

# Learning Impact Fund

## Classification of the security of findings from E4L evaluations

Last Updated: 10 April 2018

### 1. Background

This document presents a classification system and accompanying procedure for judging the security of findings from Evidence for Learning (E4L) evaluations. Where appropriate, examples and supporting information are drawn materials developed by the UK's Education Endowment Foundation (EEF) (E4L partner). This document provides:

- Purpose of the E4L security ratings system
- Process and overview of the E4L security ratings system
- Quality markers for evaluating a trial (Appendix A)
- Criteria for judging security ratings
- Examples of security ratings reports

### 2. Purpose of the E4L Security Ratings System

The primary purpose of this system is to communicate the likelihood of finding the same result in a similar context again in another well-conducted evaluation. The ratings have been designed specifically to differentiate between E4L evaluations, most of which are set up as randomised controlled trials (RCT).

The ratings are from 5 padlocks (🔒🔒🔒🔒🔒), the best kind of evidence that could be expected from a single study, to 0 padlocks which denotes a study that adds little or nothing to the evidence base. The ratings take no account of whether the intervention itself was successful. The reported average effect size (translated into month's progress) explains the likely scale of the difference in attainment (or other outcome measure) and is reported next to the rating in the executive summary of each report (see Table 1).

**Table 1: Example of an impact table in the executive summary of an E4L report**

Outcome	No. of students	Effect size (95% confidence interval)	Estimated months progress	Security Rating	Cost Rating
English	1,326	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. However, a trial cannot achieve the top rating without having some external validity, defined as the trial having been conducted in a large number of schools or at least two different contexts (e.g. two local authorities).

The system is only to be used to classify the security of findings for E4L evaluations where the primary purpose is to determine impact (efficacy and effectiveness trials), not where the primary purpose is formative or to establish feasibility, which is the case in pilots.

### 3. Process









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

1. Two peer reviewers will use this guidance to provide a security rating
2. E4L arbitrates between peer reviewer ratings and presents this to the author
3. The author has an opportunity to comment
4. E4L make the final decision based on recommendations from reviewers

The security rating is determined by five criteria (see Figure 1). These are not the only things that are important in determining the quality of any evaluation. They are, however, the key factors that E4L provides on best practice in evaluation that it expects evaluators to fulfil as standard. These best practice standards are described in detail in the 'Quality markers' in Appendix 1. These Quality markers are not included in the classification criteria because they should be standard practise in all E4L evaluations.

#### Figure 1: Security rating template

1. **Criteria for rating:** In each column highlight the relevant cell in green
2. **Initial score:** Write how many padlocks the trial has received based on the first 3 columns ("x 🗝") and highlight in green (initial score is the lowest rating from the first three columns – see guidance on security classification for more detail)
3. **Adjust:** record adjustment for balance and threats for validity in the adjust column.
4. **Final score:** Write the number of padlocks ("x 🗝") in the relevant cell and highlight green
5. Provide a brief summary of your classification, following the bullet point prompts below.

Rating	Criteria for rating			Initial Score	Adjust	Final Score
	<b>Design</b>	<b>Power (MDES*)</b>	<b>Attrition<sup>1</sup></b>			
5 	Well conducted experimental design with appropriate analysis	MDES <0.2	0-10%		Adjustment for Balance [ ]	
4 	Fair and clear quasi-experimental (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES <0.3	11-20%			
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES <0.4	21-30%		Adjustment for threats to internal validity [ ]	
2 	Weakly matched comparison or experimental design with major flaws	MDES <0.5	31-40%			
1 	Comparison group with poor or no matching (e.g. volunteer versus others)	MDES <0.6	41-50%			
0 	No comparator	MDES <0.6	Over 50%			

- Initial padlock score: lowest of the three ratings for design, power and attrition = [x] padlocks
- Reason for adjustment for balance (if made):
- Reason for adjustment for threats to validity (if made):
- Final padlock score: initial score adjusted for balance and internal validity = [y] padlocks

<sup>1</sup> Attrition should be measured at the student level (even for clustered trials) and from the point of randomisation to the point of analysis.

The classification is of the security of the findings is for the main analysis. Subgroup analyses, such as students receiving financial support (e.g. School Cards scheme in South Australia), are not included in the security ratings unless otherwise stated.

The five 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.
- **Power:** 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, and potential for causing bias.
- **Balance:** The amount of balance achieved on pupil attainment at baseline in the pupils analysed.
- **Threats to internal validity:** How well-defined and consistently delivered the intervention was, and whether the findings could be explained by anything other than the intervention.

These criteria are combined to generate an overall padlock rating:

<b>Step 1:</b>	The first three criteria – Design, Power and Attrition – are awarded
<b>Step 2:</b>	An interim padlock rating is determined by the lowest of these three ratings.
<b>Step 3:</b>	The interim padlock rating can be adjusted downwards by assessing Balance and Threats to internal validity
<b>Step 4:</b>	The final padlock rating is determined.

