YOUTH ENDOWMENT FUND



Analysis guidance

This document outlines the Youth Endowment Fund's policy on statistical analysis and effect size calculations.

YEF analysis guidance for efficacy and effectiveness trials

This document outlines the YEF's policy on statistical analysis and effect size calculations. This guidance has been adapted from the EEF's[1] statistical analysis guidance and in collaboration with the YEF's Technical Advisory Group, other experts, and some members of YEF'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 YEF evaluations is to provide high quality information to practitioners and policy-makers on the most effective approaches to preventing young people from getting involved in crime and violence, which offer good value for money. Results from individual trials should not be 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, YEF has developed this guidance.

1 The Youth Endowment Fund are very grateful to the Education Endowment Foundation for their comprehensive work on analysis that provided the essential framework for this guidance.



Contents

Introduction 4

Key principles of the statistical analysis guide 4

Primary outcome analysis 10

Additional analyses 11

How to analyse multi-site trials 16

How to report analysis in efficacy and effectiveness reports 17

Further reading 20

Annex 1 23

Annex 2 24

Introduction

Effect size estimates can vary widely as a result of the choices that evaluators make (Xiao, Kasim & Higgins, 2016). Trial results should ideally be as comparable as possible, and for this reason this guidance provides a basic framework including key principles and minimum requirements with respect to the conduct and presentation of YEF-funded analyses and results, that we request all evaluators to follow.

All evaluators are expected to submit a detailed statistical analysis plan (SAP). This is peer-reviewed and published alongside the evaluation protocol on the YEF's 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 guide

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. An important predictor should be controlled for using a regression model;
4. Analytical methods should reflect the study design and take account of clustering;
5. Effect sizes (ES) for cluster randomised and multi-site trials should be standardised using unconditional variance in the denominator;
6. Some measure of 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; Abadie et al, 2017) and, where relevant, the nested structure of the 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.[2]

Randomisation should normally be undertaken after baseline testing. Randomisation can involve some form of stratification or minimisation[3] that helps to obtain balanced groups in terms of characteristics that are deemed to be important predictors of the outcome (e.g. prior arrests and convictions for interventions aiming to reduce re-offending) 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. The primary impact estimate should be based on observed data and should not use imputed outcome data. Where missing data analyses are undertaken (e.g. multiple imputation), these should be reported as sensitivity analyses alongside, rather than replacing, the primary impact estimate.

In addition, means and standard deviations of continuous baseline and outcome measures should be summarised for each trial arm. Histograms of baseline and outcome data distribution should also be presented. For categorical data report counts (the numerator and denominator) and percentages in each category.

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, subgroup analyses and missing data).

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

3. Control for important predictors 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 participant level outcomes to increase power and reduce bias[4] (van Breukelen, 2013, p. 907), with clustering accounted for in the model where relevant (Gelman et al., 2012).

The estimate reported in the executive summary, the headline estimate, should control for one or two important predictors using regression (e.g. an ANCOVA model using post-test as the outcome). Controlling for predictors increases the precision of the estimate and increases statistical power. Where available, the baseline measure of the primary outcome should normally be included as the primary covariate. Additional covariates should only be included where they were used in randomisation or are predictive of the outcome and have been pre-specified in the SAP.

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 one or two important predictors, the group status and design characteristics as covariates (for a discussion, please see Xiao, Kasim & Higgins, 2016 and Olken, 2015, p. 67)[5]. 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 us 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 or sensitivity analyses and should be specified in the SAP. However, the headline estimate should always be based on the primary model specified above.

4 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 SDQ scores 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.

5 Adding further participant level covariates reduces some of the total variance "to be explained" (Nakagawa & Cuthill, 2007, p. 597). However, it is not an approach that the YEF currently recommends in its statistical analysis guidance.

4. Analytical methods should account for clustering

As noted above, analytical methods should reflect study designs and, if applicable, the nested structure of the data. These include cluster randomised trials (CRT), simple randomised trials (SRT) and multi-site trials (MST)[6].

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[7].

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 participant level within clusters (e.g. schools, local authorities or pupil referral units) see further guidance in the "How to analyse multi-site trials" section below.

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

Impact estimates should be reported as ES with CI. For comparability between YEF projects and with the wider literature, YEF requires ES calculations to be standardised. As Hedges' g[8] is the ES used by the Campbell Collaboration, it is a suitable choice for these comparisons.

