# Classification of the security of findings from EEF evaluations

Version 2.0 – July 2019

Education Endowment Foundation

## Table of Contents

## A. PURPOSE

This document presents a classification system and accompanying procedure for judging the security of findings from EEF evaluations. The rating also makes recommendations around the design and analysis of evaluations that should be considered for these studies.

The ratings have been designed specifically to differentiate between EEF evaluations, most of which are set up as randomised controlled trials (RCT). The ratings are from 5 padlocks ( ), the most robust evidence that could be expected from a single study, to 0 padlocks ( ) which denotes a study that adds little to the evidence base. The ratings take no account of whether the intervention itself was successful in achieving an impact. The reported average effect size explains the likely scale of the differences in outcomes (translated into month's progress in the case of academic attainment measures) and is reported next to the rating in the executive summary of each report (see Table 1).

**Table1: Example of an impact table in the executive summary of an EEF report**

| OUTCOME | NO. OF PUPILS | EFFECT SIZE (95% CONFIDENCE INTERVAL) | ESTIMATED MONTHS' PROGRESS | SECURITY RATING | COST RATING |
|---|---|---|---|---|---|
| English | 5,376 | 0.033 (0.004, 0.064) | +1 month | | |

The padlock ratings largely refer to the internal rather than external validity of the findings. There needs to be some judgement on the part of the audience as to whether the finding might be generalisable to their context. The system is only to be used to classify the security of findings for EEF evaluations where the primary purpose is to determine impact (i.e. efficacy and effectiveness studies), not where the primary purpose is formative or to establish feasibility, which is the case in pilots[1].

Padlock ratings are assigned to the primary outcome only. For studies that identify more than one primary outcome, separate padlock rating will be assigned for each of these outcomes. Separate padlock ratings could be assigned for studies involving more than one primary outcome.

## B. WHAT IS NEW IN THIS VERSION (VERSION 2.0)?

The padlock rating classification system was originally developed in 2014 and has undergone minor adjustments to date as our thinking develops and methodological improvements are made. This version, as of **July 2019**, introduces 3 refinements to the padlock rating. The changes are:

1. The wording in the "Design Pillar" has been modified to better describe the ways in which non-randomised designs deal with different types of confounding. As most studies commissioned by the EEF to date have used randomised designs, this would not affect the security ratings of most studies. However, it would provide a more accurate description of the robustness of alternative designs.
2. The "Adjustment for Imbalance at Baseline" criterion has been eliminated and is instead included as a threat to validity in RCTs. This change has the potential to *increase* the security rating of studies that found small baseline imbalances in measures of previous attainment, if those were controlled for in the analytical model used.
3. A descriptive table of threats to validity has been introduced to improve the transparency with which these are reported. This version describes threats to validity depending on the design and analytical method chosen and proposes criteria and signalling questions to assess the likely relevance of these threats. These signalling questions have been adapted from the Risk of Bias Tool (Sterne et al, 2016a, 2016b), but we encourage reviewers to also consult those tools. These are designed to improve the consistency and transparency reporting threats to validity but is unlikely to affect the security ratings of most studies as these were considered as part of the Quality Markers and Threats to Internal Validity included in previous versions of the classification system.

[1] http://educationendowmentfoundation.org.uk/uploads/pdf/EEF_evaluation_approach_for_website.pdf

**This version of the padlock rating classification system is used to assess the security of results for reports peer reviewed starting from July 2019**. Previous studies will not be retrospectively rated using this version of the classification system. Given the nature of the changes introduced in this version, security ratings should remain unchanged for most studies. However, an interested reader could use the information in the reports in conjunction with this document to assess the reasons that reduce the security of any given study.

These changes to the classification system are the product of a wide consultation with relevant stakeholders.

First, these were discussed in multiple iterations between Summer 2018 and Summer 2019 with the Evaluation Advisory Group and EEF Staff. Second, these changes were retrospectively piloted with 10 studies to assess the extent to which these would affect the security ratings of previous studies – only small changes were identified, and most adjustments were made upwards in cases were small chance baseline imbalances (that had been controlled for analytically) were penalised dropping padlocks. Other observed differences were related to the *transparency* reporting threats to validity in the padlock classification system. However, these rarely granted changes to the ratings assigned. Third, the new classification system was also piloted with three on-going projects where three peer reviewers were requested to use this version and provide suggestions. These peer reviewers are active members of EEF panel of evaluators, to whom the EEF thanks for their useful and thoughtful contributions. Finally, this classification system was subject to a final round of internal consultation before approving its implementation from July 2019.

## C. Process

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 EEF arbitrates between peer reviewer ratings and presents this to the author,
3. The author has an opportunity to respond,
4. The EEF assigns the final security rating[2].

The security rating is determined by four criteria (see Figure 1). 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 EEF-funded studies.

The classification is of the security of the findings for the primary outcome. Subgroup analyses are not included in the security ratings unless otherwise stated[3].

The four criteria are:

- **Design:** The quality of the design used to create a comparison group of pupils with which to determine an unbiased measure of the impact on attainment. 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 pupil level 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.

[2] On the rare occasions where unsurmountable disagreements were to arise between the peer reviewers, the EEF, and the author, the EEF in consultation with the Evaluation Advisory Group will make the final decision.
[3] In effectiveness trials, the primary analysis might be done for the Free School Meals (FSM) sub-group. In that case, a padlock rating would be assigned to the result from that analysis.

These criteria are combined to generate an overall padlock rating:

| | |
|---|---|
| **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 padlock rating is determined by the lowest of these three ratings. |
| **Step 3**: | The interim padlock rating can be adjusted downwards by assessing **Threats to internal validity**. |
| **Step 4**: | The final padlock rating is determined. |

Appendix 1 provides more detail on how the reviewer should use each criterion in order to determine the final rating for the relevant primary outcome.

Appendix 2 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.

