

March 2018

This document outlines the EEF’s policy on statistical analysis and effect size calculations. This guidance has been developed in association with the EEF’s Evaluation Advisory Group, other experts, and some members of EEF’s panel of evaluators. We are grateful for all the feedback we have received. This is a working document that we will continue to review and update to take account of methodological and analytical developments as well as evaluators’ experiences.

The main purpose of EEF evaluations is to provide high quality information to schools on the most effective approaches to teaching and learning that offer good value for money. Results from trials are not seen in isolation but reviewed and compared across projects. For this reason it is important that, whenever possible, analyses should be comparable across studies. With this aim, EEF has developed this guidance.

## Contents

INTRODUCTION .....	1
KEY PRINCIPLES OF THE STATISTICAL ANALYSIS GUIDANCE .....	1
PRIMARY OUTCOME ANALYSIS .....	5
ADDITIONAL ANALYSES .....	6
HOW TO ANALYSE MULTI-SITE TRIALS.....	11
FURTHER READING .....	12

## INTRODUCTION

The EEF believes that trial results should be as comparable as possible, and as such there is a basic framework with minimum requirements that we request all evaluators to follow. The Statistical Analysis Guidance includes key principles as well as guidance for the presentation of results and additional analyses.

All evaluators are expected to submit a detailed statistical analysis plan (SAP) three months after randomisation. This is peer-reviewed and published alongside the evaluation protocol on the EEF’s website. More details on the SAP can be found on [the EEF website](#).

In some circumstances, the current guidance may differ from the analysis that was specified in protocols or SAPs, particularly in those SAPs published prior to this guidance being updated. Where that is the case, both analyses can be reported; however, the effect sizes reported in the executive summary should be based on this guidance and deviations from the original protocol and SAP should be documented in the report.

## KEY PRINCIPLES OF THE STATISTICAL ANALYSIS GUIDANCE

The key principles of the guidance are:

1. Analyses must reflect study design and randomisation choices;
2. Analyses of primary and secondary outcome(s) should be undertaken on an ‘intention to treat’ basis;
3. Prior attainment should be controlled for using a regression model when the outcome is attainment. When the outcome is not attainment a parallel prior measure should be used;

4. Analytical methods should reflect the nested structure of educational data (i.e., clustered data);
5. Effect sizes (ES) for cluster randomised and multi-site trials should be standardised using unconditional variance in the denominator;
6. Sampling errors and uncertainty should be reported around all ES as confidence intervals (CI), or credibility intervals; and,
7. Intra-cluster correlations (ICCs) should be reported for post-test outcomes (and pre-test if available).

### 1) Analysis must reflect the design

The validity of a trial is dependent upon its design. Analytical methods should reflect study designs, randomisation choices (Rubin, 2008b) and the nested structure of educational data (Gelman, Hill, & Yajima, 2012; Gelman & Hill, 2007, pp. 245–246). Much of the guidance here applies to randomised controlled trials (RCTs), although the guidelines on using the intention to treat approach, clustering, subgroup analysis, ICCs and CIs are also relevant to analysis of quasi-experimental designs.<sup>1</sup>

Randomisation should normally be undertaken after baseline testing. Randomisation can involve some form of stratification or minimisation<sup>2</sup> that helps to obtain balanced groups in terms of characteristics that are deemed to be important predictors of the outcome [e.g., English as an Additional Language (EAL) for trials targeted at EAL pupils] or to aid intervention delivery (e.g., guarantee the same number of units assigned to each group across geographical areas). This is particularly important when the size of the sample is small enough that simple randomisation might yield groups with very different characteristics.

### 2) Use intention to treat analysis

Analyses of primary and secondary outcome(s) should be undertaken on an ‘intention to treat basis’, meaning that all those allocated to treatment and control conditions in the randomisation are included, wherever possible, in the final analysis, even if they drop out of the treatment (Torgerson & Torgerson, 2008). This means that, for all analyses, the maximum N should be used (as opposed to imposing a common sample where all analyses are based on the same pupils where there is no missing data for any of the variables used in the analytical model). This provides the most conservative estimate of impact and helps to preserve fully the benefit of randomisation.

In addition, means and standard deviations of pre-test and post-test scores should be summarised for each trial arm. Histograms of pre- and post-tests data distribution should also be presented.

Further analyses should be undertaken to estimate the potential benefit of the intervention as set out elsewhere in this guidance (e.g., treatment effects in the presence of non-compliance, sub-group analyses and missing data).

### 3) Control for prior attainment using a regression model

In a randomised design, the impact estimate on the primary outcome should be calculated using a regression model (e.g., ANCOVA) with pupil level outcomes to increase power and reduce bias<sup>3</sup> (van Breukelen, 2013, p. 907), with clustering accounted for in the model (Gelman et al., 2012).