In multilevel models or mixed effect models, we assume that variations in outcomes 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 clusters, sites or 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:

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{s^*}$$

6 In cluster randomised trials, the unit of randomisation is the cluster. Individual randomised trials are those where randomisation occurs at the participant level over a single cluster or "site". Multi-site trials are those where participants are randomised within clusters or over more than one cluster or "site".
7 Generalised estimating equations (GEE) is an alternative method to analyse clustered data. YEF discourages its use because it precludes the calculation of ICCs as required later in this guidance (7).
8 The difference between Hedges' g and Cohen's d is minimal for samples over 30 so either could be used in practice.

Where,

$(\overline{Y}_T - \overline{Y}_C)_{adjusted}$ adjusted denotes ANCOVA difference in means between study groups adjusting for one or two important predictors and other stratification variables as specified in the relevant model.

s* denotes the pooled[9] unconditional standard deviation of the two groups. Using the pooled unconditional standard deviation 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)[10].

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 offending is used as a covariate, the ES estimator using the conditional variance would be akin to the effects found by an experiment where participants of the same prior offending 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 standard deviation (σ) instead of the pooled standard deviation (s*)[11]. For transparency, evaluators should provide all parameters

$((\overline{Y}_T - \overline{Y}_C)_{adjusted}, s^*, \sigma^2, {s_1}^2, {s_2}^2, n_1, n_2)$ 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 the (unconditional or population) standard deviation is valid.

9 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.

10 This is less precise than Hedges' g using the pooled standard deviation due to the smaller sample size of the control group in comparison to the full sample (Thompson, 2007)

11 Using a population standard deviation implies that the inference is made for the population instead of restricted to those on the sample.

Where an outcome is defined as a binary variable, ES should be presented as risk ratios and natural frequencies as they are simpler to interpret than other commonly used options such as odds ratios. See Ferguson (2009) for a description of risk ratios and alternatives to present results using binary data. Odds ratios from analyses with dichotomous outcomes can be transformed into ES comparable to Hedges' g using the Cox Index as in the equation below which does not require additional assumptions (What Works Clearinghouse, 2017, p12)[12].

$$d_{Cox} = \frac{\left[\ln(\frac{p_t}{1-p_t}) - \ln(\frac{p_c}{1-p_c})\right]}{1.65}$$

Where p_t is the probability of occurrence in the treatment group and p_c is the probability of occurrence in the control group.

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

Evaluators must present a measure of uncertainty 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 (Wasserstein & 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), which do not rely on the assumption of random sampling, or a Bayesian compatibility intervals which rely on less stringent assumptions. The ASA's Special Issue, Statistical inference in the 21st century: A world beyond p<0.05, offers some suggestions. P-values, if used, should be presented as a continuous probability, and any dichotomous interpretation around 0.05 should be avoided.

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. practitioner, when only clustering at the site level was assumed).

12 Estimating the Cox Index and then transforming it into a measure of months of progress assumes a normal distribution which even if not necessarily true, is a reasonable assumption made by most statistical models.

Primary outcome analysis

Number of primary outcomes

The YEF will usually identify violence or offending, or a predictor of violence or offending as the primary outcome in the trials it funds. It is considered best practice for trials to have one primary outcome[13]. 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.

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

- The YEF does not recommend combining measures to create a composite except when there is a precedent to do this (such as combined SDQ scores). Outcome measures should be recognisable and understandable to practitioners. When composite outcomes are used, additional exploratory analyses should be included to ascertain if the results on some of the subjects are driving the results.
- 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 rationale for choosing a specific outcome for impact analysis.
- For effectiveness trials, the YEF will insist on one primary outcome.

Multiple testing

Many YEF evaluations include multiple outcomes, time points, or subgroup analyses. Multiple testing increases the likelihood of false-positive findings and should be considered when designing analyses and interpreting results. YEF recommends a proportionate approach: a clearly pre-specified primary outcome and primary time point should remain the focus of the main analysis, while secondary and exploratory analyses should be interpreted with appropriate caution. Evaluators should justify their approach to multiple testing in the statistical analysis plan and refer to Annex 2 for detailed guidance.

13 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.

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 analyses to explore interaction effects or other appropriate tests for heterogeneity using the whole sample (e.g. using gender, treatment allocation and treatment allocation*gender) and include the estimated difference between each subgroup with confidence intervals. Evaluators might also want to interrogate the subgroups using a separate model.