**Figure 1: Security rating template**

**OUTCOME:** *ADD NAME OF THE PRIMARY OUTCOME*

| Rating | Criteria for rating | | | Initial score | | Adjust | | Final score |
|---|---|---|---|---|---|---|---|---|
| | **Design** | **MDES** | **Attrition** | | | Adjustment for threats to internal validity<br>**[X]** | | |
| **5** | Randomised design | <= 0.2 | 0-10% | | | | | |
| **4** | Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs) | 0.21 - 0.29 | 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) | 0.30 - 0.39 | 21-30% | | | | | |
| **2** | Design for comparison that considers selection only on some relevant confounders | 0.40 - 0.49 | 31-40% | | | | | |
| **1** | Design for comparison that does not consider selection on any relevant confounders | 0.50 - 0.59 | 41-50% | | | | | |
| **0** | No comparator | >=0.6 | >50% | | | | | |

| Threats to validity | Threat to internal validity? | Comments |
|---|---|---|
| **Threat 1: Confounding** | Low/Moderate/High | Add relevant comments based on signalling questions and criteria |
| **Threat 2: Concurrent Interventions** | | |
| **Threat 3: Experimental effects** | | |
| **Threat 4: Implementation fidelity** | | |
| **Threat 5: Missing Data** | | |
| **Threat 6: Measurement of Outcomes** | | |
| **Threat 7: Selective reporting** | | |

- **Initial padlock score:** [NUMBER] Padlocks - Insert Description
- **Reason for adjustment for threats to validity**: [- NUMBER] Padlocks - Insert Description
- **Final padlock score:** initial score adjusted for threats to validity = [NUMBER] Padlocks

*Please use this template to assign the security rating for each primary outcome*

## APPENDIX 1 SECURITY CRITERIA

This section describes the four criteria for judging the security rating and provides detailed guidance on each.

### 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. EEF impact evaluations are expected to be designed to attain at least 3 🔒, 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).

| 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 that considers 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) are able to achieve 4 🔒 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 by the use of matching.

Methods that only attempt to control for observable characteristics (for example, matching/weighting) can only achieve 3 🔒 or less.

### 2. 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). Most EEF studies should be powered to detect either of an effect size of 0.2 or whatever is considered a cost-effective impact. For targeted interventions, this is likely to be higher than 0.2.

The EEF's aim is reducing the attainment gap. Consistent with this remit, most EEF studies are designed with a measure of **educational attainment** as the primary outcome. Given the focus on broad attainment outcomes, small effects can be expected. However, for interventions that focus on socio-emotional skills or use narrow outcome measures aligned with the intervention, larger effect sizes may be expected and thus would be appropriate. However, the same the rating system will be used.

| Rating | MDES |
|---|---|
| 5 | <= 0.2 |
| 4 | 0.2- 0.29 |
| 3 | 0.3- 0.39 |
| 2 | 0.4-0.49 |
| 1 | 0.5-0.59 |
| 0 | >=0.6 |

## 3. ATTRITION

Attrition should be measured at the pupil level regardless of the level of randomisation (i.e. pupil 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.

The What Works Clearinghouse (WWC) have developed a system for determining the level of attrition based on the levels of overall and differential attrition (the difference between the attrition between the treatment and control groups).[4] EEF has decided to develop its own attrition scale based purely on overall attrition, rather than relying on that created by the WWC. This is because attrition reduces the sample size and can also introduce biases: a large amount of balanced attrition might not introduce biases, but reduces the sample size and the ability to detect effects of a given size. In contrast, a small amount of very extreme missing data could cause bias when its missingness is not random. One way of judging the potential for bias is to look at the balance on observable baseline characteristics after attrition and explore potential reasons for missing data. However, these analyses cannot inform whether bias has occurred on unobservable characteristics (e.g. motivation). For this reason, the only way to ensure that biases have not been introduced is to be strict about overall attrition.

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

[4] http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf

## 4. THREATS TO INTERNAL VALIDITY

This final criterion allows for adjustment of the security rating based on any other concerns that mean that the findings are less secure. Adjustments of up to two padlocks can be made according to the following table:

| Adjustment to Padlocks | Criteria |
|---|---|
| No adjustment made | Up to two threats are classified as Moderate Risk and the direction of the likely biases is unknown or operates in opposite directions. |
| Drop one padlock | Up to four Threats are classified as Moderate Risk but the directions of biases are unknown; **OR**<br>Up to two Threats are classified as Moderate Risk with the same likely direction of bias; **OR**<br>Up to one Threat is classified as High Risk with all other deemed as Low Risk. |
| Drop two padlocks | One Threat is classified as High Risk and two Threats are classified as Moderate Risk **OR**;<br>Two or more Threats are classified as High Risk. |

Any estimate of intervention impact is subject to threats that could affect the validity of those results. Some of them are common to any design, whilst others are contingent to the analytical method chosen to generate such estimate. The objective of this section is characterising them and providing a basis to assess the validity of impact estimates. We also provide design-specific signalling questions to complement those included in the ROBINS-I Tool (Sterne et al, 2016) and Cochrane risk of bias tool for randomised trials and for cluster randomised trials (Eldridge et al, 2016; Higgins et al, 2016). We encourage evaluators and reviewers to use these signalling questions alongside ROBINS-I.

Adapting the characterisation made by these tools, we can identify seven threats to the internal validity of a study that are relevant for EEF-funded impact evaluations. The first domain pertains to the allocation of individuals to the treatment or comparison group. The other threats can occur after the intervention is implemented and apply to any study design.

Threats before the intervention starts

1. Confounding

Threats after the intervention starts

2. Concurrent interventions
3. Experimental effects and contamination
4. Implementation fidelity
5. Missing data
6. Measurement of outcomes
7. Selective reporting and data availability

Assessing the threats to validity in each of these domains requires peer reviewer's judgment of the likelihood of the problem, and how this could affect the magnitude of the impact estimate. The likelihood of the bias can be qualified as Low, Moderate and High or No Information Available, depending on the descriptions made for each domain. Studies should be designed to provide information on these different threats to validity, but EEF acknowledges that information on some might not be available for studies commissioned before these padlock revisions are published. In those cases where the incomplete or missing information does not allow the reviewer to assess the likelihood of a given threat to validity, this should be clearly stated. Note that not all the domains mentioned above would be equally important for all studies depending on the features upon which these studies were designed.

Using the signalling questions for each domain, the independent peer reviewer is expected to assess the severity and likely direction of the bias and include a brief explanation as this affects the level of security we can

assign to an estimate of impact. These assessments require a degree of expert judgement and knowledge, and thus are flexibly worded to accommodate the views of the reviewer.