The estimate reported in the executive summary, the headline estimate, should control for prior attainment using regression (e.g., an ANCOVA model using post-test as the outcome). This provides a

<sup>1</sup> RCTs and comparative observational studies should be seen on a continuum rather than a dichotomy in terms of suitability for causal inference (see Rubin 2008b, p. 810).

<sup>2</sup> For example, see the Minim Software available at <https://www-users.york.ac.uk/~mb55/guide/minim.htm>

<sup>3</sup> ANCOVA is better than CHANGE (the gain score approach) even if assignment of treatment is conditionally random on pre-test scores, for instance, pupils with lower prior attainment are more likely to be treated. As van Breukelen (2013, p. 907) argues, CHANGE takes pre-test imbalance “too seriously” and fails to take into account the regression to the mean phenomenon, which is accounted for by ANCOVA. According to Donald Rubin, even when the distributions of covariates are similar (this is what we mean by “balance”), it is still wise to adopt ANCOVA, because it has “possibly substantial positive effects” (2008a, p. 1352) on the precision of the point estimate.

more conservative and comparable method that can be used across EEF studies. Controlling for prior attainment also increases the precision of the estimate and increases statistical power. This measure of prior attainment can be either a commercial test (e.g. NGRT) or prior performance from the National Pupil Database (NPD) where the outcome is attainment. Currently, EEF prefers evaluators to use NPD data for this purpose.<sup>4</sup>

This pre-test should be included as a raw score at the pupil level, unless there are clear reasons to transform the data<sup>5</sup>. This will account for part of the variance explained at the pupil and school levels (Bloom, 2006). Alternatively, pre-tests can be controlled for by including school-level means, and pupil-level scores subtracting the school-level mean (Hedges & Hedberg, 2013). In either case, the share of variance explained at each level should be accounted for in the sample size calculations<sup>6</sup>.

Where additional variables have been used as part of randomisation (e.g., if randomisation is stratified on factors other than the treatment) these should be included in the primary analysis and should be pre-specified in the protocol and SAP (Rubin, 2008a, p. 1352).

For comparability, unless there are clear reasons otherwise, evaluations should only use the pre-test scores, the group status and design characteristics as covariates (for a discussion, please see Xiao, Higgins & Kasim, 2016 and Olken, 2015, p. 67)<sup>7</sup>. This way, we can best avoid the “fishing” problem (Humphreys, Sanchez de la Sierra, & Van der Windt, 2013; Simmons, Nelson, & Simonsohn, 2011) and the “curse of dimensionality” (Hayes, 2011). Moreover, this allows to promote transparency and reproducibility in scientific studies (Miguel et al., 2014; Nosek et al., 2015).

In addition, other specifications can be included as robustness checks and should be specified in the SAP. However, the headline estimate should always be based on the primary model specified above.

#### **4) Take account of clustering**

As noted above, analytical methods should reflect study designs and the nested structure of educational data. These include cluster randomised trials (CRT), simple randomised trials (SRT) and multi-site trials (MST)<sup>8</sup>.

Methods for cluster analysis include multilevel modelling [also known as ‘hierarchical linear modelling’ (HLM)] as advised by U.S. Department of Education’s What Works Clearinghouse (n.d.); and, variance components analysis<sup>9</sup>.

If clustering is not accounted for, the point estimates will be accurate, but the standard errors will be downward biased and resulting confidence intervals would be too narrow. This would inflate the potential contribution of the study in meta-analyses that use standard errors to weight the contribution of individual studies. This implies greater certainty than may be warranted.

For interventions randomised at the pupil level see further guidance in the “How to analyse multi-site trials” section below.

---

<sup>4</sup> See Allen et al (2018) on using commercial tests as pre-tests for EEF trials. Available at [https://educationendowmentfoundation.org.uk/public/files/Publications/EEF\\_Research\\_Papers/Research\\_Paper\\_1\\_-\\_Properties\\_of\\_commercial\\_tests.pdf](https://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf)

<sup>5</sup> One such exemption to use a standardised score is when there are multiple cohorts using different pre-tests or varying levels of difficulty. Any transformation of the data needs to be specified in the SAP prior to analysis.

<sup>6</sup> As Optimal Design assumes variance explained only at higher levels, we discourage its use for this purpose.

<sup>7</sup> The EEF acknowledges that adding further pupil level covariates reduces some of the total variance “to be explained” (Nakagawa & Cuthill, 2007, p. 597). However, it is not an approach that the EEF currently supports as part of its statistical analysis guidance.

<sup>8</sup> In cluster randomised trials, the unit of randomisation is the school, with pupils clustered in each school. Individual randomised trials are those where randomisation occur at the pupil level over a single school or “site”. Multi-site trials are those where pupils are randomised within schools over more than one school or “site”.

<sup>9</sup> Generalised estimating equations (GEE) is an alternative method to analyse clustered data. EEF discourages its use because it precludes the calculation of ICCs as required later in this guidance (7).

