored independently, or in ways that minimise differences between groups?
- Are the outcome tests capable of identifying differences across the whole distribution, i.e. is there strong existing evidence on floor/ceiling effects, for example a measure with many participants with a score/outcome of zero?
- If floor/ceiling effects are found, do the researchers discuss the implications of the problem and run sensitivity analyses that consider this?
- Does the proposed causal direction make logical sense given the existing theory and evidence base?
- Are the timings of the intervention, outcome, and potential confounders clearly recorded? Is there a clear and consistent "time zero" for participants from which follow-up begins, and is intervention status determined after this point? Is there any inherent aspect of their measurement that could introduce ambiguity in their sequence?
- Was the outcome measured at baseline, before the intervention was introduced to ensure that the outcome was not already present?
- Is there evidence that the intervention or exposure (presumed cause) occurred before the outcome(s) (presumed effect)?

TABLE 12. RISK LEVEL BASED ON MEASUREMENT OF OUTCOMES

Reliability and validity with target population	**Ceiling/floor effects**	**Scorers blind to intervention allocation**	**Ambiguous temporal ordering**	**Risk level**
Outcome tests have been thoroughly justified in relation to reliability, validity, utility and acceptability with target population	No ceiling/floor effects are found	Tests are administered and scored blinded to allocation or with very minor judgments	Causal direction makes logical sense given existing theory and evidence base; **AND** evidence that the intervention (presumed cause) occurred before the outcome measurement (presumed effect).	Low *Risk level is low only if all of these conditions are met (**AND** logic).*
Outcome tests have been shown to be moderately reliable and valid with the target population, or evidence on reliability and validity comes from a different population	Minor ceiling/floor effects are found and controlled for analytically	Tests involve minor judgement from assessors who are not blinded to allocation, but safeguards are included to ensure quality	• Causal direction ambiguous given existing theory and evidence base; **AND** the order of the intervention and outcome are measured at the same time but controlled for analytically (e.g. using time series data and appropriate lagged values of the relevant variables in the analysis); **OR** • No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	Moderate *Risk level is moderate as soon as one of these conditions is met (**OR** logic).*
Outcome tests have poor validity or reliability for the target population	Large ceiling/floor effects are found (e.g., 20% of the sample are	Tests involve important judgement from assessors who are not blinded to allocation with no safeguards in place to guarantee independence	Causal direction ambiguous given existing theory and evidence base; **AND** the intervention and outcome are measured at the same time or the order is unclear; **AND**	High *Risk level is high as soon as one of these*

	in the top/bottom 10% of marks)[20]		it is impossible to determine the sequence of events from the analysis.	*conditions is met (**OR** logic).*

[20] Please note that this is only indicative. The decision of the relevance of the threat would depend on the judgement of the peer reviewer depending on the intensity of the floor/ceiling effect.

Threats to internal validity (7): Selective reporting

There may be a greater risk of selective reporting in SDA studies, given there are likely to be more potential variables that can be tested and many combinations of models that can be fitted. We consider selective reporting for those cases where results are presented only for i) a particular outcome measure and specification; ii) a specific analytical approach; or, iii) a subset of participants; contravening what is specified in the study plan.

It may not be possible for SDA impact study teams to pre-specify all statistical methods in advance before seeing the data. We ask research teams to prepare a prospective study plan with their preferred models and specifications, including the assumptions and justifications for their preferred approach and any decision rules they will be applying. Any uncertainties at this stage should be identified and listed, and there should be clear strategies for how the research teams will resolve or investigate these. For established datasets, we expect that the statistical analysis plan can be pre-specified, using data dictionaries and published descriptive statistics, prior to accessing data. For less established datasets, this may need to be done after teams gain access to the data and produce descriptive statistics, i.e. at the interim reporting stage.

At the interim reporting stage, we expect the research team to follow what is in the prospective documentation as much as possible and to clearly report on and fully justify any deviations from the plan. Unlike trials, deviations may be likely and should not necessarily be rated as a threat to validity if they are fully justified and plausible. In all cases, this process should be documented in detail in the final report to support assessment of this criterion.

