Youth Endowment Fund

Magnifying Glass Guidance

YOUTH
ENDOWMENT
FUND

YOUTH ENDOWMENT FUND

Contents

Acknowledgement

We are very grateful to the Education Endowment Foundation for their comprehensive work on security ratings that provided the essential framework for this guidance. We would also like to thank our Technical Advisory Group for their thoughtful reflection on the guidance. Finally, we would like to extend our gratitude to Camilla Neville. Her technical expertise, dedication and thoughtful comments have been invaluable in adapting the guidance to our context.

Purpose

This document describes the process to arrive at a security rating for YEF evaluations. It is written **for members of the YEF panel of evaluators** who conduct the assessment for evidence security of our evaluations.

The YEF's magnifying glass (MG) evidence security rating assessment system is **based on the padlock system developed by the Education Endowment Foundation (EEF) but is adapted to the youth justice sector** and associated outcomes. All adaptations have been discussed and approved by our Technical Advisory Group, a panel of world-class experts in evaluation.

Like the EEF's system, the magnifying rating from 0-5 represents **to what extent one might expect to find the same outcome of an intervention in a similar context.** It does not include an assessment of the size or direction of effect.

While information reduction is always controversial in scientific contexts, to achieve the YEF's mission of preventing young people becoming involved in crime **it is crucial that we can communicate to practitioners to what extent they can trust the findings of an evaluation.** This guidance describes how peer reviewers can arrive at a security rating for an evaluation.

Process

YEF assigns the final security rating, considering assessments by two peer reviewers from YEF's panel of evaluators, the author's opinion, and where needed, arbitration through YEF's Technical Advisory Group.

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[1].

The security rating is determined by four criteria: design, minimum detectable effect size, attrition, and threats to internal validity. These are not the only things

[1] On the rare occasions where unsurmountable disagreements were to arise between the peer reviewers, the YEF, and the author, the YEF in consultation with the Technical Advisory Group will make the final decision.

that are important in determining the security of the results. They are, however, the key factors that differentiate the security of findings for YEF-funded 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.

The four criteria are:

- **Design:** The quality of the design used to create a comparison group with which to determine an unbiased measure of the impact on the primary outcome(s). Higher padlocks are given for designs better suited to deal with confounding.
- **MDES:** The minimum detectable effect (MDES) that the trial was powered to achieve at randomisation, which is heavily influenced by sample size.
- **Attrition:** The level of overall drop-out from the evaluation treatment and control groups, measured at the level of the young person regardless of the level of randomisation.
- **Threats to internal validity:** A series of markers that explain whether the results could be explained by anything other than the intervention.

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

- **Step 1:** The first three criteria – Design, MDES, and Attrition – are awarded a rating on a scale from 0 to 5.

- **Step 2:** An interim magnifying glasses rating is determined by the lowest of these three ratings.

- **Step 3:** The interim magnifying glasses rating can be adjusted upwards or 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 to your assigned evaluation, please complete the Error! Reference source not found.. **Please complete an assessment form for each rating – we hope it is clear enough to guide you through the assessment without re-reading the full guidance for every assessment.** In the rare cases where an evaluation has multiple primary outcomes, each outcome will be assigned a security rating. Please complete the assessment form separately for each outcome.

Appendix 1 shows three worked examples. Once the security rating has been agreed, the appendix will be added into the final report to summarise the reasons for the decision

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*	**Attrition**	**Initial score**		**Adjustments**		**Final score**
5 🔍	Randomised design	Offending: <=0.1 SDQ tot: <= 0.3 Other: <= 0.2	0-10%		→	*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: 0.11 – 0.19 SDQ tot: 0.31 – 0.39 Other: 0.21 – 0.29	11-20%					
3 🔍	Design for comparison that considers selection on all relevant observable	Offending: 0.2 – 0.29 SDQ tot: 0.4 – 0.49	21-30%					

	confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	Other: 0.3 – 0.39				*+1*		
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	31-40%			*0* *-1*		
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	41-50%			*-2*		
0	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6	>50%					

*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		
2	Concurrent interventions	Low/moderate/high/ n/a		
3	Experimental effects and contamination	Low/moderate/high/ n/a		
4	Implementation fidelity and compliance	Low/moderate/high/ n/a		