## 5) Report effect sizes (ES) based on total variances

Impact estimates should be reported as ES with CI. For comparability between EEF projects and with the wider literature, EEF requires ES calculations to be standardised. As Hedges'  $g^{10}$  is the ES used by both the What Works Clearinghouse and Campbell Collaboration, it is a suitable choice for these comparisons.

In multilevel models or mixed effect models, we assume that variations in a post-test outcome are due to different sources, which must be fully accounted for in a statistical model. By using total variance in the calculation of ES, we account for the nested structure of the data and potential differences between settings. This prevents inadvertent over estimation of ES (Xiao, Kasim & Higgins, 2016).

ES for cluster randomised trials with equal cluster size and using total variance can be calculated as:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}}}{s^*}$$

Where,

$(\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}}$  denotes ANCOVA difference in means between study groups adjusting for pre-test score and other stratification variables as specified in the relevant model.

$s^*$  denotes the pooled<sup>11</sup> unconditional variance of the two groups. Using the pooled estimate of variance assumes that the variances of both groups are estimates of the same population value. When there are reasons to believe this assumption is untenable (the treatment is expected to affect the *dispersion* of results), pooled estimates might not be adequate. In this case, the variance of the control group could be used instead which is equivalent to the calculation of ES in Glass (1976)<sup>12</sup>.

The choice of conditional or unconditional variance of outcomes as the denominator in the ES calculation has implications for the interpretation of results. If prior attainment is used as a covariate, the ES estimator using the conditional variance would be akin to the effects found by an experiment where pupils of the *same* previous attainment were randomised to treatment and control (Tymms, 2004). Even if this is a valid experiment, it is unlikely to be the policy parameter of interest (Schagen & Elliot, 2004, p. 56). Hence, evaluators should use the unconditional variance in the calculation of ES. Whenever available, ES should also be calculated with the population variance ( $\sigma^2$ ) instead of the pooled variances ( $s^{*2}$ )<sup>13</sup>. For transparency, evaluators should provide all parameters ( $(\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}}, s^*, \sigma^2, s_1^2, s_2^2, n_1, n_2$ ) to allow third parties to compute the ES of their interest. When using a different model from the one mentioned above (e.g., a multi-level model), please refer to further guidance in Hedges (2007) alongside the principles outlined here.

Note that the denominator of the ES calculation could be estimated with errors. However, Schochet & Chiang, (2011, p. 324) demonstrate that correcting for this error as suggested by Hedges (2007) has a trivial effect and can be ignored. Hence, dividing the adjusted ANCOVA difference in means by (unconditional or population) variance is valid.

<sup>10</sup> The difference between Hedges'  $g$  and Cohen's  $d$  is minimal for samples over 30 so either could be used in practice.

<sup>11</sup> This is a weighted average of the variance of both groups, not the estimate of the variance of all individuals pooled (See Coe, 2002). It can be calculated as:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Where,

$s_1^2$  is the variance of group 1; and equally defined for the other group.

$n_1$  is the number of individuals in group 1; and equally defined for the other group.

<sup>12</sup> This is less precise than Hedges'  $g$  using the pooled variance due to the smaller sample size of the control group in comparison to the full sample (Thompson, 2007)

<sup>13</sup> Using a population variance implies that the inference is made for the population of English pupils instead of restricted to those on the sample. This parameter can be obtained from NPD data when the outcome measure chosen are scores from National Curriculum Assessments. Most EEF trials use convenience samples that might not be generalisable to the broader population of English schools.

In the uncommon case where an outcome is defined as a binary variable, ES should be presented as risk ratios and difference in probabilities as they are simpler to interpret than other commonly used options. See Ferguson (2009) for a description of risk ratios and alternatives to present results using binary data. Also, in the presence of non-normal distributions or categorical data, other ES can be computed (for example, Mann-Whitney U test). See Fleiss (1994) and Fritz, Morris, & Richler (2012) for further guidance on this topic.

## 6) Report uncertainty

Statistical uncertainty should be reported around all ES. It is important to take into account the variation that is associated with any estimate using sampled data in understanding the minimum uncertainty associated with an estimate of impact (Wassertein & Lazar, 2016). However, acknowledging some limitations of frequentist CI and their associated hypotheses, evaluators may report uncertainty using other methods like bootstrapped CI, permuted  $p$  value (minimum of 1000 bootstrap or permutation runs) or a Bayesian credibility interval.

## 7) Report ICCs

For cluster randomised trials, the ICC should be calculated for the post-test (and pre-test, if there is one). Evaluators should report ICC at each level of clustering assumed in their design, but can report more if appropriate (e.g., form, when only clustering at school level was assumed).

## PRIMARY OUTCOME ANALYSIS

### Attainment

The EEF will usually identify academic or school attainment as the primary outcome in the trials it funds. This is because increased educational attainment is seen as a key aim for all EEF programmes.

### Number of primary outcomes

It is considered best practice for trials to have one primary outcome<sup>14</sup>. This is because multiple inferences are more prone to producing false-positive errors. Having one primary outcome also helps to minimise the risk of a false-negative error by providing the basis for the estimation of the sample size necessary for an adequately powered study.