As done for YEF evaluation reports, peer reviewers will compare the final report against the pre-specified analyses in the study plan, to track deviations and ensure any deviations made are fully justified.

Considerations

- Are analyses pre-specified and conducted according to plan?
- If the team have pivoted away from the plan in the final report, is there a plausible rationale, such as high data missingness, for their strategy?

TABLE 13. RISK LEVEL BASED ON SELECTIVE REPORTING

Criteria	Risk level
• A comprehensive prospective document is published and followed; **OR** • A comprehensive, prospective document is published, and all deviations are fully justified.	Low
A comprehensive prospective document is published, but with minor deviations that are not fully justified.	Moderate
• A comprehensive prospective document is not published; **OR** • Important deviations from the proposed analysis that are not fully justified.	High

Appendix 1: Template assessment form

Please complete this form for each primary outcome. Magnifying glasses will only be assigned to the primary outcome. Separate padlock ratings may be assigned where there is more than one primary outcome.

Project name	
Name of reviewer	
Date assessment submitted	
What is/are the primary outcome(s) of the evaluation?	

Assessment Outcome 1:

Please highlight the cells that represent the rating you've given the evaluation. The initial score is the lowest magnifying glass rating out of all scores assigned. See also the worked examples.

Rating	**Design**	**MDES** Outcome: Threshold*	**Initial score**	**Adjustments**	**Final score**
5 🔍	Randomised design	NA		*Adjustment for threats to internal validity* *(Please select and describe threats in the table below)*	
4 🔍	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	Offending: <= 0.19 SDQ tot: <= 0.39 Other: <= 0.29			
3 🔍	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	Offending: 0.2 – 0.29 SDQ tot: 0.4 – 0.49 Other: 0.3 – 0.39			
2 🔍	Design for comparison that considers selection only on some relevant confounders	Offending: 0.3 – 0.39 SDQ tot: 0.5 – 0.59 Other: 0.4 – 0.49			

1 🔍	Design for comparison that does not consider selection on any relevant confounders	Offending: 0.4 – 0.49 SDQ tot: 0.6 – 0.69 Other: 0.5 – 0.59			*0* *–1*		
0 🔍	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6			*–2*		

*MDES requirements vary by outcome measurement. Offending: Offending data collected through self-report or admin data; SDQ tot = SDQ total difficulties score; Other: all other outcomes, incl. SDQ externalising and internalising

Adjustment due to threats to internal validity needed?

Threat		**Threat assessment**	**Comments**	**Direction of effect**
1	Confounding	Low/ moderate/ high risk		
2	Concurrent interventions	Low/ moderate/ high risk		
3	Contamination, spillover effects and misclassification of interventions	Low/ moderate/ high risk		
4	Implementation fidelity and compliance with the intervention	Low/ moderate/ high risk		
5	Attrition and missing data	Low/ moderate/ high risk		
6	Measurement of outcomes	Low/ moderate/ high risk		
7	Selective reporting	Low/ moderate/ high risk		

Please use this table to assess the previous table and identify how the initial rating needs to be adjusted. Then add the adjustment to the scoring table.

Weighting of threats by level of risk and direction of bias	**Adjustment to magnifying glasses**
Up to two threats classified as 'moderate risk' or 'no information available'; **AND** the direction of any likely biases is unknown or operates in opposite directions; **AND** all other threats deemed as 'low risk'	No adjustment made
• Up to four threats classified as 'moderate risk' or 'no information available' and the directions of any likely biases are unknown; **OR** • Up to two threats classified as 'moderate risk' or 'no information available' and the direction of any likely biases operates in the same direction; **OR** • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1
• One threat classified as 'high risk' and two threats are classified as 'moderate risk' or 'no information available'; **OR** • Two or more threats are classified as 'high risk'	-2

Appendix 2: Worked example

Please complete this form for each primary outcome. Magnifying glasses will only be assigned to the primary outcome. Separate padlock ratings may be assigned where there is more than one primary outcome.