5	Attrition adjustments	Low/moderate/high/ n/a		
6	Measurement of outcomes	Low/moderate/high		
7	Selective reporting and data availability	Low/moderate/high		

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**
Missing data due to attrition is classified as 'low risk of bias'.	+1
Up to two threats are classified as 'moderate risk' and the direction of the likely biases is unknown or operates in opposite directions.	No adjustment made
• Up to four threats are classified as 'moderate risk' but the directions of biases are unknown; OR • Up to two threats are classified as 'moderate risk' with the same likely direction of bias; OR • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1
• One threat is classified as 'high risk' and two threats are classified as 'moderate risk'; OR • Two or more threats are classified as 'high risk'	-2

Criterion 1: Design

The quality of the design is the validity of the comparison group used as an estimate of the counterfactual.

Table 1 summarises the scale for rating quality of design. YEF impact evaluations are expected to be designed to attain at least 3 magnifying glasses (MG) except in rare circumstances.

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, reports of unbiased randomisation, 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. All YEF impact studies will be designed to achieve at least 3 MGs, except in rare circumstances.

Criterion 2: Minimum Detectable Effect Size (MDES)

This is the ability of the study to detect a given impact. MDES is highly dependent upon the sample size but is also influenced by the intra-cluster correlation (ICC) and correlation between the baseline covariates and the post-test.

The rating on this criterion should be determined by the **MDES at the start of the study** (i.e. at randomisation for an RCT). The YEF's 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 point of randomisation, and provides a useful guide to evaluators on YEF expectations of study size and power. But it cannot replace detailed sample size calculations using assumptions based on evidence. Evaluators must also include a measure of the ultimate statistical uncertainty around all ES in the final report (e.g. using a confidence interval, see YEF analysis guidance).

The YEF encourages evaluators to use the DELTA 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. Justification can be made to adjust MGs up or down by one where a strong rationale using the DELTA guidance can be provided.

The MDES of all YEF studies should adhere to the thresholds indicated in the table below, 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.

Table 1. MDES at design stage and associated magnifying glasses rating.

Magnifying glasses (MGs)	**Offending** *(measured through admin data or SRDS)*	**SDQ Total difficulties**	**Other outcomes**
5	<= 0.1	<= 0.3	<= 0.2
4	0.11- 0.19	0.31- 0.39	0.21- 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
0	>=0.5	>=0.7	>=0.6

Criterion 3: Attrition

Attrition should be measured at the level of the young person regardless of the level of randomisation (i.e. individual level attrition should be used for cluster randomised trials) and should be measured as the drop-out from the initial sample (i.e. those included in the randomisation for RCTs) to the point of analysis.

YEF has decided to use an overall attrition scale, rather than a combination of overall and differential attrition (such as the What Works Clearinghouse uses). The scale is shown in *Table 3. Attrition thresholds for the six magnifying glasses ratings.*

Table 3. Attrition thresholds for the six magnifying glasses ratings.

Attrition	**Rating**
0-10%	5🔍
11-20%	4🔍
21-30%	3🔍
31-40%	2🔍
41-50%	1🔍
>50%	0🔍

While the attrition thresholds are ambitious, we recognise the challenging contexts in which we commission evaluations and the vulnerable populations our programmes serve. Therefore, evaluators can gain a padlock under Criterion 4e if there is no differential attrition and authors can show that analyses accounting for missing data yield similar results as complete-case analyses (i.e., the risk of bias through attrition is low).

Criterion 4: Threats to internal validity

The magnifying glass ratings for our evaluations are dynamic and can be adjusted upward or downward in response to the changing risk levels e.g. threats to internal validity allow for a downwards adjustment of the magnifying glasses rating, or, when the risk of bias through attrition is low, for an increase in rating.

Threats to internal validity before the intervention starts:

1. Confounding

Threats to internal validity after the intervention starts:

2. Concurrent interventions
3. Experimental effects and contamination
4. Implementation fidelity and compliance with the intervention
5. Attrition
6. Measurement of outcomes
7. Selective reporting and data availability

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.

Please use your expert judgement and the signalling questions for each criterion to estimate whether these threats are low, moderate or high, and in which direction they likely bias results. If incomplete or missing information does not allow you to assess the likelihood of a given threat to validity, please clearly state this in the assessment form. Overall, take a **'benefit of the doubt' approach**: If there is no indication that the respective threat was present, rate it as low and include a respective comment.