The primary outcome needs to be defined at the time the study is designed. Selecting a primary outcome should be relatively straightforward when programmes being evaluated are clearly working to improve attainment in a specific subject, such as Maths or English. However, not all programmes evaluated by the EEF will be clearly linked to a single primary attainment outcome. This is particularly the case for interventions that aim to boost certain aspects of 'character', or programmes that aim to improve underlying aspects of learning such as metacognition, as these should have impacts on several areas of attainment.

The following guidance should be considered when defining the primary outcome:

- The EEF does not recommend combining measures across subjects to create a composite measure – such as Maths and English – for studies at the statutory school age, except when there is a precedent to do this (such as combined GCSE scores or when using measures of cognitive development in early years<sup>15</sup>). Outcome measures should be recognisable and understandable to practitioners.
- In efficacy trials, evaluators should aim for one primary outcome, but may need to allow for co-primary outcomes **if** the logic model and prior evidence support this and there is not a clear

<sup>14</sup> If a trial collects more than one primary outcome, yet is powered for the measurement of a single outcome and produces 95% confidence intervals for two outcomes, it is equivalent to multiple testing, as the probability of at least one type I error increases from 0.05 to somewhere between 0.05 and 0.0975 depending on the extent of correlation between the two outcomes.

<sup>15</sup> Demetriou et al (2017) suggest that, in the early years, cognitive measures like reading and maths are essentially not separable

rationale for choosing a specific outcome – such as Maths or English – for impact analysis. This would need to be reflected in sample size calculations.

- For effectiveness trials, the EEF will insist on one primary outcome.

When there are multiple primary outcomes, the EEF would expect multiple testing to be taken into account regardless of the type of trial. This can be done in a number of ways, including, but not limited to<sup>16</sup>:

#### *Applying a Bonferroni correction*

Using the Bonferroni correction, the standard 5% threshold for significance is divided by the number of outcomes. So if there are two outcomes, then the calculation is  $0.05/2 = 0.025$ , which is equivalent to 97.5% CI. As noted by (Benjamini & Hochberg, 1995), the Bonferonni correction is very conservative and does not account for correlated outcome measures.

*Insisting that two 95% CI from separate analyses must not overlap with zero, in the correct direction, to conclude success*

Trials are usually powered at 80%, so the probability of correctly concluding success when there is a genuine effect, over two outcomes, reduces to as low as 64% depending on the correlation between the two.

#### *Combining resulting ES*

This involves carrying out the two analyses separately and then taking the mean of the two resulting ES. Care is needed when calculating the variance of the combined ES using the formula in Borenstein, Hedges, Higgins, & Rothstein, (2009, p. 227). In a cluster randomised trial, the correlation between the two outcomes, required by the formula, is calculated at the school level.

## **ADDITIONAL ANALYSES**

### **Specification robustness checks**

In expectation, due to the randomisation of treatment, altering the regression specification should not have any effect on the point estimates of impact, but may change CI. Evaluators may propose additional secondary specifications to test the robustness of the results. These specifications may include:

- A model controlling for covariates that were imbalanced at baseline
- A simple model, including only the treatment assignment as covariate
- A saturated model, controlling for a vector of pre-treatment characteristics (gender, FSM, EAL, prior attainment, etc.)

However, the headline estimate of treatment effects should be that specified in the main model as referred in (3).

### **Subgroup analyses**

Subgroup analyses should be supported by theory and usually only conducted if pre-specified in the protocol and SAP. Evaluators should run analysis to explore interaction effects using the whole sample (e.g., using FSM, treatment allocation and treatment allocation\*FSM). Evaluators might also want to interrogate the subgroups using a separate model.

Any pre-specified sub-group analysis that is underpowered should be reported as exploratory. Likewise, results from any subgroup analysis that was not pre-specified in the protocol and SAP should be reported as a post-hoc exploratory analysis.

---

<sup>16</sup> EEF acknowledges that some of these approaches can be considered too conservative.

FSM should always be pre-specified as a subgroup to be analysed as an interaction effect and a sub-sample. The ES on the primary outcome for FSM pupils should be calculated using a separate model. FSM pupils are EEF's key target group and, although the findings will be less secure for this group when the trial is not powered to detect an effect on FSM, it is important that it is reported (with appropriate caveats) <sup>17</sup>.

The EverFSM indicator (EVERFSM\_6\_P) in the NPD should be used to conduct this analysis. This is likely to be a good indicator of the potential impact of disadvantage upon attainment. The table in Annex outlines the advantages and disadvantages of the different FSM measures available in the National Pupil Database (NPD).

Evaluators are also encouraged to explore differential treatment effects depending on previous attainment.