A template for peer reviewers is included in the peer reviewer guidance on trials (pilots do not need security ratings) <sup>2</sup>.

Appendix 3 provides more detail on how the reviewer should use each criterion in order to determine the final rating for an evaluation.

Appendix 4 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. In addition, we have developed a list of ‘quality markers’ that are described in the Appendix 1 that the reviewer should check have been fulfilled.



**EVIDENCE  
FOR LEARNING**



info@evidenceforlearning.org.au | evidenceforlearning.org.au | @E4Ltweets

# Appendices

## Appendix 1: Quality Markers

This section summarises a set of ‘Quality markers’. The quality markers are split into best practice categories and references.

### Trial set-up

1. All E4L trials are expected to be registered at: <http://www.anzctr.org.au/>
2. Evaluators are expected to provide an Evaluation Protocol using a template based on CONSORT, agreed with the developer, which is published on E4L’s website, including pre-specified primary and secondary outcome measures and analyses.

### Outcome measures

3. All outcome measures in E4L trials should be reliable, externally valid, and predictive of future performance in national tests and non-intervention specific.
4. Outcome testing should be conducted independently and marked blind to treatment. This is an important factor that is also included in the ‘Threats to validity’ criterion for judging security.

### Randomisation

5. Randomisation is always expected to be conducted independently by a statistician within the independent evaluation team.
6. Guidance on appropriate randomisation methods can be found on the EEF’s website in the Randomised Trials in Education Handbook.<sup>2</sup>
7. The randomisation method is expected to be described fully in the report.

### Analysis and reporting

8. All E4L trials are expected to be reported to CONSORT standards and E4L provides a template based on CONSORT for this purpose.
9. E4L also provides best practice guidance to evaluators on analysis and Statistical Analysis Plans (SAPS)

### Publicly available data and long-term follow-up

10. Where possible a long-term control group should be built into the design.

When assessing the security of findings of an evaluation, the reviewer should check through the list above to ensure that these quality markers have also been fulfilled. If there are any concerns these should be noted. In the worst-case scenario this may result in a down-grading of the security rating (for example in the case that a project has managed to use a measure that is inherent to treatment as the primary outcome measure).

---


<sup>2</sup> [https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Randomised\\_trials\\_in\\_education\\_revised.pdf](https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Randomised_trials_in_education_revised.pdf)

## Appendix 2: Security criteria

This section describes the five criteria for judging the security rating of an E4L evaluation and provides detailed guidance on each.







### 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. All E4L impact evaluations should have a design of 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, evidence that valid methods were used to identify the control group (reports of unbiased randomisation, appropriate methods to reduce chance imbalance, appropriate and successful matching) and (3) the level of baseline imbalance in key covariates, especially the pre-test.




**Table 1: Thresholds for rating quality of design**



Rating	Design
5 	Well conducted experimental design with appropriate analysis
4 	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity
2 	Weakly matched comparison or experimental design with major flaws
1 	Comparison group with poor or no matching (E.g. volunteer versus others)
0 	No comparator

### **Notes:**

The randomisation method is expected to be described fully in the report. The report should also describe the level of balance achieved from the randomisation at baseline.


Regression discontinuity design (RDD) is able to achieve 4 , because it attempts to balance unobservable as well as observable characteristics, which is the advantage it has over matched designs. However, the effectiveness of RDD is highly dependent upon assumptions such as the

‘sharpness’ of the cut point. Only a ‘sharp’ RDD would be able to achieve 4 , whereas a ‘fuzzy’ RDD might achieve only 3  or 2 . Some judgement will be required here. (So far the EEF has only funded one RDD<sup>11</sup> so this is not a common design).







Quasi-experimental designs can only achieve 3  because they cannot seek to balance unobservable characteristics. Other quasi-experimental designs that would be able to achieve 3  include regression, instrumental variables and difference in difference approaches.

## 2. Power

This is the ability of the study to detect the impact on attainment. Power 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. In the case of the EEF, it is advised that evaluators always control for prior attainment using National Pupil Database (NPD) data in order to increase the precision of estimates.

The rating on this criterion should be determined by the actual power of the study at the start of the trial (i.e. at randomisation for an RCT) calculated using the actual ICC and pre-post test correlation. One option would be to use Optimal Design software to do this.<sup>13</sup> Most E4L trials should be powered to detect either of an effect of 0.2 or whatever is considered to be a cost-effective impact. In addition, in order to achieve 5 , as well as being powered to 0.2, the trial needs to have been delivered across at least two contexts.