## THREATS BEFORE THE INTERVENTION STARTS

To generate an estimate of impact, we would like to observe the outcomes of a pupil when they receive a particular intervention and when they don't. The comparison between the outcomes in both "states of the world" would be the effect of the treatment. However, as we only observe one of those "states of the world", we need to identify a valid counterfactual of what *would have happened* in the absence of the intervention.

Different analytical methods attempt to identify that counterfactual, but differences between the treatment and comparison might continue to exist, biasing the estimates of impact. This section discusses how each methodology tries to account for that potential selection into the treatment.

## 1. CONFOUNDING

A confounder is a variable that i) is correlated with the receipt of an intervention; and, ii) has an independent impact on the outcomes. In an educational setting, one common confounder is a pupil's motivation. For example, those with higher latent motivation are more likely to seek support but are also more likely to obtain high scores. Thus, not considering a pupil's latent motivation can confound it with the effect that an intervention would have on their scores. Importantly, note that variables correlated with the receipt of the intervention (Condition i), but not correlated with the outcome (Condition ii) will not introduce any bias in the estimate of impact.

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 interest in maths. Furthermore, confounding can be based on variables that are observable and measurable, or on variables that are unobservable and unmeasurable.

The potential biases created by confounding can be addressed either by design or using statistical methods. However, the extent to which this issue can be minimised depends on the method (including design and analysis) chosen and can be classified in four broad groups (Waddington et al, 2017):

| | |
|---|---|
| Random assignment | • Randomised Control Trials |
| As-if randomised | • Regression Discontinuity Designs<br>• Valid Instrumental Variables<br>• Natural Experiments |
| Non-randomised, with selection on time-invariant observables and unobservables | • Difference-in-Differences<br>• Repeated Measures Fixed Effects |
| Non-randomised, with selection on observable | • Matching/Weighting<br>• Synthetic Controls<br>• Multivariate Regression |

First, in an RCT, randomisation implies that in expectation intervention and control groups will have the same characteristics. Thus, an RCT is, *in expectation*, freed from time-invariant and time-variant confounders, whether they are observable or unobservable.

Second, some methods are able to generate a comparison group that is "as if-randomised". For example, in a Regression Discontinuity Design, groups at both sides of the relevant arbitrary cut-off within an infinitesimally small window of width *w* are expected to be identical. Thus, this should also break the link between allocation

to the intervention and any other variable with an independent impact on the outcomes; i.e. it attempts to control for observable and unobservable confounders, even if this is valid only around the cut-off.

Third, Difference-in-Differences assumes that in the absence of treatment, the average change in outcomes for those treated would have been equivalent to the change in outcomes for those in the comparison group. This assumption of "parallel trends" attempts to control for observable and unobservable *time-invariant* confounders, but it is not capable of dealing with sources of time-varying confoundedness. Even if these trends are identical before the intervention, the assumption would be violated if individuals select into each group based on characteristics that change between the pre and post treatment periods which also affect the outcomes.

Fourth, a group of designs that employ observable characteristics to identify the comparison group (Matching/Weighting, Multivariate Regressions, Synthetic Controls) rely on the conditional independence assumption, which implies that after taking into account a group of variables, the treatment allocation is independent of other variables affecting the outcomes. Nonetheless, this is only able to consider confoundedness created by observable variables.

In addition, when variables are measured with error, confoundedness can still appear if the measurement error is correlated with the receipt of treatment and the outcomes.

*General Recommendations (design specific recommendations are detailed below):*

1.1 Specify potential confounders for the intervention and their likely effect on outcomes. This could be helpful to assess the potential presence of bias when important confounders have not been considered.

1.2 State what type of confoundedness is controlled by the method chosen, which are the identification assumptions, and the specific tests used to support it.

1.2.1 Provide evidence supporting the identification assumption. More detailed guidance is given below depending on the method used.

1.2.2 Assess whether observable confounders might be measured with errors in a way that is correlated with the intervention and outcomes.

1.2.3 Assure that variables that might be affected by the treatment (mediating variables) are not controlled for in the statistical model used as that is likely to produce biased estimates of impact.

1.3 Assess balance between treatment and intervention groups

1.3.1 Note that these are falsification tests as they cannot rule out imbalance in unobservable characteristics. If imbalances are found, 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?)

1.3.2 Run sensitivity analyses where important confounders are controlled for, especially those for which imbalances are found

1.3.3 Note that some designs (e.g. Difference-in-Differences) do not require balance between study groups to fulfil the relevant identification assumptions.

1.3.4 Acknowledging that small studies are more likely to have imbalance due to chance, consider sample size when assessing balance.

## CONSIDERATIONS DEPENDING ON THE DESIGN:

### RANDOMISED CONTROL TRIALS (RCT):

Type of confoundedness: controls for observable and unobservable confounders.

Key assumption: randomised groups are identical, in expectation.

Recommendations for RCTs:

RCT.1. Randomisation should always be conducted independently by a member of the evaluation team using appropriate methods[5] 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[6].

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[7].

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

Signalling Questions for RCTs:

- How was the allocation sequence conducted, and by whom?
- Is there evidence of any imbalance in relevant characteristics at baseline? Is there any imbalance in pre-tests?
- 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?

Assessment:

| Threat Assessment | Criteria |
|---|---|
| Low risk | Adequate allocation sequence with concealed assignment; **AND**, Imbalance of 0.00 - 0.05 SD in pre-test. |
| Moderate Risk | Imbalance of 0.05 – 0.10 SD in pre-test; **AND**, controlled for in a regression model. |
| High Risk | Inadequate description of allocation sequence; **OR**,<br>Imbalance of 0.05 – 0.10 SD in pre-test AND not controlled for in a regression or that meaningfully affects the estimate of impact; **OR**,<br>Imbalance >0.1 SD in pre-test. |

[5] http://educationendowmentfoundation.org.uk/uploads/pdf/Randomised_trials_in_education_revised.pdf

[6] As suggested by EEF Statistical Analysis Guidance, please express differences as effect sizes, with particular emphasis on the attainment baseline, if applicable, as this is a key predictor of future attainment. Austin (2009) suggest that a standardised difference of 0.1 denotes meaningful imbalance in a baseline covariate. Ho et al (2007) suggest that balance is more relevant in variables that are highly predictive of outcomes. Thus, imbalances larger than 0.1 in variables highly predictive of outcomes (for example, but not limited to, previous attainment) could be considered more problematic. Even if large imbalances in relevant variables occur when using simple randomisation, an analysis adjusting for those variables will be as efficient as other methods that achieve better balance (Hewitt and Togerson, 2006). Thus, according to the aforementioned Statistical Guidance, Evaluators should include sensitivity analysed that control for variables where imbalances are found as this improves the efficiency of the estimate.