### **Treatment effects in the presence of non-compliance**

To avoid underestimating the potential benefit from interventions, further analysis according to compliance may be appropriate. This is because the intention to treat analysis may underestimate the efficacy of an intervention because some individuals, in either trial arm, will not adhere to their assigned treatment. Analyses in the presence of non-compliance give an indication of the treatment effects amongst those who participate in the intervention.

For this purpose, an Instrumental Variables (IV) approach should be used (Angrist & Imbens, 1995) because it tends to be more rigorous than per-protocol or on-treatment analyses (McNamee, 2009; Tillbrook et al. 2014). This will use a Two Stage Least Square (2SLS) approach with group allocation as the instrumental variable for the compliance indicator. Results for the first stage<sup>18</sup> should be reported alongside with i) the correlation between the instrument and the endogenous variable; and, ii) a F test.

The definition of compliance and how it will be measured should be agreed between the developer and the evaluation team, and discussed with EEF. The compliance indicator will be aligned with all inputs and activities that define the intervention as reflected in the logic model. If more than one input is used to define compliance, the compliance indicator could be a composite of all inputs. For instance, if an intervention includes teacher training and access to a software package, both activities will be required to define a unit as fully compliant with treatment.

The compliance indicator should be either binary or continuous. If consistent with the logic model, continuous variables can be used to explore the effects of partial compliance. Alternatively, minimal and optimal compliance thresholds can be defined and used to estimate bounds for the treatment effects. See Gerber & Green (2012, p. 165) for further details.

Depending on the characteristics of the intervention, compliance could be defined at either pupil, teacher or school level depending on who is responsible for the activities that define compliance. This does not need to coincide with the level of randomisation, or with the level at which outcomes are measured.

### **Missing data**

Although considerable resources should be invested in the follow-up of randomised schools and pupils, missing data is almost inevitable in an RCT. Two factors are important when analysing missing data: the extent of missingness and the patterns of missingness. Evaluators must specify the number of complete cases (i.e. those without any data missing), attempt to establish the missingness mechanism (i.e. what variables in the data are predictive of non-response) and discuss both in the report. The latter should be explored through a logistic regression model (multi-level, if required) where the presence of

---

<sup>17</sup> EEF acknowledges that most sub-group analyses would be underpowered as sample sizes are chosen to detect effects in the whole sample.

<sup>18</sup> The first stage predicts the compliance indicator using the treatment allocation as instrumental variable alongside all other covariates included in the second stage (See Angrist and Imbens, 1995).

missing data is modelled with additional information that might be predictive of missingness (i.e. not just variables in the main model). Interaction effects could be included at the evaluator's discretion.

Although not possible to conclude from the data alone, it is useful to consider the possible reasons for missing data as the appropriate analytical approaches differ depending on the patterns of missingness. The logistic model of missingness will help with this.

There are three types of missing data, described in Table 1. If a small number of cases are missing not at random (MNAR) and they have true values at either end of the distribution of valid cases, they can have a substantial influence on the main substantive model result. However, we would not typically expect the primary impact estimate to change when less than 5% of cases are missing, regardless of the pattern of missingness. Conversely, if a large proportion of data is missing, this would introduce biases depending on the pattern of missingness. It is possible, although arguably unlikely, that all missing data would be missing completely at random (MCAR), which is not expected to introduce biases. Hence, it is not advisable to use a threshold above which inference is not possible under any circumstances, as it would depend on the pattern of missingness.