Any pre-specified subgroup 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.

When analysing subgroup effects, race and ethnicity should be included where meaningful to do so.

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; Tilbrook 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[14] should be reported alongside with i) the correlation between the instrument and the endogenous variable; and, ii) an 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 the YEF. The compliance indicator should 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 attendance at an event and access to a software package, both activities will be required to define a unit as fully compliant with treatment.

Depending on the characteristics of the intervention, compliance may involve measures of quality and quantity and could include components at different levels (e.g. participant, practitioner or cluster) depending on who is responsible for the activities that define compliance. This does not need to coincide with the level of randomisation, or the level at which outcomes are measured. The compliance indicator could be either binary, categorical 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.
The model should account for the clustered structure of the data. There are two broad alternatives to do so: use a structural modelling approach with a multi-level setting or use a 2SLS approach clustering the standard errors.

14 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

Although considerable resources should be invested in the follow-up of randomised participants, 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, to account for the clustered structure of the data, if required) where the presence of 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. This may be done separately for outcome variables and covariates included in the headline model.

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.

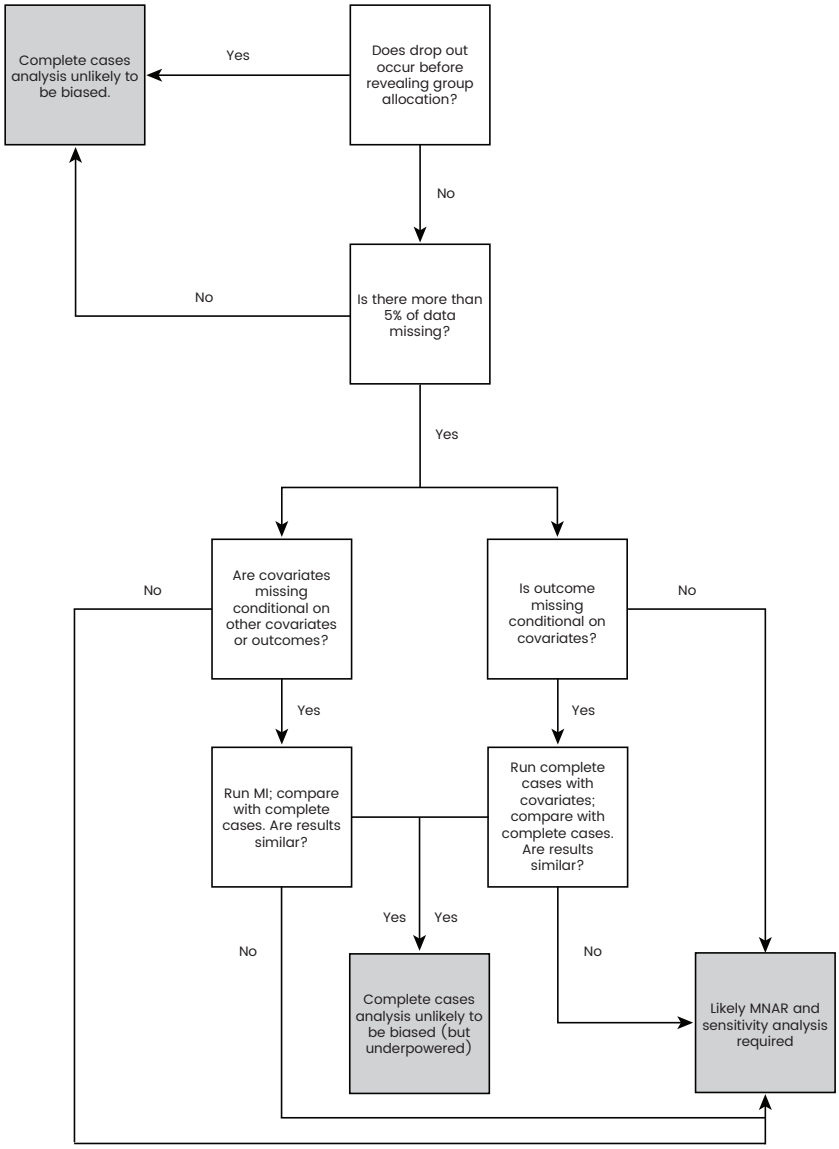
There are three types of missing data, described in Table 1. If a small number of cases are missing not at random (MNAR) and they have true values at either end of the distribution of valid cases, they can have a substantial influence on the main substantive model result. However, we would not typically 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