[7] This is inherently hard and requires judgment, but information on who conducted the randomisation and how could be helpful. For instance, risks of subversion of the allocation are higher when randomisation is not undertaken by an independent team.

## REGRESSION DISCONTINUITY DESIGNS (RDD):

Type of confoundedness: controls for observable and unobservable confounders locally (in the area around the threshold).

Key assumption:

i) Discontinuity in treatment assignment around the threshold;
ii) Continuity in assigning variable around the threshold. This means that as one approaches the threshold from either side, individuals become more alike.

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[8].

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.

Signalling Questions for RDDs:

- 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?

[8] Note this is only a falsification test as there is still the possibility that individuals at each side of the cut-off control the value of the assignment variable in ways that nets off switches between sides of the cut-off (see Barnighausen et al 2018).

Assessment:

| Assessment | Criteria |
|---|---|
| Low risk | Sharp discontinuity in treatment allocation around the cut-off; **AND,** no evidence of discontinuity in the assignment variable and other covariates around the cut-off; **AND,** appropriate robustness checks are presented with similar results. |
| Moderate Risk | Fuzzy discontinuity in treatment allocation around the cut-off; **OR,** limited evidence of discontinuity (manipulation) in assignment variable or other covariates around the cut-off; **OR**, appropriate robustness checks are presented with some differences in the impacts of estimates. |
| High Risk | No evidence of the discontinuity in treatment allocation around the cut-off; **OR**, evidence suggestive of discontinuity in assignment variable (manipulation) and other covariates around the cut-off; **OR**, large differences in impact estimates when using different robustness checks. |

## DIFFERENCE-IN-DIFFERENCES (DD):

Type of confoundedness: controls for observable and unobservable *time-invariant* confounders

Key assumptions: "Parallel trends" and "Common shocks"

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.

Signalling Questions for DDs:

- 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?

Assessment

| Threat Assessment | Criteria |
|---|---|
| Low risk | Evidence suggestive of parallel trends is presented including in-time and/or in-space placebo tests; **AND,** Matched Diff-in-Diffs is used. |
| Moderate Risk | Evidence suggestive of parallel trends is presented including in-time and/or in-space placebo tests. |
| High Risk | Weak, or no evidence of parallel trends is presented. |

## MATCHING/WEIGHTING:

Type of confoundedness: controls for observable confounders.

Key assumptions: conditional independence based on observable confounders.

Recommendations for Matching/Weighting:

MAT.1. Explain how different variables are expected/hypothesised to be correlated with the treatment status and outcomes[9] (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[10].

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[11].
>
> 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[12] 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[13].

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.

[9] Variables included in the matching/weighting should be correlated with the intervention and the outcomes (confounders), or variables that affect the outcomes, but not influence the exposure can be included as those reduce the variance to be explained without introducing biases. See Austin et al (2007), Brookhart et al (2006) and Caliendo and Kopeinin (2005, p.6) for a discussion and recommendations regarding the choice of variables

[10] See Baser (2006), Caliendo and Kopeinin (2005, p. 11), Stuart (2010) and Austin (2014) for a discussion on the conditions under which some methods are considered superior. To keep this document accessible to a broader audience, those details can be provided in an appendix or as footnotes.

[11] A summary of the methods that can be used to assess balance can be consulted in Austin (2009). If using many-to-one matching, consult Austin (2008) for a summary of diagnostic appropriate for these methods.

[12] For example, Doubly Robust Methods (Funk et al, 2010)

[13] See trimming: Lee, Lessler and Stuart, 2011

Signalling Questions for 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?

Assessment

| Threat Assessment | Criteria |
|---|---|
| Low risk | Includes description of variables to be included in the matching/weighting which are predictive of the intervention and outcomes; **AND** finds good balance in observable characteristics between groups; **AND** explores multiple specifications of the matching/weighting approach finding similar results; **AND** considers further robustness checks. |
| Moderate Risk | Includes satisfactory description of variables to be included in the matching/weighting which are predictive of the intervention and outcomes with minor concerns; **OR**, there are small differences in observable characteristics between groups after matching/weighting, but those are controlled for analytically with alternative methods; **OR** Includes a range of specifications, but these results appear are moderately different depending of the method chosen. |
| High Risk | Includes unsatisfactory description of variables included in the matching/weighting, failing to consider some relevant confounders; **OR**,<br>Large imbalances remain in the observable characteristics between groups after the matching/weighting which are not accounted for. |

### THREATS AFTER THE INTERVENTION STARTS

After an intervention starts, deviations from the expected delivery, problems with data collection or reporting of results can compromise the validity of the impact estimates. These apply to all designs, whether randomisation is used or other assumptions are employed to create a credible counterfactual.

Assessing the threats to validity in each of these domains requires peer reviewer's judgment of the likelihood of the problem, and how this could affect the magnitude of the impact estimate. The likelihood of the bias can be qualified as Low, Moderate and High or No Information Available, depending on the descriptions made for each domain. Studies should be designed to provide information on these different threats to validity, but EEF acknowledges that information on some might not be available for studies commissioned before these padlock revisions are published. In those cases where the incomplete or missing information does not allow the reviewer to assess the likelihood of a given threat to validity, this should be clearly stated.

## 2. CONCURRENT INTERVENTIONS

This problem arises when treatment allocation is correlated with the receipt of another programme that is not part of the intended intervention. For example, schools in Opportunity Areas participating in one of the Teaching and Leadership Innovation Fund (TLIF) projects are likely to also be involved in other related initiatives (e.g. receiving additional funding). Thus, if treatment schools are more likely to be receiving this other intervention (e.g. receiving additional funding), the impact estimate can be capturing the effects of both the specific TLIF-funded project under investigation and the additional funding, thus potentially overestimating the treatment effects of the TLIF-funded project[14].