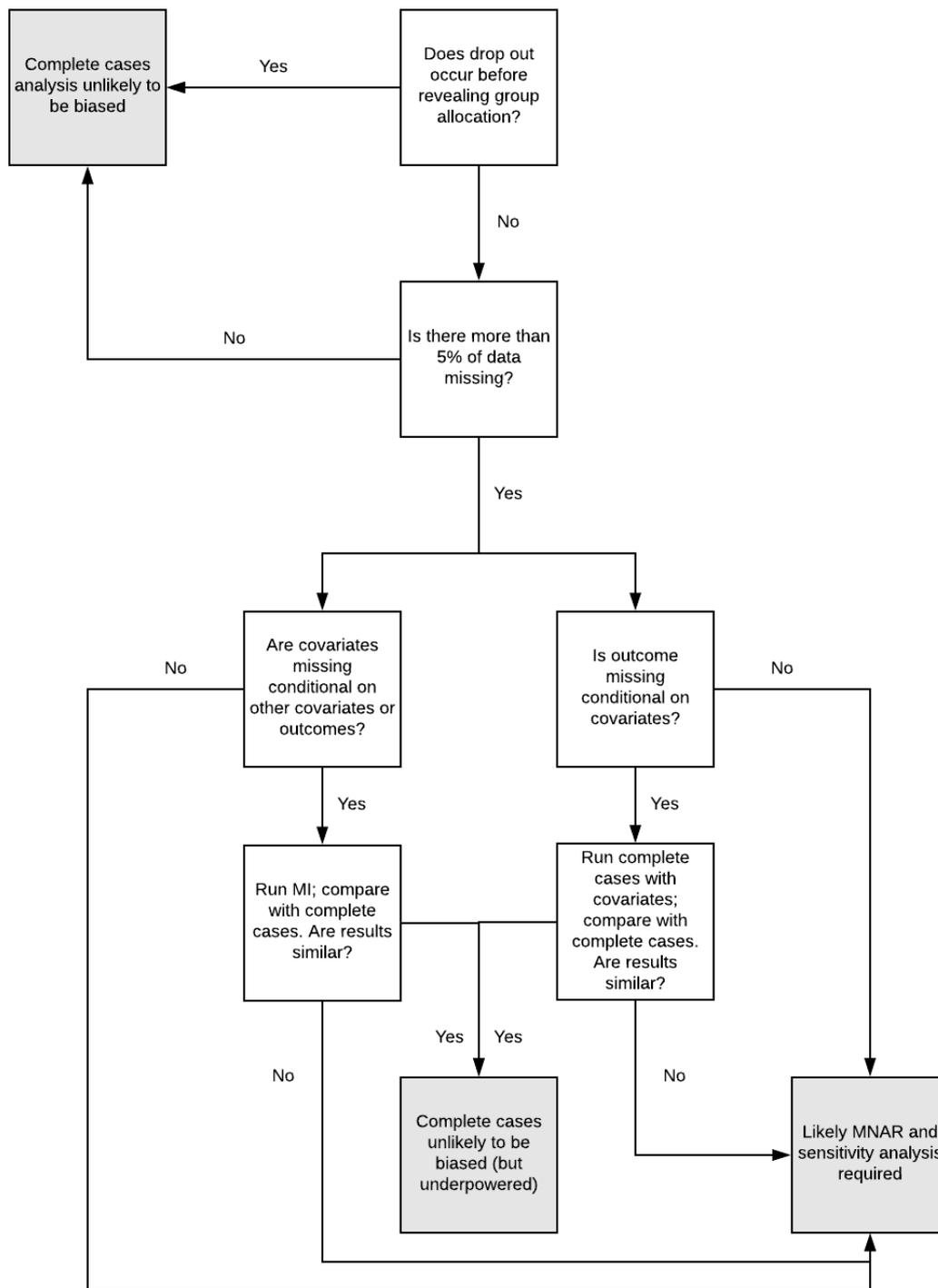
Table 1: Types of missing data and further analysis options

<b>Types of missing data</b>	<b>Description</b>	<b>Example</b>	<b>Further analysis</b>
<i>Missing completely at random (MCAR)</i>	If the reason for missing data is unrelated to any inference we wish to draw, missing observations are MCAR.	Pupils not attending due to sickness	Analysing only cases with observed data gives sensible, although less precise, results.
<i>Missing at random (MAR)</i>	If, given the observed data, the reason for missing does not depend on unseen data, then the missing observations are MAR. In this case, simply analysing the observed data is invalid.	Less able children (as measured by pre-test score) are more likely to be missing at follow-up and this is the only factor (associated with substantive model outcome) that is relevant.	<p>To obtain valid estimates, we have to include in an additional analysis the variables predictive of non-response. If only the outcome variable in a substantive model is MAR given covariates, no further work is needed but the model's interpretation is conditional on these covariates being included. Implications for this will need to be discussed clearly in the final report.</p> <p>However, if a covariate in the substantive model is MAR given other covariates, analysis should be done after multiple imputation (MI) of that covariate.</p> <p>Visit <a href="http://www.missingdata.org.uk/">http://www.missingdata.org.uk/</a> for more information on MI.</p> <p>Results from MI will need to be reported in addition to the headline impact estimates. Implications of</p>

			this analysis will need to be discussed clearly in the final report.
<i>Missing not at random (MNAR)</i>	If the reason for missing depends on an unobserved variable, even after taking into account all the information in the observed variables, then the missing observations are MNAR.	Less able children (as measured by post-test score) are more likely to be missing at follow-up and this tendency is not completely explained by pre-test score.	It is not possible to fix this scenario with MI alone and some sensitivity analysis needs to be reported alongside the headline impact estimates. Carpenter, Kenward, & White (2007) and Carpenter & Kenward (2007, p. 119) suggest some of these sensitivity analyses to assess results under MNAR. Implications of this analysis will need to be discussed clearly in the final report.

The following flow chart documents likely missing data scenarios, other than missing completely at random (MCAR), during an education RCT and possible solutions. Please note that drop-out after randomisation, but before allocation is revealed to schools or pupils, should be reported in the participant flow diagram, but not included in the intention to treat analysis. Note also that evaluators should focus on a robust MI model for the primary outcome rather than investing resources into MI for secondary outcomes and subgroups, for which results are more tentative anyway.

Figure 1: Flow chart for missing data analyses



Common *ad hoc* methods of dealing with missing data, which we do not recommend, include replacing missing values with the mean of the variable, creating a dummy variable to flag missing cases, 'last observation carried forward' and mean imputation using regression. These can be biased, lead to incorrect estimates of variance, or both, and should be avoided.