**Table 2: Thresholds for rating quality power of the trial**


Rating	Power	Generalisability
5 	Powered to detect 0.2 or lower	Delivered in at least two diverse contexts
4 	Powered to detect 0.3 or lower	
3 	Powered to detect 0.4 or lower	
2 	Powered to detect 0.5 or lower	
1 	Powered to detect 0.6 or lower	
0 	Not powered to detect 0.6	

### Notes:

Generalisability or ‘external validity’ is highly dependent upon the audience and requires a judgement on behalf of the reader. For an individually randomised trial it might be possible to detect an effect of 0.2 or lower having been delivered in only a very small number of schools and the EEF believes, while



difficult to define, generalisability is important enough that trials should not be able to achieve 5  unless they are at least somewhat generalisable.







For this reason an additional condition has been added that in order to achieve 5  a trial needs to have been delivered in at least two diverse settings or contexts (e.g. two Local Authorities or area hubs). Some judgement is required in determining whether this has been achieved.

### 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 trails) 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).<sup>14</sup> Having discussed this with EEF panel members, the EEF has decided to develop its own attrition scale based purely on overall attrition, rather than relying on that created by the WWC. The reason for this is that it is difficult to judge whether attrition has led to bias. A low amount of balanced attrition could cause bias because it is not random. Whereas a high amount of differential attrition that occurs randomly could simply cause a reduction in the sample size. One way of judging the level of bias is to look at the balance on observable baseline characteristics after attrition (which is dealt with in the next section). However, this still does not tell us whether bias has occurred on unobservable characteristics (e.g. motivation). For this reason, the only way to be sure that bias has not occurred is to be strict about overall attrition.

**Table 3: Thresholds for overall attrition**

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

### Balance

The evaluator and reviewer should check whether or not the sample included in the final analysis is balanced on:

- Prior attainment (all trials)
- Free School Meals (FSM) (if effectiveness trial with FSM as the primary outcome) (or an equivalent in Australia's context e.g. School Card in South Australia)
- Other characteristics which may be critical for the analysis

A certain degree of judgement will need to be used in determining how the security rating should be adjusted to take account for the level of balance achieved in the final analysis.

Reviewers should consider whether the imbalance is likely to be due to chance (by referring the first criterion on Design and the level of imbalance in the total sample that was randomised) or due to differential attrition. Imbalance due to chance is less problematic, and can be more reliably accounted for by appropriate analysis methods.

The criterion allows the security rating to be adjusted downwards by a maximum of 2 s. One way of judging the level of balance is the size of the difference (using an effect size) between the treatment and control group on the above characteristics where appropriate. We provide suggestions for how ratings might be adjusted based on amount of balance in Table 4. However, this is just guidance.

**Table 4: Guidance on using balance to adjust security rating**

Balance (ES at baseline for primary analysis)	Adjustment guidelines
0.05 to 0.1 difference	Drop 1 s
0.1 or more difference	Drop 2 s

It is also important to take into account the efforts that an evaluator has made to use non-experimental techniques to create balance (e.g. matching techniques). So the balance calculations should be the amount of balance in the final sample analysed. The following rules should be taken into account:

- *The security rating cannot go above minimum of the first two criteria*
- *Evaluations that use non-experimental techniques to create balance cannot achieve more than 3 s*

### 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. The evidence of the fidelity of implementation and whether or not there are threats to the validity of the evaluation will be partly assessed in the process evaluation.

The reviewer should check whether there is any evidence of the following threats:

- 1. Insufficient description of the intervention:** The intervention should be clearly defined to enable replication, with the core components outlined, as well as any aspects that are adaptable. Appropriate fidelity checks should be carried out and any deviations from delivery should be documented and explained in order to know what the evaluation has tested. This should be available in the published report and manuals at the end of the trial.
- 2. Diffusion (also known as ‘contamination’):** Where treatment effects spread from the treatment to the control group. This is particularly a concern where the control group is within the same school and teachers or staff may share what they believe to be good practice with colleagues. This should be checked through the process evaluation.
- 3. Compensation rivalry or resentful demoralisation:** Where the control group behaviour alters as a result of the evaluation. For example, the participants in the control group may work harder or become less motivated as a result of their allocation. This is more of a concern when pupils are aware of their allocation, such as in individually randomised trials, and can be avoided by ensuring that participants in the control group are equally incentivised by providing a wait-list or alternative treatment. This should be checked through the process evaluation.
- 4. Evaluator or developer bias:** This occurs when the bias occurs due to either the evaluator or developer inadvertently behaving differently with the treatment and control group. For this reason it is very important that the randomisation, testing, and analysis are conducted by an independent evaluator blind to treatment. Evidence should be provided that appropriate measures were taken.
- 5. Testing bias:** School or delivery staff in the treatment arm may ‘teach to the test’ or helping children differentially in the treatment group. For this reason it is important that tests are blind. Also, it should be double checked that there is no possibility that the outcome measure was ‘inherent to treatment’. This should not be the case with E4L trials as all outcome measures should be reliable, valid and predictive measures of general literacy and numeracy.
- 6. Selection bias:** The concern that some differences on the primary outcome measure can be accounted for by the treatment group being systematically different from the control group on some unobservable characteristic, due to the way they were selected into the evaluation. This is only a threat for quasi-experimental designs.