Under randomisation, these cases should be rare as schools would be balanced in observable and unobservable characteristics, in expectation.

[14] This assumes that both programmes have positive effects, but that is not necessarily the case.

Schools are also likely to be implementing multiple interventions as part of their 'Business as Usual' provision. If concurrent interventions are common across both study groups as part of their 'Business as Usual' provision, this does not introduce biases nor reduces the security of findings of the study. However, this may affect the interpretation of the results as the study may be comparing a given intervention with whatever is being implemented in schools. This may also imply that the effects that could be expected are smaller[15].

For example, in Years 11 and 12, schools are likely to have many interventions to improve maths attainment which would imply that creating an additional, differential change in attainment would be harder. Consequently, the intervention effects should be appraised in comparison to what is done in comparison schools and thus the effect sizes that the researcher can expect would be smaller.

*Recommendations*:

2.1 Use contextual knowledge and elements of the Implementation and Process Evaluation to explore whether other interventions are implemented, and in which study group.

2.2 When concurrent interventions are expected as Business as Usual provision, this would affect the interpretation of the results and could be fed into the sample size calculations.

2.3 When data on concurrent interventions is available and those are expected to be correlated with the intervention and with outcomes, indicators for these interventions could be included in sensitivity analyses.

*Signalling Questions:*

- What entails business as usual in intervention and comparison schools?
- Are schools implementing similar interventions as part of their 'Business as Usual' provision?
- Are there any other interventions or approaches, other than the intervention, more common in one of the groups? If yes, are those accounted for in the analysis?

Assessment:

| Risk Assessment | Criteria |
|---|---|
| Low risk | 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. |
| Moderate Risk | Concurrent interventions are explored and there is evidence of minor differential uptake between groups which is not controlled for analytically. |
| High Risk | Concurrent interventions are explored and there is evidence of large differential uptake between groups. |
| No Information Available | No information was collected as part of the study, or its quality was deemed insufficient to make any judgement. |

## 3. EXPERIMENTAL EFFECTS AND CONTAMINATION

Experimental effects refer to actions taken by study groups upon recognition that they are part of an experiment. Changes in the treatment group are known as Hawthorne Effects and those in the control group are named after John Henry. These can operate in both directions; for example, the control group can gain knowledge of similar programmes and decide to implement them (compensatory rivalry) or might choose to reduce the amount of time allocated to similar activities (resentful demoralisation).

In addition, units in the comparison group might receive the intervention (double-sided non-compliance or "contamination") which would also bias the estimates of impact. For example, when within school

[15] When intense concurrent evaluations are expected as part of 'Business as Usual' provision, Evaluators could consider powering studies to detect smaller MDES compared to without these intense concurrent interventions

randomisation is used, teachers assigned to the treatment might share materials with other teachers assigned to the control.

In the context of educational interventions, it is difficult to blind recipients to their treatment allocation as is common practice in medical sciences. However, assessing the likelihood of experimental effects should remain a key aspect of evaluating the internal validity of the study.

Recommendations

3.1 Collect data on usual practice in study groups before the intervention starts and after it has been implemented. Compare potential changes in practice between study groups and how this might be affected by their involvement in the study.

3.2 Judge how these changes in practice may affect the magnitude and direction of outcomes in each group, and the potential effect for the impact estimate.

3.3 Whenever data on these behaviours are available, include relevant sensitivity analyses, for example, excluding those that

3.4 Using the findings from the Implementation and Process Evaluation, assess whether the comparison group is likely to be affected by the intervention (contamination of the comparison group; double-sided non-compliance)

Signalling Questions

- Is there evidence that the control group behaved differently because of their inclusion in the study?
- 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?

Assessment

| Threat Assessment | Criteria |
|---|---|
| Low risk | Experimental effects are explored and there is no evidence suggesting such behaviours. Contamination of the control group is explored and there is no evidence suggesting such behaviour. |
| Moderate Risk | Experimental effects or contamination are explored and there is evidence of minor differential changes (e.g. 20%[16] of the control units implementing something similar) **AND** sensitivity analyses accounting for these behaviours find effects similar to the main analysis. |
| High Risk | Experimental effects are explored and there is evidence of meaningful differential changes (e.g. 50% of the control units implementing something similar), ***OR*** sensitivity analyses accounting for these behaviours obtain different results than the main analysis. |
| No Information Available | No information was collected as part of the study, or its quality was deemed insufficient to make any judgement. |

[16] 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.

## 4. IMPLEMENTATION FIDELITY AND COMPLIANCE WITH INTERVENTION

Interventions are rarely delivered as the developers envisioned. For example, in an intervention where trainers train teachers to deliver a literacy programme in a classroom setting, these trainers may deviate from the planned content and process of intervention implementation, and teachers might adapt the intervention or not implement certain components of it if they feel the pupils are not responding as expected and are not interested. Furthermore, pupils might not comply with all required activities, among others. Low implementation fidelity (both regarding how training is delivered to teachers and the intervention delivered in classrooms) can be a potential moderator of the relationship between interventions and their intended outcomes which could affect the observed impact estimates in comparison to the impact estimate that could be expected under ideal practice[17]. Participant responsiveness can also be a key factor moderating the degree of implementation fidelity achieved for a given intervention.

It is important to make judgements about the level of implementation fidelity and compliance in both effectiveness and efficacy trials as low levels could affect the interpretation of results. Low implementation fidelity could be a threat to the validity of both efficacy and effectiveness trials when the essential components of an intervention are absent, i.e. a 'null' result with high implementation fidelity suggests that the key components of an intervention fail to create the changes expected while a 'null' result with low implementation fidelity does not necessarily preclude that an approach could work if implemented appropriately.

Recommendations

4.1 Describe the *'implementation logic'* stating its main components and features (e.g. activities that the trainers need to complete and how; activities that the teachers need to complete and how).

4.2 Triangulate the results from the Implementation and Process Evaluation to understand the relationship between the level of fidelity to the 'implementation logic' and how it may affect the relationship between the intervention and desired outcomes. This might also be moderated by other contextual factors which could be explored.

Signalling Questions

- 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. pupil, teacher, school) clearly specified?
- Was the intervention content and process delivered as intended (including implementation fidelity and compliance)?