Types of missing data	Description	Example	Further analysis
Missing completely at random (MCAR)	If the reason for missing data is unrelated to any inference we wish to draw, missing observations are MCAR.	Participants not attending a workshop due to sickness	Analysing only cases with observed data gives sensible, although less precise, results.
Missing at random (MAR)	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.	Children with lower SDQs (as measured at a baseline) are more likely to be missing at follow-up and this is the only factor (associated with substantive model outcome) that is relevant.	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. However, if a covariate in the substantive model is MAR given other covariates, analysis should be done after multiple imputation (MI) of that covariate. Visit http://www.missingdata.org.uk/ for more information on MI. Results from MI will need to be reported in addition to the headline impact estimates. Implications of this analysis will need to be discussed clearly in the final report.
Missing not at random (MNAR)	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.	Children with lower SDQs (as measured by post-assessment 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 RCT and possible solutions. Please note that drop-out after randomisation, but before allocation is revealed to participants, 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. The headline impact estimate should be based on observed outcome data only. Multiple imputation analyses should be reported as sensitivity analyses.

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 participants within clusters, such as schools, sites or clinics across more than one cluster 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 sites 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 sites from which trial sites 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.

Fixed site effect	Random site effect with site-by-treatment interaction
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' w, the proportion of variance that is due to the site-by-treatment interaction (Hedges and Rhoads, 2009).

How to report analysis in efficacy and effectiveness reports

Most of YEF's expectations around reporting are outlined in the Trial Reporting Template. This section highlights three areas that are particularly important for the application of YEF's security rating system and the preparation of YEF executive summaries.

Reporting of sample size and MDES at randomisation

Evaluators should report the sample size randomised and the minimum detectable effect size (MDES) implied by that sample size and the assumptions specified in the protocol and SAP.

In YEF-funded trials, this can be more complicated than in conventional trials because many studies use rolling recruitment and randomisation. As a result, there may be a substantial period between the first and final participant being randomised. Consequently, the MDES reported in the original protocol or statistical analysis plan (SAP) will differ from the MDES achieved once recruitment and randomisation have been completed.

Where possible, evaluators should update the SAP to report the final sample size and MDES achieved at randomisation. Where an updated SAP is not produced, reports should include an updated table showing the final numbers randomised and the corresponding MDES. This information is required to support interpretation of the study findings and YEF's assessment of the strength of the evidence.

Being clear on the valence of scales and direction of effects

YEF evaluations use a wide range of outcome measures. For some outcomes, higher scores indicate better outcomes (for example, school attendance or attainment). For others, higher scores indicate worse outcomes (for example, offending, behavioural difficulties).

To aid interpretation, evaluators should clearly state the valence of each primary and secondary outcome measure and explain the direction of any reported effects. Reports should make clear whether an increase in score represents an improvement or deterioration in outcomes.

This is particularly important when reporting standardised effect sizes. A positive effect size does not always indicate a positive impact. For outcomes where lower scores are desirable, a negative effect size may represent an improvement. For example, on measures where higher scores indicate greater behavioural difficulties, a negative effect size may indicate a reduction in difficulties and therefore a positive impact.

Evaluators should provide sufficient information for readers to understand both the direction of change and whether that change represents an improvement, deterioration, or no meaningful difference between groups.

Evaluators are also encouraged to check results carefully to ensure that results have been interpreted correctly in the context of the measure properties. Care should be taken when interpreting outcomes where reductions are desirable, as misinterpreting the direction of the scale can lead to incorrect conclusions about programme impact.

An example approach to reporting outcome valence and direction of effects is provided below:

> *… between baseline and follow-up, an improvement in behavioural problems was seen for both the intervention and control groups. Table X shows the results for the headline ITT analysis of the primary outcome; teacher SDQ externalising score. The measure is scored from 0–20 and lower scores represent fewer behavioural problems; thus a negative effect size indicates an effect in the desired direction. The effect size for the primary outcome was -0.07 (CI low: -0.22; CI high 0.07), providing tentative evidence of positive effect for the programme, though this did not reach statistical significance (p=0.33).*

Reporting of effect sizes as Hedges' g

Evaluators should report Hedges' g for all primary outcomes and secondary outcomes. Reporting a common standardised effect size supports consistency across YEF evaluations and enables comparison of findings across studies using different outcome measures.