Project name	Example 1
Name of reviewer	Rose Tyler
Date assessment submitted	20/03/25
What is/are the primary outcome(s) of the evaluation?	Offending

Assessment Outcome 1:

Please highlight the cells that represent the rating you've given the evaluation. The initial score is the lowest magnifying glass rating out of all scores assigned. See also the worked examples.

Rating	**Design**	**MDES** Outcome: Threshold*	**Initial score**	➡	**Adjustments**	➡	**Final score**
5 🔍	Randomised design	NA	4		Adjustment for threats to internal validity (Please select and describe threats in the table below)		4
4 🔍	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	Offending: <= 0.19 SDQ tot: <= 0.39 Other: <= 0.29 **MDES 0.14**					
3 🔍	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression	Offending: 0.2 – 0.29 SDQ tot: 0.4 – 0.49 Other: 0.3 – 0.39					

	Analysis with variables descriptive of the selection mechanism)				0		
2 🔍	Design for comparison that considers selection only on some relevant confounders	Offending: 0.3 – 0.39 SDQ tot: 0.5 – 0.59 Other: 0.4 – 0.49					
1 🔍	Design for comparison that does not consider selection on any relevant confounders	Offending: 0.4 – 0.49 SDQ tot: 0.6 – 0.69 Other: 0.5 – 0.59					
0 🔍	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6					

**MDES requirements vary by outcome measurement. Offending: Offending data collected through self-report or admin data; SDQ tot = SDQ total difficulties score; Other: all other outcomes, incl. SDQ externalising and internalising.*

Adjustment due to threats to internal validity needed?

Threat		Threat assessment	Comments	Direction of effect
1	Confounding	Low	This was designed as a matched difference-in-differences study. Variables included in the matching are well detailed and argued, achieving good balance in relevant variables (all with standardised differences smaller than 0.06SD, see Table 3). Evidence supportive of parallel trends before intervention is provided (Figure 1) and improved by the additional matching between intervention and control participants (Figure 2).	No bias (low risk)
2	Concurrent interventions	Moderate	No information of concurrent interventions was available.	No bias (no info)
3	Contamination, spillover effects and misclassification of interventions	Low	As the intervention group participants were identified using administrative data, there is no expectation of potential experimental effects in the comparison group.	No bias (low risk)

4	Implementation fidelity and compliance with the intervention	Moderate	Compliance and fidelity appropriately defined and in line with the causal mechanism (p.34), but no information available on these aspects.	No bias (no info)
5	Attrition and missing data	Low	Missing data was low (3% outcomes, 9% overall, see Table 6) so the complete case analysis is expected to be unbiased.	No bias (low risk)
6	Measurement of outcomes	Low	The outcome measure is a high-stakes national assessment for this year group so it can be deemed as independent to the intervention. There were no relevant changes to the assessment during the study period.	No bias (low risk)
7	Selective reporting	Low	This study was pre-registered and the analytical approach was identified before outcomes were observed.	No bias (low risk)

Please use this table to assess the previous table and identify how the initial rating needs to be adjusted. Then add the adjustment to the scoring table.

Weighting of threats by level of risk and direction of bias	**Adjustment to magnifying glasses**
Up to two threats classified as 'moderate risk' or 'no information available'; **AND** the direction of any likely biases is unknown or operates in opposite directions; **AND** all other threats deemed as 'low risk'	No adjustment made
• Up to four threats classified as 'moderate risk' or 'no information available' and the directions of any likely biases are unknown; **OR** • Up to two threats classified as 'moderate risk' or 'no information available' and the direction of any likely biases operates in the same direction; **OR** • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1
• One threat classified as 'high risk' and two threats are classified as 'moderate risk' or 'no information available'; **OR** • Two or more threats are classified as 'high risk'	-2