Assessment

| Threat Assessment | Criteria |
|---|---|
| Low risk | Implementation fidelity and/ or compliance are well defined and aligned with the implementation logic and the causal mechanism identified in the logic model; **AND,** implementation fidelity and/ or compliance with the intervention are high. |
| Moderate Risk | Implementation fidelity and/ or compliance are well defined and aligned with the implementation logic and the causal mechanism identified in the logic model **AND,** implementation fidelity and/ or compliance are moderate. |
| High Risk | Implementation fidelity and/ or compliance are not well defined or poorly aligned with the logic model; **OR,** Implementation fidelity and/ or compliance are very low. |
| No Information Available | No information was collected as part of the study, or its quality was deemed insufficient to make any judgement. |

[17] Including its individual aspects of it such as content, dose, frequency and duration if applicable.

## 5. MISSING DATA

Missing data is pervasive in impact evaluations. It may arise due to attrition, missed data collection points or exclusion of some participants. Attrition can be problematic in two ways: i) It reduces the statistical power to capture an ES of a given magnitude, and ii) It can introduce biases to the impact estimate. Differential attrition poses a greater threat to validity than data missing completely at random because the former would bias the impact estimate (ii) while the latter would lead to an analysis with less statistical sensitivity (i). The Pillar 3 (Attrition) is mainly related to the loss of statistical sensitivity (i), while this threat to validity also explores the potential for bias introduced by missing data.

Analysis can account for some forms of missing data, but as it is impossible to ascertain whether data is missing based on *unobservable* characteristics that can lead to bias, evaluators should strive to minimise missing data as the only way of guaranteeing that no biases are introduced.

*Recommendations:*

5.1 Use data collection processes that minimise missing data (i.e. administrative data of statutory tests like GCSE results).

5.2 Assess the total share of missing data at the unit of treatment and the unit of analysis[18].

5.3 Assess the share of missing data by intervention group, summarising reason for missing data at the unit of treatment and the unit of analysis.

5.4 Define a model to explore the patterns of missingness. This would generally use a logistic model that defines drop out using the variables of the main model and additional variables available.

5.5 Run models that consider missing data[19] clearly stating their assumptions and comparing the results with the complete-cases analysis.

5.6 Consider further sensitivity analyses if the results of the complete case analysis and the preferred missing data approach differ.

Signalling Questions

- What was the total amount of missing data?
- 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?

Risk Assessment

| Risk Assessment | Criteria |
|---|---|
| Low risk | Total missing data is low; **AND**, there is no differential missing data between treatment groups; **AND,** analyses accounting for missing data are similar to the complete-cases analyses. |
| Moderate Risk | Total missing data is moderate; ***AND*****,** analyses accounting for missing data are similar to the complete-cases analyses; **OR**,<br>There is differential missing data between treatment groups, ***AND*** analyses accounting for missing data are similar to the complete-cases analyses; **OR,**<br>Analyses accounting for missing data have minor deviations to the complete-cases analyses. |
| Serious Risk | Analyses accounting for missing data differ from the complete-cases analyses. |
| No Information Available | No information was collected as part of the study, or its quality was deemed insufficient to make any judgement. |

[18] Usually the unit of treatment is a school and the unit of analysis is the pupil

[19] For example, complete cases including covariates predictive of missingness, Multiple imputation, inverse probability weighting, full information maximum likelihood. See EEF's Statistical Analysis Guidance for recommendations for these analyses.

## 6. MEASUREMENT OF OUTCOMES

When choosing outcome tests, evaluators are expected to consider their validity, reliability, relation with future outcomes, educational value, and should strive to minimise the burden on schools and pupils[20].

Schools and assessors could behave in ways that differentially affect the outcomes in different treatment groups. For this reason, it is desirable to choose outcome measures that can be blinded to treatment allocation, are marked independently (e.g. SATs) or involve limited judgement from the assessors. Even if these are preferred, these characteristics are not necessary as might not be adequate in certain settings (e.g. practitioner-led tests in the early years). Including safeguards to guarantee the independence of data collection in those cases could be considered (e.g. by conducting independent checks on the marking or invigilation of a sample of tests).

Some analytical models make assumptions around the distribution of variables. In some occasions, outcome measurement can be subject to flukes that can compromise our ability to use them for evaluation purposes as they are inconsistent with the distributional assumptions made by the analytical models used. For example, some analytical models are not meant to be used when the distributions of results are very skewed or censored. Thus, the distribution of outcome measures should be explored.

Additionally, outcomes might be measured with errors. When those errors are correlated with the intervention or other relevant confounders, they could introduce biases. In contrast, when errors are random, it would reduce the precision of the impact estimate, but it would remain unbiased. Thus, it is relevant to investigate whether outcomes are likely to be measured with errors.

*Recommendations:*

6.1 Evaluators should consider the validity, reliability, breath, predictive capacity and educational value of outcome tests chosen.

6.2 Tests should ideally be administered and marked blinded to treatment allocation. However, this is not possible in many settings. In those cases, additional safeguards to guarantee that tests are independent are recommended.

6.3 Whenever possible and adequate, markers should have limited influence on how scores are assigned or quality assurance methods could be introduced (e.g. double-marking).

6.4 Provide details on the distribution of results using histograms and QQ-plots to explore potential floor or ceiling effects. If those are found, include relevant robustness checks that account for the censored structure of the data (e.g. Tobit models).

6.5 Assess whether measurement error might be related to the intervention or other confounders (differential measurement error).

Signalling questions

- Are the outcome tests a valid and reliable measure of the relevant construct for the population of interest?
- Are the outcome tests administered and marked 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?

[20] Please consult EEF's Guidance on choosing attainment tests for further guidance

Assessment

| Threat Assessment | Criteria |
|---|---|
| Low risk | Outcome tests have been thoroughly justified in relation to reliability, validity, utility and acceptably with target population; **AND**, Tests are administered and marked blinded to allocation or with very minor judgments; **AND**, no ceiling/floor effects are found. |
| Moderate Risk | 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. |
| High Risk | Outcome tests have poor validity or reliability for the target population **OR,**<br>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. |

## 7. SELECTIVE REPORTING AND DATA AVAILABILITY

EEF 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. EEF 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 EEF-funded studies

Additionally, all EEF-funded studies will be expected to submit all data and analysis syntax to EEF's data contractor for the Data Archive. To identify potential errors and minimise deviations on the estimates of impact, results will be re-analysed by a team at University of Durham led by Prof. Steve Higgins.