Where other effect measures are more appropriate for the outcome being analysed, such as incidence rate ratios (IRRs), evaluators should report these measures alongside Hedges' g. For example, outcomes relating to offending or other count data may be most naturally analysed and reported using an IRR. However, an equivalent Hedges' g should also be provided. Evaluators are not required to use YEF's internal effect size categories when reporting findings. However, interpretations should be consistent with them and should accurately reflect the magnitude and direction of the observed effect. Using different thresholds for communicating scale of impact, such as Cohen's original suggestions around small, medium and large effects should be avoided. Reports should focus on the size of the estimated effect and the associated uncertainty, rather than relying solely on statistical significance.

The thresholds YEF applies through the drafting of each Executive Summary are presented in the tables that follow. They are based on detailed analytical work described in the YEF Effect Size Database. This means they are specific to YEF reporting and are used to support consistent interpretation across evaluations. They should not be interpreted as universal benchmarks for effect size magnitude.

Positive effects

Cohen's *d* or Hedges *g*	YEF impact rating
0.00 ≤ d < 0.02	No impact
0.02 ≤ d < 0.10	Small
0.10 ≤ d < 0.25	Moderate
0.25 ≤ d	Large

NB In this table, the start point of each given range is part of the relevant effect size, while the end point marks the start of the next interval

Negative effects

Cohen's *d* or Hedges *g*	YEF impact rating
0.00 ≥ d > −0.02	No impact
−0.02 ≥ d > −0.10	Small
−0.10 ≥ d > −0.25	Moderate
−0.25 ≥ d	Large

NB In this table, the start point of each given range is part of the relevant effect size, while the end point marks the start of the next interval



Further reading

The following resources should be referred to for further guidance on analysis. In addition, the YEF 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/Randomised_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

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). 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 (NCSER 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

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., Trewhela, 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

Wasserstein, 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

Annex 1

Advantages and disadvantages of using different FSM measures for subgroup analysis

Measure	Potential advantages	Potential disadvantages
FSM	• Simple, easy to understand • Most likely to remain over time	• 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)
FSM6	• Used for pupil premium allocation • Pupils in FSM6 have lower attainment than FSM • Larger group than FSM • Those 'extra' pupils included in FSM6 are more similar in terms of attainment to FSM pupils than Non-FSM.	• More likely to change in future than FSM • Is a function of the age of the child and the time they have been in the state funded education system • Some analytic approaches rely on historical data and FSM6 is only available from 2009/10
FSMever	• Larger group than FSM and FSM6 • 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	• More likely to change in future than FSM • 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 • Some analytic approaches rely on historical data and FSMever is only available from 2009/10

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

Treadaway, M (2014) 'Pupil Premium and the invisible group'. FFT Research Paper No. 5, June 2014. http://www.fft.org.uk/FFT/media/fft/News/FFT-Research-Pupil-Premium-and-the-Invisible-Group.pdf

Annex 2

Context and multiple testing at YEF

- Multiple testing occurs when we have more than one statistical test in the same study. At YEF, this is common as we often pre-specify:
  - Multiple secondary outcomes (and in rare cases multiple primary outcomes)
  - The measurement of outcomes at multiple time points (e.g., T1, T2, and T3)
  - Test effects for different subgroups
- Each test has a chance of detecting a false positive – saying there is an effect when there isn't one (Type 1 error, capped at 5% per test). However, when we run multiple tests, the chance of at least one false positive increases.
- Often, correcting for multiple testing is recommended (e.g., Romano-Wolf correction) to avoid Type 1 errors. However, overcorrecting can lead to missed true effects (increased Type 2 error).[15]
- Deciding whether to correct for multiple testing requires balancing key trade-offs. Table A2.1 below summarises the main considerations.