## HOW TO ANALYSE MULTI-SITE TRIALS

Trials that randomise pupils within schools across more than one school have specific analysis considerations. Such trials can be termed multi-site or randomised block designs. Before embarking on model choice, it is necessary to decide the type of inference we wish to draw (Hedges and Vevea, 1998). ‘Conditional inference’, where we do not attempt to generalise beyond the sample of schools within a trial, is more appropriate for efficacy trials and requires the use of a fixed effects model. ‘Unconditional inference’, where we wish to generalise to the population of schools from which trial schools were sampled, is more appropriate for effectiveness trials and requires the use of a random effects model and site-by-treatment interactions. Using random effects to derive conditional inferences will result in CI that are too wide (Hedges and Vevea, 1998). Table 2 compares the features of these two model types.

Table 2. Analysis considerations for the fixed and random effects modelling of multi-site trial data (site is usually a school).

<b>Fixed site effect</b>	<b>Random site effect with site-by-treatment interaction</b>
Single-level model (dummy variable for site)	Multi-level model (highest level is site); intervention coefficient random at site level
Intervention effect estimate based on within-site variation	Intervention effect estimate based on both within-site and between-site variation
Assumes no site-by-treatment interaction	Models site-by-treatment interaction
No statistical basis for generalisation	Can generalise result to the population of schools from which trial schools were (randomly) sampled
Sample size calculations with no design effect are usually slightly conservative	In addition to the ICC, sample size depends on a further parameter: Hedges’ $\omega$ , the proportion of variance that is due to the site-by-treatment interaction (Hedges and Rhoads, 2009).

## FURTHER READING

The following resources should be referred to for further guidance on analysis. In addition, the EEF can refer evaluators to an expert statistician, if desired.

Torgerson, C. J., Torgerson, D. J., & Styles, B. (2013). Randomised Trials in Education: An Introductory Handbook. p.16-26. EEF.

[https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting\\_up\\_an\\_Evaluation/Rand\\_omised\\_trials\\_in\\_education-revised031213.pdf](https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Rand_omised_trials_in_education-revised031213.pdf)

What Works Clearinghouse (n.d). Procedures and Standards Handbook, Version 3.0, p.22-32: [https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc\\_procedures\\_v3\\_0\\_standards\\_handbook.pdf](https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf)

Borenstein, M., Hedges, L. V., Higgins, J. P T., & Rothstein, H. R. (2009). Introduction to meta-analysis. London: Wiley, pp. 21-32.

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale Journal of Biology and Medicine*, 86(3), 343-358.

## REFERENCES

Allen, R., Jerrim, J., Parameshwaran, M. & Thompson, D. (2018). Properties of commercial tests in the EEF database. EEF Research Paper No. 001.

Angrist, J., & Imbens, G. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *American Statistical Association*, 90(430), 431-442.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1), 289-300.

Bloom, H. (2006). Learning more from social experiments. *Evolving analytical approaches*. Russell Sage Foundation.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. Sussex: John Wiley & Sons.

Carpenter, J., & Kenward, M. (2007). *Missing data in randomised controlled trials - a practical guide*. London: LSHTM.

Carpenter, J., Kenward, M., & White, I. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16, 259-275.

Demetriou, A., Merrell, C., & Tymms, P. (2017). Mapping and predicting literacy and reasoning skills from early to later primary school. *Learning and Individual Differences* 54: 217-225.

Ferguson, C. (2009). An effect size primer: A guide for clinicians and researchers. *American Psychological Association*, 40(5), 532-538.

Fleiss, J. (1994). Measure of effect size for categorical data. In H. Cooper, & L. Hedges, *The handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.

Fritz, C., Morris, P., & Richler, J. (2012). 2012. *Journal of Experimental Psychology*, 141(1), 2-18.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. <http://doi.org/10.2277/0521867061>

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. <http://doi.org/10.1080/19345747.2011.618213>

Gerber, A., & Green, D. (2012). *Field Experiments: Design, analysis and Interpretation*. W.W. Norton & Company.

Glass, G. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Research*, 5(10), 3-8.

Hayes, B. (2011). An adventure in the Nth dimension. *American Scientist*, 99, 442–446. <http://doi.org/10.1511/2011.93.442>

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics* 32, 4:. 341 - 370 <https://doi.org/10.3102/1076998606298043>

Hedges, L. V. and Rhoads, C. (2009). *Statistical Power Analysis in Education Research* (NCSE 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Available: <https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf> [4th December 2017].

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39-65.

Hedges, L., & Hedberg, E. (2013). Intraclass correlations and covariate outcome correlations for planning two and three level cluster randomized experiments in education. *Evaluation Review*, 37(6):445-489.

Hedges, L.V. & Vevea, J.L. (1998). Fixed and Random-Effects Models in Meta-Analysis. *Psychological Methods* 3: 486-504.