*Recommendations:*

7.1 Write prospective protocol and analysis plans based on templates provided by EEF and according to CONSORT checklists[21] and other templates for transparent reporting.
- 7.1.1 Describe the intervention to enable replication, with core components outlined as well as any aspects that are adaptable.
- 7.1.2 Pre-specify primary and secondary outcome measures, including any planned transformation of the data.
- 7.1.3 Pre-specify analytical methods used to generate the estimates of impact.
- 7.1.4 Pre-specify all sub-group analyses.

7.2 Register studies in the relevant indexing platforms like http://www.controlled-trials.com/isrctn/submission/

7.3 Flag analyses as exploratory if not pre-specified and described in the relevant prospective document.

7.4 Flag analyses as exploratory if multiple testing is not accounted for.

7.5 Submit all data to EEF's Data Archive partner for re-analysis of results to assess whether results are replicable by a third party (University of Durham).

Signalling questions

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

[21] http://educationendowmentfoundation.org.uk/uploads/pdf/Evaluation_protocol_templatePDF.pdf

Risk Assessment

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

## APPENDIX 2 WORKED EXAMPLES

| Rating | Criteria for rating | | | Initial score | | Adjust | | Final score |
|---|---|---|---|---|---|---|---|---|
| | **Design** | **MDES** | **Attrition** | | | | | |
| **5** | Randomised design | <= 0.2 | 0-10% | **5** | | | | 5 |
| **4** | Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs) | 0.21 - 0.29 | 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) | 0.30 - 0.39 | 21-30% | | | Adjustment for threats to internal validity **[0]** | | |
| **2** | Design for comparison that considers selection only on some relevant confounders | 0.40 - 0.49 | 31-40% | | | | | |
| **1** | Design for comparison that does not consider selection on any relevant confounders | 0.50 - 0.59 | 41-50% | | | | | |
| **0** | No comparator | >=0.6 | >50% | | | | | |

| Threats to validity | Threat to internal validity? | Comments |
|---|---|---|
| **Threat 1: Confounding** | Low | Randomisation procedure was appropriate, conducted independently and disclosed in the report. There was a very small imbalance in pre-test in favour of the intervention group (0.03) which was controlled for in the model. |
| **Threat 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. |
| **Threat 3: Experimental effects and contamination** | Low | The IPE suggests that there were no important instances of compensatory rivalry or resentful demoralisation. |
| **Threat 4: Implementation fidelity** | 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. |
| **Threat 5: Missing Data** | Moderate | The proportion of missing data was moderate (9%). Reasons for missing data were detailed and explored including a detailed drop-out model. Multiple imputation was conducted with qualitatively similar results. Further sensitivity analysis was conducted to explore the impact of data missing not at random. |
| **Threat 6: Measurement of Outcomes** | Low | Outcome testing was conducted and marked independently and blinded to treatment allocation. This test has been standardised for the UK population and was deemed appropriate by all stakeholders. |
| **Threat 7: Selective reporting** | Low | Trial was registered and all analyses were pre-specified. |

- **Initial padlock score:** [5] Padlocks – This was a well conducted cluster randomised trial with MDES at randomisation of 0.18 and 9% attrition
- **Reason for adjustment for threats to validity**: [- 0] Padlocks – There is no evidence of relevant threats to validity
- **Final padlock score:** initial score adjusted for threats to validity = [5] Padlocks

| Rating | Criteria for rating | | | Initial score | | Adjust | | Final score |
|---|---|---|---|---|---|---|---|---|
| | Design | MDES | Attrition | | | | | |
| 5 | Randomised design | <= 0.2 | 0-10% | | | | | |
| 4 | Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs) | 0.21 - 0.29 | 11-20% | 4 | | | | |
| 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) | 0.30 - 0.39 | 21-30% | | | Adjustment for threats to internal validity **[ -1 ]** | | 3 |
| 2 | Design for comparison that considers selection only on some relevant confounders | 0.40 - 0.49 | 31-40% | | | | | |
| 1 | Design for comparison that does not consider selection on any relevant confounders | 0.50 - 0.59 | 41-50% | | | | | |
| 0 | No comparator | >=0.6 | >50% | | | | | |

| Threats to validity | Threat to internal validity? | Comments |
|---|---|---|
| **Bias 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. |
| **Bias 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. |
| **Bias 3: Experimental effects** | **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.** |
| **Threat 4: Implementation fidelity** | 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 delivered the number of sessions expected and analysis accounting for non-compliers produced similar results. |
| **Bias 5: Missing Data** | 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. |
| **Bias 6: Measurement of Outcomes** | Low | The outcome test is a valid and reliable commercial test that was administered independently and blinded to allocation. |
| **Bias 7: Selective reporting** | Low | This trial was registered and all analyses were conducted as specified in the Protocol and SAP. |

- **Initial padlock score:** [4] Padlocks – This was a well conducted cluster randomised trial with MDES at randomisation of 0.23 and 17% attrition
- **Reason for adjustment for threats to validity**: [- 1] Padlocks – There is evidence consistent with compensatory rivalry which suggests that the impact estimate could be underestimated
- **Final padlock score:** initial score adjusted for threats to validity = [3] Padlocks

| Rating | Criteria for rating | | | Initial score | | Adjust | | Final score |
|---|---|---|---|---|---|---|---|---|
| | Design | MDES | Attrition | | | | | |
| 5 | Randomised design | <= 0.2 | 0-10% | | | | | |
| 4 | Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs) | 0.21 - 0.29 | 11-20% | 4 | | | | 4 |
| 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) | 0.30 - 0.39 | 21-30% | | | Adjustment for threats to internal validity **[0]** | | |
| 2 | Design for comparison that considers selection only on some relevant confounders | 0.40 - 0.49 | 31-40% | | | | | |
| 1 | Design for comparison that does not consider selection on any relevant confounders | 0.50 - 0.59 | 41-50% | | | | | |
| 0 | No comparator | >=0.6 | >50% | | | | | |

| Threats to validity | Threat to internal validity? | Comments |
|---|---|---|
| **Bias 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. |
| **Bias 2: Concurrent Interventions** | No information | No information of concurrent interventions was available in the comparison schools. |
| **Bias 3: Experimental effects** | Low | As schools in the intervention group were identified using administrative data, there is no expectation of potential experimental effects in the comparison group. |
| **Threat 4: Implementation fidelity** | 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. |
| **Bias 5: Missing Data** | Low | Missing data was remarkably low (3%) so the complete case analysis is expected to be unbiased. |
| **Bias 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. |
| **Bias 7: Selective reporting** | Low | This study was registered and the analytical approach was identified before outcomes were observed. |