Table A2.1: Summary of main considerations:

Considerations	Correcting for multiple tests	Not correcting (with precise interpretation)
False positives (Type 1 error)	Controlled – reduced chance of spurious findings from chance alone	Potentially inflated – greater chance of Type 1 error
Risk of missing true effects (Type 2 error)	Higher – more conservative and can mask true effects	Lower – more power to detect true effects
Interpretation caution (avoid selective reporting)	Built in via adjusted thresholds and discourages cherry picking results	Must be managed through transparent reporting and careful language
Secondary outcomes	Risk of downgrading important findings Correcting for multiple testing is often considered best practice	Allows flexibility and consideration of pre-specified secondary outcomes
Distinguishing the 'type' of outcome	Can collapse all non-primary outcomes into 'exploratory' outcomes	Preserves meaningful distinctions between secondary and exploratory outcomes
Ease of communication	Easier to interpret with cleaner significance thresholds	Requires nuance and care in interpretation

15 It is worth noting that even trials published in top medical journals lack consensus as to the optimum way to report primary and secondary outcomes of RCTs - https://www.jacc.org/doi/10.1016/j.jacc.2021.06.024

YEF guidelines on multiple testing

These principles are designed to support consistent and transparent analysis and interpretation across YEF evaluations. They apply to multiple outcomes, time points, subgroups and repeated measures.

Principle 1: The primary outcome must be clearly pre-specified

- The protocol should pre-specify the primary outcome(s), and the specific time point at which it will be assessed, linking to the theory of change and expected timing of impact.
- YEF consider it best practice to have one primary outcome.[16] Where there is a single primary outcome, it will be used for the main impact analysis and will not require correction for multiple testing.
- In rare examples where there are multiple primary outcomes, YEF expect multiple testing to be taken into account as part of the primary analysis. Noting that this becomes less relevant where the initial analysis shows no impact.

Principle 2: Secondary outcomes do not require formal correction but must be interpreted with caution

- Correction is **not required** if there is a clearly pre-specified primary outcome, and secondary outcomes are hierarchically pre-specified in the protocol (e.g., by choosing what is best powered and/or what is most clinically important).
- Findings are interpreted precisely (e.g., intervention X had an effect on Y outcome at Z time). Reports must not accidentally overstate effects on secondary outcomes by using general statements about impact.
- Correction **is recommended** where there are many secondary outcomes without clear hierarchical pre-specification and /or where secondary outcomes are considered confirmatory (e.g., a pre-specified hypothesis test that has adequate statistical power to detect an expected effect size) and therefore may drive policy/ practice recommendations.

Principle 3: Where multiple time points are included, the primary analysis must be pre-specified

- Pre-specify one time point for the primary analysis – e.g., T2.
- Other time points should be treated as secondary unless clearly justified.

16 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.

Principle 4: Subgroup and interaction analyses are exploratory in most cases

- Subgroup and moderation tests must be pre-specified and powered to be considered confirmatory.
- Otherwise, they are considered exploratory, and results must be clearly labelled as such.
- **Correction is encouraged** if multiple subgroups are tested to avoid overstating effects.

Principle 5: SAPs should include a section on multiple testing that

- States whether correction will be applied.
- Justifies the choice.
- Specifies how many outcomes, timepoints, and subgroup tests are included.
- Describes how findings will be interpreted.

Principle 6: All outcomes must be reported transparently, with multiplicity risks reflected

- All pre-specified outcomes (primary, secondary and exploratory) must be reported. Selective reporting of outcomes or time points is not acceptable. Evaluators must clarify in the report text which outcomes are pre-specified, and which are exploratory.
- The number of tests conducted must be acknowledged.
- Appropriate discussion about Type 1 and Type 2 error trade-offs should be reported.

Table A2.2: Summary table – when to correct for multiple testing

Scenario	Correction required?	Notes
One primary outcome and few secondary outcomes	No	Ensure secondary outcomes are hierarchically pre-specified in the protocol **and** interpret secondary outcomes with caution
One primary outcome and many secondary outcomes[17]	No	Correction is not required if the secondary outcomes are pre-specified, interpreted individually, and not used to support broad claims.
	Yes	Correction is recommended if there are many secondary outcomes without clear hierarchical pre-specification and /or where secondary outcomes are considered confirmatory and therefore may drive policy/practice recommendations.
Multiple primary outcomes / equally weighted outcomes	Yes	Correction controls the inflated error risk and avoids cherry-picking. Alternately, use hierarchical outcome structure
Multiple time points, with one pre-specified (e.g., T2)	No	Treat the pre-specified time point as primary; interpret other time points as secondary and with caution
Subgroup or interaction tests	Yes - if confirmatory and powered	Required if making a formal claim that the intervention has a differential effect by group
	No – if exploratory	Tests must be clearly labelled and interpreted cautiously

17 Please note, YEF do not recommend including a large battery of secondary outcomes.

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

Youth Endowment Fund

This document was last updated in **August 2026.**

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

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413