Weighting of threats by level of risk and direction of bias	**Adjustment to magnifying glasses**
Missing data due to attrition is classified as 'low risk of bias'. (see Criterion 4.5)	+1
Up to two threats are classified as 'moderate risk' and the direction of the likely biases is unknown or operates in opposite directions.	No adjustment made
• Up to four threats are classified as 'moderate risk' but the directions of biases are unknown; **OR** • Up to two threats are classified as 'moderate risk' with the same likely direction of bias; **OR** • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1

- One threat is classified as 'high risk' and two threats are classified as 'moderate risk'; OR - Two or more threats are classified as 'high risk'	-2

1. Confounding (before the intervention starts)

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?
2. What type of confoundedness is controlled by the chosen design?
   - Which are the identification assumptions?
3. Variables 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.
4. If imbalances on observable variables occur, try to assess whether those are due to chance or a deviation from a random assignment. E.g., do they occur in many variables and always in the same direction? (Cannot rule out imbalance in unobservable characteristics.)
5. Are sensitivity analyses run where important confounders are controlled for, especially those for which imbalances are found?
6. Consider sample size when assessing balance as small studies are more likely to have imbalance due to chance.

RCTs

Recommendations for RCTs

RCT.1. Randomisation should always be conducted independently by a member of the evaluation team using appropriate methods which should be fully described in the protocol and the statistical analysis plan (SAP) to enable replication. It is advisable to disclose the code used to generate the allocation as an appendix in these documents.

RCT.2. Run balance tests based on observable pre-intervention characteristics recognising that this does not rule out imbalances in unobservable characteristics.

RCT.3. In the case that an imbalance is found, assess whether this is likely to be due to chance or because the randomisation procedure was subverted.

RCT.4. Run sensitivity analyses controlling for variables where imbalance was found by including these variables and assessing the stability of the main results.

Considerations depending on the design: RCTs

Please determine risk of bias using the following criteria and thresholds:

- How was the allocation sequence conducted, and by whom?
- Is there evidence of imbalance of demographic characteristics and/or outcome measure at baseline? If yes, what is its size (in SD)?
- If an imbalance was found, did the evaluator conduct a sensitivity analysis? Was this method appropriate to account for the imbalance? Were the results different?

Criteria	**Risk level**
Adequate allocation sequence with concealed assignment **AND** imbalance of **0.00 – 0.05 SD** in variables identified as important predictors of outcome	Low
Imbalance of **0.05 – 0.10 SD** in variables identified as important predictors **AND** controlled for in a regression model	Moderate
Inadequate description of allocation sequence **OR** imbalance of **0.05 – 0.10 SD** in variables identified as important predictors AND not controlled for in a regression or that meaningfully affects the estimate of impact **OR** imbalance **>0.1 SD** in variables identified as important predictors	High

RDDs

Recommendations for RDDs

RDD.1. Describe the nature of 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 (RDD)

- Is there evidence of a discontinuity in the probability to be assigned to treatment around the cut-off? Is the discontinuity sharp?
- 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?

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).*

Difference in Difference designs

Recommendations for DDs

DD.1. Provide contextual information describing the quasi-experimental variation that creates a feasible comparison group, including definition of groups and the precise timing of the intervention period. 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.

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?

Parallel trends assumption	**Risk level**
Evidence suggests assumption is met (including in-time and/or in-space placebo tests) **AND** matched Diff-in-Diffs 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 Designs

Recommendations for 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, consider including additional robustness checks of the sensitivity to hidden bias, e.g. using Rosenbaum Bounds.

MAT.6. Select the approach to used based on its ability to reduce imbalance. It is strongly preferred that this choice is made before outcomes are observable to the research team.

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?

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.

2. Concurrent interventions

For this criterion, we are concerned about participation of treatment units in other interventions. If concurrent interventions are common across both study groups as part of 'Business as Usual' provision, this does not introduce biases nor reduces the security of findings of the study (although it affects the interpretation of results).