Further information on all the nine threats to internal validity identified by Shadish, Cook and Campbell (2002) can be read in *Experimental and Quasi-Experimental Designs for Causal Inference*.<sup>15</sup> However, the ones above have been identified as those that are of particular concern for E4L evaluations.












If any of the above issues are identified as a cause for concern some judgement should be used in adjusting the security rating to account for any issues identified. The following are some suggested rules:

- If there is evidence of any substantial/significant threats the rating should drop 1 🔒
- If there is evidence of more than two substantial/significant threats the rating should drop 2 🔒

## Appendix 4: Worked examples











### EG1 Security rating summary: Chatterbooks


Rating	Criteria for rating			Initial Score	Adjust	Final Score	
	<b>Design</b>	<b>Power (MDES*)</b>	<b>Attrition<sup>3</sup></b>				
5 	Well conducted experimental design with appropriate analysis	MDES <0.2	0-10%		Adjustment for Balance [0]		
4 	Fair and clear quasi-experimental (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES <0.3	11-20%				
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES <0.4	21-30%	3 		Adjustment for threats to internal validity [ ]	3 
2 	Weakly matched comparison or experimental design with major flaws	MDES <0.5	31-40%				
1 	Comparison group with poor or no matching (e.g. volunteer versus others)	MDES <0.6	41-50%				
0 	No comparator	MDES <0.6	Over 50%				

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 3 
- **Reason for adjustment for balance (if made):** none made
- **Reason for adjustment for threats to validity (if made):** none made
- **Final padlock score:** initial score adjusted for balance and internal validity = 3

<sup>3</sup> Attrition should be measured at the student level (even for clustered trials) and from the point of randomisation to the point of analysis.











## EG2 Security rating summary: Calderdale



Rating	Criteria for rating			Initial Score		Adjust		Final Score
	<b>Design</b>	<b>Power (MDES*)</b>	<b>Attrition<sup>4</sup></b>					
5 	Well conducted experimental design with appropriate analysis	MDES <0.2	0-10%					
4 	Fair and clear quasi-experimental (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES <0.3	11-20%			Adjustment for Balance [0]		
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES <0.4	21-30%			Adjustment for threats to internal validity [0]		
2 	Weakly matched comparison or experimental design with major flaws	MDES <0.5	31-40%	2 				2 
1 	Comparison group with poor or no matching (e.g. volunteer versus others)	MDES <0.6	41-50%					
0 	No comparator	MDES <0.6	Over 50%					

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 2 
- **Reason for adjustment for balance (if made):** none made
- **Reason for adjustment for threats to validity (if made):** none made
- **Final padlock score:** initial score adjusted for balance and internal validity = 2

<sup>4</sup> Attrition should be measured at the student level (even for clustered trials) and from the point of randomisation to the point of analysis.

### EG3 Security rating summary: Fictional trial with adjustments needed for balance and threats to validity

Rating	Criteria for rating			Initial Score		Adjust		Final Score	
	<b>Design</b>	<b>Power (MDES*)</b>	<b>Attrition<sup>5</sup></b>						
5 	Well conducted experimental design with appropriate analysis	MDES <0.2	0-10%			Adjustment for Balance [-1]			
4 	Fair and clear quasi-experimental (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES <0.3	11-20%	4 					
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES <0.4	21-30%				Adjustment for threats to internal validity [-1]		
2 	Weakly matched comparison or experimental design with major flaws	MDES <0.5	31-40%						2 
1 	Comparison group with poor or no matching (e.g. volunteer versus others)	MDES <0.6	41-50%						
0 	No comparator	MDES <0.6	Over 50%						

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 4 
- **Reason for adjustment for balance:** much greater proportion of FSM pupils in control group than in the treatment group
- **Reason for adjustment for threats to validity:** process evaluation suggested that some teachers used some of the intervention techniques with the control children
- **Final padlock score:** initial score adjusted for balance and internal validity = 2 

<sup>5</sup> Attrition should be measured at the student level (even for clustered trials) and from the point of randomisation to the point of analysis.

## **Acknowledgements**

This document has been developed based on the EEF's guidance document: '[Classifying security ratings from EEF evaluations](#)'. Its classification system is developed by the EEF Panel of Evaluators and Evaluation Advisory Board who contributed their expertise and advice during subsequent period of consultation.