- **Initial padlock score:** [4] Padlocks – This was a well-conducted matched difference-in-differences with MDES of 0.26 and 3% attrition
- **Reason for adjustment for threats to validity**: [-0] Padlocks – There is no evidence of serious threats to the internal validity of the study
- **Final padlock score:** initial score adjusted for threats to validity = [4] Padlocks

## REFERENCES


Abadie, Alberto and Javier Gardeazabal (2003). The economic costs of conflict. A case study of the Basque Country. *The American Economic Review 93(1)*: 113-132.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association 105(490)*: 493-505.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59(2)*: 495-510.

Arpino, Bruno and Massimo Cannas (2015). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine 35*: 2074-2091.

Austin, Peter (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity score. *Pharmacoepidemiology and drug safety 17*: 1270-1225.

Austin, Peter (2009). Balance Diagnostic for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples. Statistics in Medicine 28: 3083-3107.

Austin, Peter (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46(3)*: 399-424.

Austin, Peter (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics 10*: 150-161.

Austin, Peter (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine 33*: 1057-1069.

Austin, Peter; P. Grootendorst and G. Anderson (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine 26(4)*:734-753.

Baser, Onur (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health 9(6)*: 2006.

Brookhart, Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn and Til Sturmer (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology 163(12)*.

Caliendo, Marco and Sabine Kopeinig (2005). Some practical guidance for the implementation of propensity score matching. *DIW Discussion Papers No. 485.*

Eldridge, Sandra Marion Campbell, Michael Campbell, Amy Dahota, Bruno Giraudeau, Julian Higgins, Barney Reeves and Nandi Siegfried (2016). Revised Cochrane Risk of Bias Tool for Randomized Trials (RoB 2.0). Additional considerations for cluster-randomised trials.

Ferman, Bruno, Cristine Pinto and Vitor Possebom (2016). Cherry Picking with Synthetic Controls. *Sao Paulo School of Economics Working Paper 420.*

Funk, Michele; Daniel Westreich, Chris Wiesen, Til Sturmer, Alan Brookhart and Marie Davidian (2009). Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology 173 (7):* 761-767.

Higgins, Julian; Jelena Savović, Matthew J Page, Jonathan AC Sterne (2016). Revised Cochrane Risk of Bias Tool for Randomized Trials (RoB 2.0).

Hewitt, Catherine and David Togerson (2006). Is restricted randomisation necessary? *British Medical Journal Research Methods 332:* 1506-1508.

Ho, Daniel; Kosuke Imai; Gary King and Elizabeth Stuart (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis 15*: 199-236.

King, Gary; C. Lucas; and, R. Nielsen (2017). The Balance-Sample Size Frontier in Matching Methods for Causal Inference. *American Journal of Political Science 61(2)*:473-489.

Lee, Brian, Justin Lessler and Elizabeth Stuart (2011). Weight Trimming and Propensity Score Weighting. *Plos ONE 6(3)*: e18174.

Lee, David and Thomas Lemieux (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature 48*: 281-355.

Li, Fan, Alan Zaslavsky and Mary Beth Landrum (2012). Propensity Score Weighting with Multilevel Data. *Statistics in Medicine 32*: 3373-3387.

Linde, Ariel and Paul Yarnold (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice 22(6)*: 868-874.

McClelland, Robert and Sarah Gault (2017). The synthetic control method as a tool to understand state policy. *The Urban Institute research report.*

McCrary, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142(2)*: 698-714.

Steiner, Peter, Jee-Seon Kim and Felix Thoemmes (2012). Matching Strategies for Observational Multilevel Data. *Social Statistics Section JSM 2012*.

Sterne, Jonathan; Miguel A Hernán; Barnaby C Reeves; Jelena Savović; Nancy D Berkman; Meera Viswanathan; David Henry; Douglas G Altman; Mohammed T Ansari: Isabelle Boutron: James R Carpenter; An-Wen Chan: Rachel Churchill; Jonathan J Deeks; Asbjørn Hróbjartsson; Jamie Kirkham; Peter Jüni, ; Yoon K Loke; Theresa D Pigott; Craig R Ramsay; Deborah Regidor; Hannah R Rothstein; Lakhbir Sandhu; Pasqualina L Santaguida; Holger J Schünemann; Beverly Shea; Ian Shrier; Peter Tugwell; Lucy Turner; Jeffrey C Valentine: Hugh Waddington; Elizabeth Waters; George A Wells; Penny F Whiting; Julian PT Higgins (2017). ROBINS-I: a tool for assessing risk of bias in non-randomised studies if interventions. *British Medical Journal 335*: i4919.

Stuart, Elizabeth (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci 25(1)*:1-21.

Waddington, Hugh; Ariel Aloe; Betsy Jane Becker; Eric Djimeu; Jorge Garcia Hombrados; Peter Tugwell; Geoge Wells; Barney Reeves (2017). Quasi-experimental study designs series – paper 6: risk of bias assessment. *Journal of Clinical Epidemiology 89*: 43-52.

Williamson, Elizabeth, Ruth Morley, Alan Lucas and James Carpenter (2011). Propensity Scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research 21(3)*: 273-293.

Zubizarreta, Jose and Luke Keele (2017). Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System. *Journal of the American Statistical Association 112*: 547-560.

## ACKNOWLEDGEMENTS

The EEF would like to thank Professor Stephen Gorard and Professor Steve Higgins for their original thinking on the classification system. EEF would also like to thank all members of the EEF Panel of Evaluators and Evaluation Advisory Group who have contributed to their expertise and advice during the subsequent periods of consultation.