Criteria	**Risk level**
Concurrent interventions are explored and there is no evidence suggesting differential uptake of those interventions; **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.	Moderate
Concurrent interventions are explored and there is evidence of large differential uptake between groups.	High
No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.	n/a

3. Experimental effects and contamination

- Is there evidence that the control group behaved differently because of their inclusion in the study? Please consider compensatory rivalry (seeking out and participating in similar programmes) and resentful demoralisation (spend less time in similar activities).
- Is this behaviour likely to affect their outcomes positively or negatively?
- Are sensitivity analyses to account for these behaviours included? Are the results comparable to those of the main analysis?

Experimental effects in the control group	**Contamination**	**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 changes	Explored – minor changes (e.g., 20% of the control units implement something similar)[2]	Similar findings as main analysis	Moderate *Risk level is moderate as soon as one of these conditions is met (OR logic)*.*
Explored – meaningful differences	Explored – meaningful differences (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).*
No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.			n/a

4. Implementation fidelity and compliance with intervention

This criterion is concerned with how well defined and implemented the intervention was during the trial.

[2] 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 and similarity of the activities undertaken by the comparison group.

- Was the intervention appropriately described including references to its critical components and methods of delivery?
- 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 levels of compliance (e.g. young person, family, school etc.) clearly specified?
- Was the intervention content and process delivered as intended (including implementation fidelity and compliance)?

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	Low *Risk level is low only if all of these conditions are met (AND logic).*
Yes	Moderate	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).*
No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.		n/a

5. Attrition – adjustments

This criterion builds on criterion 3: attrition. Criterion 3 is mainly related to the loss of statistical sensitivity. This criterion also explores the potential for bias introduced by attrition and allows for adjustments based on differential attrition, the reason for missingness, and any analyses to account for missing data.

- What was the total amount of missing data?
- Was differential attrition present?
- Were observable variables predictive of missingness?
- Are the results of the analyses accounting for missing data similar to the main analysis?
- Are results robust to further sensitivity analyses to account for missing data?

Total amount of missing data	**Logical connection**	**Differential attrition**	**Logical connection**	**Analyses accounting for missing data**	**Risk level**	**Adjustment to MG rating**
Low	AND	No	AND	Similar to complete-cases analyses	Low	+1
Moderate	AND	No	AND	Similar to complete-cases analyses	Moderate	Needs to be considered in the round with other threats to internal validity
-	-	Yes	AND	Analyses accounting for missing data are similar to the complete-case analyses	Moderate	
-	-	-	-	Analyses accounting for missing data have minor deviations to the complete-case analyses	Moderate	
-	-	-	-	Analyses accounting for missing data differ from complete-case analyses	High	
No information was collected as part of the study, or its quality was deemed insufficient to make any judgement.					n/a	

6. Measurement of outcomes

This criterion is concerned with the use of reliable, valid and acceptable outcome tests that are free from ceiling/floor effects and where scorers are blind to allocation.

- Are the outcome tests a valid and reliable measure of the relevant construct for the population of interest?
- Are the outcome tests administered and scored independently, or in ways that minimise differences between treatment groups?
- Are the outcome tests capable of identifying differences across the whole distribution, i.e. are they free from floor/ceiling effects?
- If floor/ceiling effects are found, do the researchers discuss the implications of the problem and run sensitivity analyses that consider this?

Criteria	**Risk level**
Outcome tests have been thoroughly justified in relation to reliability, validity, utility and acceptably with target population; **AND** Tests are administered and scored blinded to allocation or with very minor judgments; **AND** no ceiling/floor effects are found.	Low
Tests involve minor judgement from assessors who are not blinded to allocation, but safeguards are included to ensure quality; **OR** minor ceiling/floor effects are found and controlled for analytically.	Moderate
Outcome tests have poor validity or reliability for the target population **OR,** Tests involve important judgement from assessors who are not blinded to allocation with no safeguards in place to guarantee independence; **OR** Large ceiling/floor effects are found.	High

7. Selective reporting and data availability

YEF consider selective reporting for those cases where results are presented only for i) a particular outcome measure; ii) a specific analytical approach; or, iii) a subset of participants; contravening what is specified in the Protocol and SAP. YEF ask evaluators to follow what is set out in these prospective documents and the peer review of reports compares the outputs produced by the author of the report against the pre-specified analyses. Thus, instances of selective reporting should be minimal across YEF-funded studies

Additionally, all YEF-funded studies will be expected to submit all data and analysis syntax to YEF's data contractor for the Data Archive. To identify potential errors and minimise deviations on the estimates of impact, results will be re-analysed.