Humphreys, M., Sanchez de la sierra, R., & Van der windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), 1–20. <http://doi.org/10.1093/pan/mps021>

Lin, W., & Green, D. P. (2015). *Standard Operating Procedures: A Safety Net for Pre-Analysis Plans*. Berkeley. Retrieved from [www.stat.berkeley.edu/~winston/sop-safety-net.pdf](http://www.stat.berkeley.edu/~winston/sop-safety-net.pdf)

McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine*, 28(21), 2639-2652.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... Van der Laan, M. (2014). Promoting transparency in social science research. *Science* (New York, N.Y.), 343(6166), 30–1. <http://doi.org/10.1126/science.1245317>

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, 82(4), 591–605. <http://doi.org/10.1111/j.1469-185X.2007.00027.x>

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <http://doi.org/10.1126/science.aab2374>

Olken, B. A. (2015). Promises and Perils of Pre-Analysis Plans. *Journal of Economic Perspectives*, 29(3), 61–80. <http://doi.org/10.1257/jep.29.3.61>

Rubin, D. B. (2008a). Comment: The Design and Analysis of Gold Standard Randomized Experiments. *Journal of the American Statistical Association*, 103(484), 1350–1353. <http://doi.org/10.1198/016214508000001011>

Rubin, D. B. (2008b). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 2(3), 808–840.

Schagen, I., & Elliot, K. (2004). But what does it mean? The use of effect sizes in educational research. Slough: NFER.

Schochet, P., & Chiang, H. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioural Statistics*, 36(3), 307-345.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366. <http://doi.org/10.1177/0956797611417632>

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44(5), 423-432.

Tymms, P (2004). Effect sizes in multilevel models. National Foundation for Educational Research.

Tilbrook, H. E., Hewitt, C. E., Aplin, J. D., Semlyen, A., Trehela, A., Watt, I., & Torgerson, D. J. (2014). Compliance effects in a randomised controlled trial of yoga for chronic low back pain: a methodological study. *Physiotherapy*, 100(3), 256-262.

Torgerson, D. J., & Torgerson, C. J. (2008). *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*. London: Palgrave Macmillan.

van Breukelen, G. J. P. (2013). ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. *Multivariate Behavioral Research*, 48(6), 895–922. <http://doi.org/10.1080/00273171.2013.831743>

Wassertein, R., & Lazar, N. (2016). The ASA's statement on p-values: context, process and purpose. *The American Statistician*, 70(2), 129-133.

Xiao Z., Kasim, A., Higgins, S.E. (2016) Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials *International Journal of Educational Research* 77: 1-14 <http://dx.doi.org/10.1016/j.ijer.2016.02.001>

Xiao, Z. Higgins, S. & Kasim, A (2017) An Empirical Unravelling of Lord's Paradox. *The Journal of Experimental Education* <http://dx.doi.org/10.1080/00220973.2017.1380591>

**ADVANTAGES AND DISADVANTAGES OF USING DIFFERENT FSM MEASURES FOR SUBGROUP ANALYSIS**

MEASURE	Potential advantages	Potential disadvantages
<b>FSM</b>	<ul style="list-style-type: none"> <li>• Simple, easy to understand</li> <li>• Most likely to remain over time</li> </ul>	<ul style="list-style-type: none"> <li>• FSM variable often retrieved from NPD at the same time as outcomes data which may not capture FSM during the period of the intervention (in the case of long term follow-ups when the intervention is longer than a year)</li> </ul>
<b>FSM6</b>	<ul style="list-style-type: none"> <li>• Used for pupil premium allocation</li> <li>• Pupils in FSM6 have lower attainment than FSM</li> <li>• Larger group than FSM</li> <li>• Those 'extra' pupils included in FSM6 are more similar in terms of attainment to FSM pupils than Non-FSM.</li> </ul>	<ul style="list-style-type: none"> <li>• More likely to change in future than FSM</li> <li>• Is a function of the age of the child and the time they have been in the state funded education system</li> <li>• Some analytic approaches rely on historical data and FSM6 is only available from 2009/10</li> </ul>
<b>FSMever</b>	<ul style="list-style-type: none"> <li>• Larger group than FSM and FSM6</li> <li>• Those 'extra' pupils included in FSMever are more similar in terms of attainment to FSM6 pupils than Non-FSM. This appears to be an important distinction, particularly for secondary schools</li> </ul>	<ul style="list-style-type: none"> <li>• More likely to change in future than FSM</li> <li>• Is a function of the age of the child and the time they have been in the state funded education system, as such, this tends to increase with year group so it is not directly comparable between, say, lower primary and upper secondary</li> <li>• Some analytic approaches rely on historical data and FSMever is only available from 2009/10</li> </ul>

Sources: Treadaway, M and Thomson, D (2014) Using longitudinal school census data, presentation to the PLUG user group workshop, 17<sup>th</sup> June 2014. <http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/treadawayandthomson2014.pdf>