- Is the study registered?
- Are analyses pre-specified and conducted according to plan?
- Was data submitted to YEF' Data Archive and subject to re-analysis?

Criteria	**Risk level**
Study is registered **AND** a comprehensive prospective document is published and followed.	Low
Study is registered **AND** a comprehensive prospective document is published, but with minor deviations.	Moderate
Study is not registered **OR** important deviations from the proposed analysis occur.	High

Annex: Worked examples

Example 1

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 project 1: School-based mentoring
Name of reviewer	John Smith
Date assessment submitted	20/03/25
What is/are the primary outcome(s) of the evaluation?	SDQ externalising behaviour

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*	**Attrition**	**Initial score**	→	**Adjustments**	→	**Final score**
5 🔍	Randomised design	Offending: <=0.1 SDQ tot: <= 0.3 Other: <= 0.2 **MDES 0.18**	0-10%	**4**		*Adjustment for threats to internal validity* **+1**		**5**
4 🔍	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	Offending: 0.11 – 0.19 SDQ tot: 0.31 – 0.39 Other: 0.21 – 0.29	11-20% **15%**					
3 🔍	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables	Offending: 0.2 – 0.29 SDQ tot: 0.4 – 0.49 Other: 0.3 – 0.39	21-30%					

	descriptive of the selection mechanism)							
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	31-40%					
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	41-50%					
0	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6	>50%					

*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**	Randomisation procedure was appropriate, conducted independently and disclosed in the report. There was a very	n/a

			small imbalance in pre-test in favour of the intervention group (0.03) which was controlled for in the model.	
2	Concurrent interventions	**Low**	The IPE suggests that other interventions were implemented in both groups, but the level of support given was similar across trial arms.	n/a
3	Experimental effects and contamination	**Low**	The IPE suggests that there were no important instances of compensatory rivalry or resentful demoralisation.	n/a
4	Implementation fidelity and compliance	**Low**	This study is an effectiveness trial and the IPE suggest that implementation fidelity was high, with a large proportion of teachers delivering a large number of sessions with small adaptations. When non-compliers were excluded from the analysis, the effect size found was similar to the headline figure.	n/a
5	Attrition adjustments	**Low**	The proportion of missing data was low (4%). Reasons for missing data were detailed, and authors showed that those who dropped out did not differ from those who remained in the trial. Authors also showed that the equivalence of treatment and control group on observable variables and demographics was maintained after drop out.	unknown
6	Measurement of outcomes	**Low**	SDQ was used and was deemed appropriate by all stakeholders.	n/a

7	Selective reporting and data availability	**Low**	Trial was registered and primary and secondary outcome analyses were pre-specified. Exploratory analyses are clearly labelled.	n/a

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**
Missing data due to attrition is classified as 'low risk of bias'.	+1
Up to two threats are classified as 'moderate risk' and the direction of the likely biases is unknown or operates in opposite directions.	No adjustment made
• Up to four threats are classified as 'moderate risk' but the directions of biases are unknown; OR • Up to two threats are classified as 'moderate risk' with the same likely direction of bias; OR • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1
• One threat is classified as 'high risk' and two threats are classified as 'moderate risk'; OR • Two or more threats are classified as 'high risk'	-2

Example 2

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 2
Name of reviewer	Sarah Smith
Date assessment submitted	20/03/25
What is/are the primary outcome(s) of the evaluation?	SDQ total score

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*	**Attrition**	**Initial score**	➡	**Adjustments**	➡	**Final score**
5 🔍	Randomised design	Offending: <=0.1 SDQ tot: <= 0.3 Other: <= 0.2 **MDES 0.23**	0-10%	**4**		Adjustment for threats to internal validity **–2**		**2**
4 🔍	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	Offending: 0.11 – 0.19 SDQ tot: 0.31 – 0.39 Other: 0.21 – 0.29	11-20% **17%**					
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	21-30%					
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	31-40%					
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	41-50%					
0 🔍	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7	>50%					

		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	Moderate	Randomisation was appropriate and conducted by an independent statistician. Imbalance was moderate in the pre-test (0.08 SD), but it was controlled for in the regression model. All other characteristics were fairly balanced between the groups with the exception of the % of FSM pupils which was higher in the intervention group. An additional sensitivity analysis controlling for this difference found similar results.	Unknown – higher % of FSM pupils might make it easier or harder to find an effect in the intervention group
2	Concurrent interventions	Low	IPE suggests that most schools had SEL practices in place. However, the magnitude and type of programmes chosen across the two groups was comparable.	Similar in treatment and control group – underestimate impact estimate
3	Experimental effects and contamination	**High**	IPE suggest that control schools took up other SEL programmes and the amount of time spent in the provision of these activities was very similar across both groups suggesting potential compensatory rivalry. For example, there was an increase in the use of SEAL or a nurture group. Randomisation was undertaken at the school level minimising the risks of contamination. **This is likely to underestimate the impact estimate.**	Underestimate impact estimate
4	Implementation fidelity and compliance	Moderate	Implementation fidelity was moderate as adaptations to the model were common, but relatively minor (e.g. changing the order in which activities were done). However, most teachers	n/a

			delivered the number of sessions expected and analysis accounting for non-compliers produced similar results.	
5	Attrition adjustments	Moderate	Missing data was moderately high, at 17%. Data was not differentially missing between treatment groups, but it was associated with weaker previous attainment. However, analysis accounting for missing data remained robust with very similar point estimates and confidence intervals.	n/a
6	Measurement of outcomes	Low	The outcome test is a valid and reliable commercial test that was administered independently and blinded to allocation.	n/a
7	Selective reporting and data availability	Low	This trial was registered and all analyses were conducted as specified in the Protocol and SAP.	n/a

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***
Missing data due to attrition is classified as 'low risk of bias'.	*+1*
Up to two threats are classified as 'moderate risk' and the direction of the likely biases is unknown or operates in opposite directions.	*No adjustment made*
• *Up to four threats are classified as 'moderate risk' but the directions of biases are unknown; OR* • *Up to two threats are classified as 'moderate risk' with the same likely direction of bias; OR* • *Up to one threat is classified as 'high risk' with all other deemed as 'low risk'*	*-1*
• *One threat is classified as 'high risk' and two threats are classified as 'moderate risk'; OR* • *Two or more threats are classified as 'high risk'*	*-2*

Example 3

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 3
Name of reviewer	Rose Tyler
Date assessment submitted	20/03/25
What is/are the primary outcome(s) of the evaluation?	SDQ externalising behaviour

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*	**Attrition**	**Initial score**	➡	**Adjustments**	➡	**Final score**
5 🔍	Randomised design	Offending: <=0.1 SDQ tot: <= 0.3 Other: <= 0.2	0-10% **3% Attrition**	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: 0.11 – 0.19 SDQ tot: 0.31 – 0.39 Other: 0.21 – 0.29 **MDES 0.26**	11-20%					

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	21-30%			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	31-40%					
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	41-50%					
0	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6	>50%					

**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). Evidence supportive of parallel trends before intervention is provided and improved by the additional matching of schools.	n/a
2	Concurrent interventions	No information	No information of concurrent interventions was available in the comparison schools.	n/a
3	Experimental effects and contamination	Low	As schools in the intervention group were identified using administrative data, there is no expectation of potential experimental effects in the comparison group.	n/a

4	Implementation fidelity and compliance	Moderate	Fidelity with the intervention was moderate as some teachers did not attend all training sessions, but they sessions were largely delivered as designed with some minor practical adaptations.	n/a
5	Attrition adjustments	Low	Missing data was remarkably low (3%) so the complete case analysis is expected to be unbiased.	n/a
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.	n/a
7	Selective reporting and data availability	Low	This study was registered and the analytical approach was identified before outcomes were observed.	n/a

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**
Missing data due to attrition is classified as 'low risk of bias'.	+1
Up to two threats are classified as 'moderate risk' and the direction of the likely biases is unknown or operates in opposite directions.	No adjustment made
• Up to four threats are classified as 'moderate risk' but the directions of biases are unknown; OR • Up to two threats are classified as 'moderate risk' with the same likely direction of bias; OR • Up to one threat is classified as 'high risk' with all other deemed as 'low risk'	-1
• One threat is classified as 'high risk' and two threats are classified as 'moderate risk'; OR • Two or more threats are classified as 'high risk'	-2

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

This document was last updated in **March 2025.**

We reserve the right to modify the document at any time, without prior notice.

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413